LREC-COLING 2024

The Joint Workshop of the 7th Financial Technology and Natural Language Processing (FinNLP), the 5th Knowledge Discovery from Unstructured Data in Financial Services (KDF), and the 4th Economics and Natural Language Processing (ECONLP) Workshop

(FinNLP-KDF-ECONLP 2024)

Workshop Proceedings

Editors
Chung-Chi Chen, Zhiqiang Ma, Udo Hahn

20 May, 2024
Torino, Italia

Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing (FinNLP), the 5th Knowledge Discovery from Unstructured Data in Financial Services (KDF), and the 4th Economics and Natural Language Processing (ECONLP) Workshop

ISBN 978-2-493814-19-7
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Message from the Organizers

Welcome to the Joint Workshop of FinNLP-KDF-ECONLP, marking a significant milestone as one of the largest gatherings in our community's history. This workshop, collocated with LREC-COLING 2024 in Turin, Italy, underscores our commitment to fostering international collaboration and knowledge exchange in the dynamic fields of NLP and AI, as they intersect with finance and economics. We extend a warm invitation to all participants, whether you are joining us in person or virtually, to immerse yourself in a productive exchange of ideas and insights throughout FinNLP-KDF-ECONLP-2024.

This edition has witnessed a notable surge in discussions surrounding ESG/CSR, reflecting its growing attention on the finance for social good. Alongside two shared tasks that echo this theme, it's evident that the conversation is significantly increasing. In line with trends observed in recent conferences, large language models (LLMs) have dominated the technical discourse, highlighting their pivotal role in advancing our field.

We extend our heartfelt appreciation to all authors who contributed to the main track and participants of the shared tasks. Your dedication to sharing groundbreaking findings and innovations is the driving force behind the workshop's continued success and expanding impact. We're also immensely grateful to the program committee members who dedicated their time and expertise in reviewing submissions and steering the selection for the workshop. Additionally, we would like to express our gratitude to our invited speakers, Flavius Frasincar (Erasmus University Rotterdam), Diyi Yang (Stanford University), and James Zhang (Ant Group), for delivering inspiring keynote speeches.

In closing, we express our deepest gratitude to Project JPNP20006, sponsored by the New Energy and Industrial Technology Development Organization (NEDO). Your generous support has been crucial in achieving FinNLP's goals and furthering research in this vibrant area of study.

We wish you an enriching and enjoyable experience at FinNLP-KDF-ECONLP.

Chung-Chi Chen, Xiaomo Liu, Udo Hahn, Armineh Nourbakhsh, Zhiqiang Ma, Charese Smiley, Véronique Hoste, Sanjiv Ranjan Das, Manling Li, Mohammad Ghassemi, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen

FinNLP-KDF-ECONLP Organizers

Organizing Committee

Chung-Chi Chen, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan

Xiaomo Liu, JP Morgan AI Research, US

Udo Hahn, TexKnowlogy, Germany

Armineh Nourbakhsh, JP Morgan AI Research, US

Zhiqiang Ma, JP Morgan AI Research, US

Charese Smiley, JP Morgan AI Research, US

Veronique Hoste, Ghent University, Belgium

Sanjiv Ranjan Das, Santa Clara University, USA

Manling Li, University of Illinois Urbana-Champaign, US

Mohammad Ghassemi, Michigan State University, US

Hen-Hsen Huang,Institute of Information Science, Academia Sinica, Taiwan

Hiroya Takamura, Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology, Japan

Hsin-Hsi Chen, Department of Computer Science and Information Engineering, National Taiwan University, Taiwan

Program Committee

E. M. Ion Androutsopoulos, Athens University of Economics and Business

Alaa Alhamzeh, University of Passau

Paulo Alves, Catolica Porto Business School

Petr Babkin, JPMorgan Chase AI Research

Ruihan Bao, Mizuho Securities Co., Ltd.

Damir Cavar, Indiana University Bloomington

Yung-Chun Chang, Taipei Medical University

Emmanuele Chersoni, The Hong Kong Polytechnic University

Nelson Correa, Andinum, Inc.

Min-Yuh Day, National Taipei University

Ruihai Dong, Insight Centre for Data Analytics, University College Dublin

Pablo Duboue, Independent Researcher

Flavius Frasincar, Erasmus University Rotterdam

Akshat Gupta, University of California, Berkeley

Sarang Gupta, Columbia University

Petr Hajek, University of Pardubice

Keiko Harimoto, Mizuho Securities Co., Ltd.

Masanori Hirano, Preferred Networks Inc.

Kiyoshi Izumi, The University of Tokyo

Jinhang Jiang, Walmart Inc.

Juyeon Kang, 3DS Outscale

Simerjot Kaur, JPMorgan Chase

Wai Lam, Chinese University of Hong Kong

Qing Li, Southwestern University of Finance & Economics

Chenyang Lyu, MBZUAI

Ismail El Maarouf, BizGive

Pekka Malo, Aalto University

Puneet Mathur, University of Maryland

Igor Mozetic, Jozef Stefan Institute

Hilal Pataci, University of Texas at San Antonio

Yulong Pei, Eindhoven University of Technology

Senja Pollak, Jozef Stefan Institute

Nicolas Pröllochs, Universität Gießen

Samuel Rönnqvist, Zefort (formerly Aivan AI)

Bryan R. Routledge, Carnegie Mellon University

Hiroki Sakaji, Hokkaido University

Kazuhiro Seki, Konan University

Selja Seppala, University College Cork

Agam Shah, Georgia Institute of Technology

Kiyoaki Shirai, Japan Advanced Institute of Science & Technology (JAIST)

Ankur Sinha, Indian Institute of Management

Kexuan Sun, University of Southern California

Joy Sain, JPMorgan Chase

Pranav Shetty, JPMorgan Chase

Zehra Sura, Bloomberg

Alan Spark, McKinsey & Company

Ke Tian, Rakuten

Anh Truong, Tubi

Arnav Wadhwa, Chainlink Labs

Chong Wang, Amazon

Ran Zmigrod, JPMorgan Chase

Chuan-Ju Wang, Academia Sinica

Shih-Hung Wu, Chaoyang University of Technology

Shiyu Wang, Emory University

Frank Z. Xing, National University of Singapore

Annie T.T. Ying, IBM

Jichang Zhao, Beihang University

Table of Contents

Construction of a Japanese Financial Benchmark for Large Language Models

Masanori Hirano
Preferred Networks, Inc.
Tokyo, Japan
research@mhirano.jp

Abstract

With the recent development of large language models (LLMs), models that focus on certain domains and languages have been discussed for their necessity. There is also a growing need for benchmarks to evaluate the performance of current LLMs in each domain. Therefore, in this study, we constructed a benchmark comprising multiple tasks specific to the Japanese and financial domains and performed benchmark measurements on some models. Consequently , we confirmed that GPT-4 is currently outstanding, and that the constructed benchmarks function effectively. According to our analysis, our benchmark can differentiate benchmark scores among models in all performance ranges by combining tasks with different difficulties.

Keywords: Large Language Model, Benchmark, Finance, Japanese

1. Introduction

Recently, Large Language Models (LLMs) have demonstrated excellent performance. In particular, the latest models, such as ChatGPT(OpenAI, 2023a) and GPT-4(OpenAI, 2023b), exhibit high performance and significant generalization abilities. The basis of these models begins with the transformer (Vaswani et al., 2017) and BERT(Devlin et al., 2019), and GPT series (Radford et al., 2018, 2019; Brown et al., 2020) were developed using the transformer. Other LLMs have also been proposed, such as Bard(Google, 2023), LLaMA(Touvron et al., 2023a,b), Dolly(Databricks, 2023), BLOOM(Scao et al., 2022), Vicuna(Vicuna, 2023), PaLM(Chowdhery et al., 2022; Anil et al., 2023), and Gemini (Team, 2023).

The major difference between the latest LLMs and previous language models, such as BERT, is that one model can answer questions in multiple languages and domains and respond to questions by following the instructions. Previously, BERT was trained separately in different languages and domains (SUZUKI et al., 2023). However, the latest LLMs, such as GPT4, can freely process multiple languages. Moreover, whereas BERT can only fill in incomplete sentences, the latest LLMs can answer questions in the same manner as humans.

Because of these improvements, the evaluation tasks should be reconstructed. The latest LLM performances far exceed those of previous language models regarding the variety and accuracy of questions they can answer. Therefore, a greater variety of questions is necessary to evaluate LLMs more accurately. Thus, evaluation tasks are important for developing high-performance LLMs.

Currently, some evaluation tasks for LLMs have already been prepared, but are insufficient as concerns domain-specified tasks and those for languages other than English. For instance, a language model evaluation harness (lm_eval) (Gao et al., 2021) was proposed for LLM evaluation using several English tasks. Moreover, several domain-specified tasks have been evaluated using GPT-4(OpenAI, 2023b). Eulerich et al. (2023) evaluated it using certified public accountant (CPA) tests, Nori et al. (2023) tested it in the medical domain, and its applications to legal services were also tested (Iu and Wong, 2023; Choi et al., 2023). However, only a small number of domain-specified tasks have been tested, and the response of LLMs to other tasks is still being investigated comprehensively.

This study focuses on evaluations of the Japanese financial domain. Financial services are relatively large as concerns money spendings. Moreover, according to World Bank data[1], Japan has the third-largest listed capital market in the world as of 2020. Therefore, the usability of LLMs in Japanese and financial domains is a crucial issue.

Several studies have been conducted on Japanese LLMs. Various models such as CyberAgent's CALM series, Rinna's model, stabilityai's stablelm series, Elyza's model, Preferred Networks' Plamo™, and LLM-jp-13B have been proposed. However, few models have been published in academic research papers, and their performances have not been thoroughly evaluated. Other studies have tuned existing English-based models to specialize in Japanese-language use(HIRANO et al., 2023; Sukeda et al., 2023; Suzuki et al., 2023). As for the Japanese task evaluation for LLMs, several benchmarks are available, including the jlm_eval(StabilityAI, 2023), llm-jp-eval(LLM-jp,

[1] https://data.worldbank.org/indicator/CM.MKT.LCAP.CD

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 1–9

2024), and Rakuda benchmarks[2].

However, no benchmarks or LLMs are specified for both Japanese and financial domain.

Thus, this study proposes a new benchmark for the Japanese financial domain and evaluates several models specified for Japanese. The benchmark and performance results of the models are publicly available at `https://github.com/pfnet-research/japanese-lm-fin-harness`.

2. Related Works

Studies on specialized language models in finance and Japanese have been conducted for a long time. The classic vector embedding technique used in language processing is word2vec (Mikolov et al., 2013). Word2vec has also been used in the financial domain HIRANO et al. (2019). After word2vec, ELMo (Peters et al., 2018), which uses a bidirectional long short-term memory (LSTM) (Schuster and Paliwal, 1997) to pre-train a distributed representation, appeared, along with transformer (Vaswani et al., 2017), which is a good alternative to LSTM in time-series processing, and transformer-based BERT (Devlin et al., 2019).

In contrast, the methodologies to fit language models to specific languages or domains are also pursued. For instance, Howard and Ruder (2018) proposed universal language model fine-tuning. Following this study, some domain- or language-specific language models were developed, such as SciBERT (Beltagy et al., 2019), MedBERT (Rasmy et al., 2021), Japanese BERT[3], and Japanese financial BERT (SUZUKI et al., 2022). Moreover, the methodologies and effects of domain-specified fine-tuning were discussed in(Gururangan et al., 2020; SUZUKI et al., 2023).

In the era of LLMs, although several transformer-based language models have been proposed, as described in the Introduction section, several unknown mechanisms of LLMs exist and numerous trials have been performed.

Several proposed LLMs that focus specifically on finance exist. For instance, BloombergGPT(Wu et al., 2023) is a private LLM focused on finance. In addition, publicly available models, such as FinLLAMA(William Todt, 2023), which is a tuned version of LLaMA(Touvron et al., 2023a), FinGPT(Yang et al., 2023), and Instruct-FinGPT(Zhang et al., 2023), exist.

Japanese-focused LLMs and benchmarks have also been developed, as mentioned in the Introduction section.

However, currently, no LLMs and benchmarks focused on the Japanese financial domain exist. Therefore, in this study, we construct a benchmark.

3. Japanese Financial Benchmark Dataset

We construct a new Japanese financial benchmark for LLMs, comprising the following five benchmark tasks:

- chabsa: Sentiment analysis task in the financial field.
- cma_basics: Fundamental knowledge questions in securities analysis.
- cpa_audit: Tasks on auditing in the Japanese Certified Public Accountant (CPA) exam.
- fp2: Multiple choice questions for 2nd grade Japanese financial planner exam.
- security_sales_1: Practice exam for the 1st grade Japanese securities broker representative test.

For chabsa and cpa_audit, we constructed a dataset using corpora from previous studies. We constructed the remaining tasks by crawling and cleansing the documents available on the Internet. In the following section, we describe these tasks in detail. For each task, an example prompt is shown below, but this is only for illustrative purposes. Several other types of prompts were also prepared, and those prompts were originally written in Japanese. For details of the prompts, please refer to the aforementioned public repository.

3.1. chabsa: Sentiment Analysis Task in the Financial Field

chabsa (Kubo et al., 2018) is a task to determine the sentiments of specific words with respect to sentences contained in securities reports. In Japan, listed companies publish securities reports annually. These data are available from `https://github.com/chakki-works/chABSA-dataset`. Three types of sentiments exist: positive, negative, and neutral. However, the number of neutral words is extremely small, which may hinder a stable performance evaluation. Therefore, we decided to treat it as a binary classification task, that is, positive or negative classification. This implies that data tagged as "neutral" will be regarded as incorrect regardless of whether the output is positive or negative. Because all the questions were two-choice questions, a random response would yield approximately 50% correct answers. For the final evaluation values, we employed the macro-f1 value.

[2] `https://yuzuai.jp/benchmark`

[3] `https://huggingface.co/tohoku-nlp/bert-base-japanese`

In this dataset, 4334 positive, 3131 negative, and 258 neutral responses were observed. Therefore, the random response yields an f1 value of 49.15 points.

An example of chabsa

Please indicate the sentiment of the targeted word in the following sentences, whether positive or negative.

Sentence: The Japanese economy continued to gradually recover during the fiscal year ending March 31, 2012.
Target Word: Japanese economy
Answer: positive

3.2. cma_basics: Fundamental Knowledge Questions in Securities Analysis

cma_basics questions basic knowledge in securities analysis. It was created by crawling and cleansing sample questions from the securities analyst examination. Therefore, it differs from the first and second rounds of the Japanese securities analyst examination administered by the Securities Analysts Association of Japan. However, it has the same characteristics as the first-round test, including a multiple-choice format. In addition, questions containing figures were deleted and the tables were translated into a markdown format. Since all questions had four choices, randomly selecting an answer results in 25.00% accuracy.

An example of cma_basics

Please answer the letter corresponding to the appropriate choice for the following question.

Question:
Which of the following statements about the Japanese economy is incorrect?
A: Real GDP (real gross domestic product) is the level of production activity excluding the effects of price fluctuations.
B: Inflation implies a sustained increase in the general price level.
C: Indirect finance is a form of financial intermediation in which banks and other financial intermediaries play a central role in mediating money lending and borrowing.
D: The fiscal policy of the Bank of Japan adjusts the price level through an increase or decrease in money supply.

Answer:
D

3.3. cpa_audit: Tasks on Auditing in the Japanese CPA Exam

cpa_audit is a collection of short-answer questions on audit theory from the Japanese CPA examination, and data from a previous study (Masuda et al., 2023) were used. It contains 360 questions with six choices and 38 questions with five choices. Therefore, 16.98% of the questions could be answered correctly if they are answered randomly.

An example of cpa_audit

Please answer the letter corresponding to the appropriate combination of symbols to answer the following questions:

Question:
Choose the most appropriate combination of the following statements regarding CPA audits.
(i) In a stock company, the management has a fiduciary responsibility to properly manage and invest the capital contributed by shareholders and provide an accounting report to shareholders regarding the results of this management responsibility. CPA audits of these financial reports contribute to proper management accountability.
(ii) CPA audit not only plays a role in ensuring the reliability of financial information but also supports corporate governance because it encourages the correction of internal control deficiencies and fraudulent acts discovered in the process.
(iii) As listed companies have a significant influence on society, special provisions are placed on CPAs who audit listed companies, such as the prohibition of independent audits, prohibition of certain non-audit attestation services, and restrictions on employment.
(iv) Because a listed company can raise funds widely from general investors, several interested parties arise, and protection against them is necessary. Therefore, establishing a management system for timely and appropriate disclosure of information to stakeholders is necessary. Therefore, CPAs must perform an internal control audit when a company is newly listed.

Choices:
A: (i) and (ii)
B: (i) and (iii)
C: (i) and (iv)
D: (ii) and (iii)
E: (ii) and (iv)
F: (iii) and (iv)

Answer:
A

3.4. fp2: Multiple Choice Questions for 2nd Grade Japanese Financial Planner exam

fp2 is the choice question for a 2nd grade Japanese financial planner exam. The past questions from the Japan FP Association's 2nd grade financial planning skills examination from May 2021 to September 2023 were obtained from the official HP[4] and processed. Questions containing figures were removed, and the tables were translated into a markdown format. Because all the questions had four choices, a random answer yielded 25.00% correct answers.

An example of fp2

Please select the appropriate answer to the following question using numbers from 1 to 4:

Question:
Which of the following statements regarding the conduct of financial planners ("FP") toward their clients is most inappropriate as concerns the relevant laws and regulations?
1. Mr. A, an FP who is not qualified as a lawyer, was consulted by a client about adult guardianship and provided a general explanation on the difference between legal and voluntary guardianship.
2. Ms. B, who is not a licensed tax accountant, received a client's consultation regarding the deduction of medical expenses for income tax purposes and explained that the amount of medical expenses paid, which is compensated for by insurance proceeds, is not deductible as a medical expense deduction.
3. Mr. C, an FP who is not a licensed social insurance consultant, received consultation from a client regarding the deferral of receipt of the basic old-age pension and estimated the pension amount in the case of deferral based on the estimated amount of pension receipt in the client's pension benefit report.
4. Mr. D, an FP who is not registered as a financial instruments business operator, concluded an investment advisory contract regarding asset management with the client and recommended the purchase of individual stocks that were expected to rise in value.

Answer:
4

[4] https://www.jafp.or.jp/exam/mohan/

3.5. security_sales_1: Practice Exam for the 1st Grade Japanese Securities Broker Representative Test

security_sales_1 is a practice exam task that corresponds to the first level of the Japanese securities broker representative test. It was created by crawling and cleansing to obtain practice examinations and sample questions for the 1st-grade Japanese securities broker representative test. Consequently, some differences in the question structure and difficulty levels from official Japanese securities broker representative tests exist. It contains 29 questions with four choices and 28 questions with two choices. Therefore, even if the questions were answered randomly, 37.28% of correct answers could be obtained.

An example of security_sales_1

Please answer the letter corresponding to the appropriate choice for the following question.

Question:
Please answer if the following statement is correct or incorrect:
A securities broker representative is deemed to have the authority to perform all judicial acts on behalf of the financial instrument firm to which they belong with respect to acts prescribed by law, such as the purchase and sale of securities.

Choices:
A: Correct
B: Wrong

Answer:
B

4. Experiments: Benchmark Calculation for LLMs

We measured the benchmarks for various models using the benchmarks described in the previous section.

Given the significant impact of prompts on performance, we prepared prompts for each task in addition to the prompts presented in the previous section. These prompts were similar to those employed in previous Japanese-specific benchmark studies (StabilityAI, 2023). Preliminary experiments with 0–4 shots were conducted using these prompts, and the best-performing prompts and numbers of shots were employed for the final experiment. Although this procedure may seem to be a type of in-sample training, in practice, we believe that such an evaluation procedure would provide a fair comparison. This is because the number of prompts was limited,

Table 1: All Benchmark Results. Some low-performance models are omitted. See full results at the repository as previously mentioned

Model	Ave.	chabsa	cma_basics	cpa_audit	fp2	security_sales_1
openai/gpt-4-32k	66.27	93.16	81.58	37.44	50.74	68.42
openai/gpt-4	66.07	93.20	78.95	37.69	50.32	70.18
openai/gpt-4-turbo	64.59	92.86	76.32	36.18	50.95	66.67
Qwen/Qwen-72B	62.18	92.36	78.95	32.91	40.00	66.67
Qwen/Qwen-72B-Chat	57.89	92.52	78.95	29.90	28.42	59.65
rinna/nekomata-14b	56.03	89.70	63.16	25.13	42.53	59.65
Qwen/Qwen-14B	55.95	90.73	63.16	22.61	38.32	64.91
Qwen/Qwen-14B-Chat	54.71	91.56	65.79	22.36	32.42	61.40
rinna/nekomata-14b-instruction	54.43	91.27	63.16	24.12	37.47	56.14
stabilityai/japanese-stablelm-base-beta-70b	53.07	90.87	60.53	22.36	33.68	57.89
stabilityai/japanese-stablelm-instruct-beta-70b	52.77	91.85	60.53	22.86	36.00	52.63
tokyotech-llm/Swallow-13b-instruct-hf	52.32	87.79	60.53	19.60	35.79	57.89
openai/gpt-35-turbo	50.27	89.98	52.63	18.09	29.26	61.40
meta-llama/Llama-2-70b-hf	50.21	89.37	57.89	20.85	30.32	52.63
lightblue/qarasu-14B-chat-plus-unleashed	50.04	89.69	57.89	20.35	31.37	50.88
rinna/nekomata-7b-instruction	49.90	90.34	47.37	22.61	27.79	61.40
Qwen/Qwen-7B-Chat	49.86	86.38	50.00	20.85	32.42	59.65
meta-llama/Llama-2-70b-chat-hf	49.53	90.29	52.63	18.84	28.00	57.89
Qwen/Qwen-7B	48.67	85.11	57.89	19.35	30.11	50.88
elyza/ELYZA-japanese-Llama-2-13b	48.37	88.37	47.37	19.35	28.84	57.89
tokyotech-llm/Swallow-13b-hf	48.31	87.59	52.63	19.60	32.63	49.12
Xwin-LM/Xwin-LM-13B-V0.2	47.53	88.11	52.63	22.11	25.68	49.12
rinna/nekomata-7b	47.12	79.18	42.11	21.61	33.05	59.65
meta-llama/Llama-2-13b-chat-hf	46.98	87.95	52.63	19.60	27.37	47.37
elyza/ELYZA-japanese-Llama-2-7b-fast	46.04	82.52	44.74	17.84	30.74	54.39
elyza/ELYZA-japanese-Llama-2-13b-fast	45.70	86.37	39.47	20.60	31.16	50.88
lmsys/vicuna-13b-v1.5-16k	45.57	85.81	52.63	19.10	28.21	42.11
mosaicml/mpt-30b-instruct	45.18	83.27	42.11	21.36	26.53	52.63
meta-llama/Llama-2-7b-chat-hf	44.86	83.70	39.47	20.35	29.89	50.88
llm-jp/llm-jp-13b-instruct-full-jaster-v1.0	44.66	85.91	39.47	20.10	26.95	50.88
elyza/ELYZA-japanese-Llama-2-13b-instruct	44.27	89.40	44.74	18.59	26.53	42.11
meta-llama/Llama-2-13b-hf	44.19	82.04	36.84	20.85	30.32	50.88
rinna/youri-7b-instruction	43.84	86.88	34.21	21.61	27.37	49.12
llm-jp/llm-jp-13b-instruct-full-dolly-oasst-v1.0	43.76	83.23	39.47	19.60	27.37	49.12
rinna/youri-7b-chat	43.67	86.67	36.84	19.60	26.11	49.12
cyberagent/calm2-7b-chat	43.67	81.09	36.84	18.09	29.68	52.63
llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0	43.60	86.83	39.47	18.59	24.00	49.12
elyza/ELYZA-japanese-Llama-2-13b-fast-instruct	43.59	87.27	42.11	18.59	26.11	43.86
lmsys/vicuna-33b-v1.3	43.44	87.81	34.21	19.60	28.21	47.37
lmsys/vicuna-7b-v1.5-16k	43.21	84.78	39.47	19.60	24.84	47.37
mosaicml/mpt-30b-chat	43.10	86.40	39.47	21.36	24.42	43.86
elyza/ELYZA-japanese-Llama-2-7b	42.99	83.48	42.11	19.60	25.89	43.86
tokyotech-llm/Swallow-7b-hf	42.91	72.27	39.47	19.60	28.84	54.39
pfnet/plamo-13b	42.87	76.97	39.47	21.61	27.16	49.12
mosaicml/mpt-30b	42.80	83.44	36.84	19.60	26.74	47.37
stabilityai/japanese-stablelm-base-alpha-7b	42.73	78.74	34.21	19.10	30.74	50.88
Xwin-LM/Xwin-LM-7B-V0.2	42.73	82.79	42.11	19.85	25.05	43.86
llm-jp/llm-jp-13b-v1.0	42.39	81.24	39.47	19.10	26.53	45.61
cyberagent/calm2-7b	41.96	80.02	42.11	17.84	24.21	45.61
rinna/japanese-gpt-neox-3.6b-instruction-ppo	41.89	74.71	44.74	20.60	23.79	45.61
rinna/youri-7b	41.84	73.60	34.21	19.10	29.68	52.63
elyza/ELYZA-japanese-Llama-2-7b-fast-instruct	41.59	82.53	39.47	20.10	25.47	40.35
stabilityai/japanese-stablelm-instruct-alpha-7b	41.43	78.94	34.21	19.35	23.79	50.88
tokyotech-llm/Swallow-7b-instruct-hf	41.36	83.61	31.58	18.09	24.42	49.12
stabilityai/japanese-stablelm-instruct-alpha-7b-v2	41.36	78.62	34.21	19.10	24.00	50.88
pfnet/plamo-13b-instruct	41.13	77.33	39.47	21.11	27.37	40.35
rinna/japanese-gpt-neox-3.6b-instruction-sft-v2	41.03	75.36	39.47	19.10	27.37	43.86
meta-llama/Llama-2-7b-hf	40.99	77.41	39.47	18.59	27.37	42.11
rinna/bilingual-gpt-neox-4b-instruction-ppo	40.71	78.38	31.58	20.60	27.37	45.61
rinna/bilingual-gpt-neox-4b-instruction-sft	40.31	78.23	34.21	19.35	25.89	43.86
llm-jp/llm-jp-1.3b-v1.0	39.70	75.48	36.84	19.85	24.21	42.11
At Random	30.68	49.15	25.00	16.98	25.00	37.28

and it was easy for a human to train the model to select the most appropriate prompts.

However, for the models provided by Open AI through its API, we decided to use only one standard prompt and only 0-shots for the number of shots because of the cost. The Open AI API was used with Azure; if a content filter was applied and no answer was obtained, it was determined to be incorrect.

To answer the multiple-choice questions, the likelihoods of the choices in the context were calculated and the choice with the highest likelihood was employed as the output. For GPT3.5 and GPT-4 series, the outputs with the temperature parameter set to 0 were obtained via API, and the choice that appeared earliest in the outputs was used as the output.

The results are summarized in Table 1.

5. Discussion

According to the results, the GPT-4 series exhibited a significantly high performance. Although the number of parameters in GPT-4 has not been determined, it is estimated to be more than 500 billion. Compared with other models, which have approximately 70 billion or fewer parameters, the number of parameters in GPT-4 is significantly larger, at least a few times. Considering that Qwen-72B exhibited the second-best results, the effect of the number of parameters in the models was important for achieving the highest results.

Compared to the existing Japanese leaderboard, Nejumi[5], our benchmark results for Japanese financial tasks almost correspond to the general Japanese task performance, but an exception exists. Nekomata-14b exhibits a high performance in financial tasks, which differs from that of the Nejumi leaderboard. Nekomata-14b is a tuned model of Qwen-14b that has not yet been evaluated on the Nejumi leaderboard. Moreover, the training corpora for the Qwen series were not revealed, but corpora of professional fields were included according to the official website. Therefore, the corpora used in the training of Qwen may include financial-related texts in their pre-training, and the performance of nekomata-14b is owing to this. However, models other than the nekomata, Qwen, and GPT series are already known to not include financial-related texts in their pre-training.

In the middle score of the benchmarks, around the model exhibiting an overall score of 35–40, no significant differences in their performances or the effect of the number of parameters in the models were present. We believe that this is also related to the corpora used in the training of the models. Currently, several LLMs do not learn financial documents. Therefore, in the future, the impact of financial texts on training should be evaluated, and developing models trained with financial documents is also important.

From the overall summary of the results, the benchmarks that we constructed exhibited considerable variation in difficulty from task to task, and it is possible that we were making an effective assessment. With respect to Chabsa, the highest-performing models approached the theoretical upper limit. For the design of this task, we believe that 95 is a realistic upper limit that can be achieved and is almost at this limit. However, room for further improvement in other tasks still exists, specifically regarding the performance of cpa_audit. A previous study (Masuda et al., 2023) reported that a combination of GPT-4 and retrieval-augmented generation is necessary to achieve a passing level of performance. The model's performance in solving the cpa_audit task without any external information sources can still be improved.

To investigate the effectivity of our benchmark, we analyzed the results, and the plots shown in Figures 1 – 5 were created. The relationships between the overall benchmark score and the individual scores for each task are plotted in Figures 1 – 5. Because 1/5 of the mean score is obtained from each task, a certain degree of correlation can be observed. In Figure 1, the scatter plot appears to be similar to that of $1 - \exp(x)$; therefore, fitting was performed using that function. This implies that the task tended to be easy and saturated for higher-performing models. The fitting function was found to fit well.

According to the plots, each task has its own difficulties. Chabsa is a relatively easy task and a good indicator that the difference in scores widens in lower-performing tiers. In addition, for cma_basics and security_sales_1, there is little difference in the scores of the lower-performing tiers, but the difference in the scores of the mid-performing tiers is increasing. In contrast, for the other indicators, that is, cpa_audit and fp2, observing differences in performance for both the lower and middle-performing tiers is difficult, and only some of the models exhibit overwhelmingly high performance. Because of the inclusion of these tasks with varying difficulty levels, our constructed benchmarks seem to be suitable for evaluating the Japanese financial performance of LLMs.

In future studies, we need to add more tasks, introduce more reasonable prompt-tuning methods, and determine whether a finance-specific language model can perform well.

[5] https://wandb.ai/wandb-japan/llm-leaderboard/reports/Nejumi-LLM-Neo--Vmlldzo2MTkyMTU0

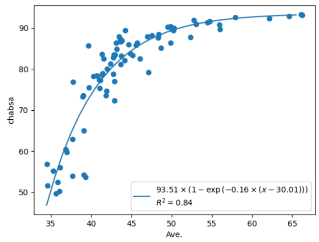


Figure 1: Relationship between Benchmark and chabsa scores

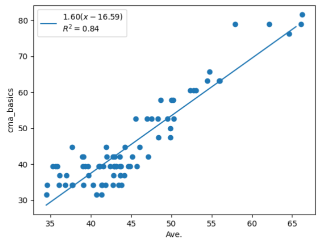


Figure 2: Relationship between Benchmark and cma_basics scores

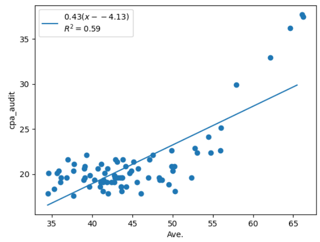


Figure 3: Relationship between Benchmark and cpa_audit scores

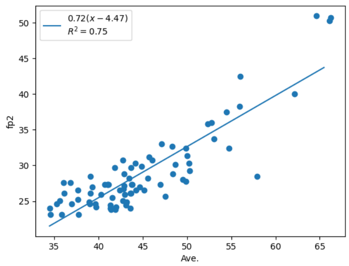


Figure 4: Relationship between Benchmark and fp2 scores

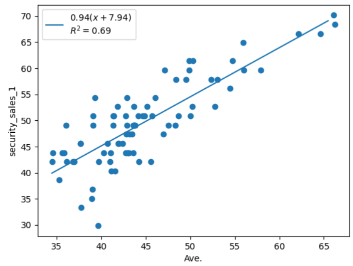


Figure 5: Relationship between Benchmark and security_sales_1 scores

6. Conclusion

In this study, we constructed a new LLM benchmark specialized for Japanese financial tasks and measured the actual benchmarks for various models. The results demonstrated that the GPT-4 series exhibited overwhelming performance. In contrast, we were also able to confirm the usefulness of our benchmark. We confirmed that our benchmark could differentiate the benchmark scores among models in all performance ranges by combining tasks with different difficulties. Future studies should also include more tasks for benchmarking to ensure a more accurate performance evaluation of LLMs.

Declarations

The author is affiliated with Preferred Networks, Inc., the developer of pfnet/plamo-13b, pfnet/plamo-13b-instruct, and pfnet/plamo-13b-instruct-nc. However, in the experiments conducted in this study, all codes were made publicly available for transparency and fair evaluation with other models.

7. Bibliographical References

Rohan Anil, Andrew M. Dai, et al. 2023. PaLM 2 Technical Report. *arXiv*. https://arxiv.org/abs/2305.10403v3.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Tom Brown, Benjamin Mann, et al. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Jonathan H. Choi, Kristin E. Hickman, Amy Monahan, and Daniel B. Schwarcz. 2023. ChatGPT Goes to Law School. *SSRN Electronic Journal*. https://papers.ssrn.com/abstract=4335905.

Aakanksha Chowdhery, Sharan Narang, et al. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv*. https://arxiv.org/abs/2204.02311v5.

Databricks. 2023. Dolly. https://github.com/databrickslabs/dolly.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the*

Association for Computational Linguistics, pages 4171–4186. Association for Computational Linguistics.

Marc Eulerich, Aida Sanatizadeh, Hamid Vakilzadeh, and David A. Wood. 2023. Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries. *SSRN Electronic Journal*. https://papers.ssrn.com/abstract=4452175.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, et al. 2021. A framework for few-shot language model evaluation. https://github.com/EleutherAI/lm-evaluation-harness.

Google. 2023. Bard. https://bard.google.com/.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Masanori HIRANO, Hiroki SAKAJI, Shoko KIMURA, Kiyoshi IZUMI, Hiroyasu MATSUSHIMA, Shintaro NAGAO, and Atsuo KATO. 2019. Related Stocks Selection with Data Collaboration Using Text Mining.

Masanori HIRANO, Masahiro SUZUKI, and Hiroki SAKAJI. 2023. llm-japanese-dataset v0: Construction of Japanese Chat Dataset for Large Language Models and its Methodology. In *The 26th International Conference on Network-Based Information Systems*, pages 442–454.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339. Association for Computational Linguistics.

Kwan Yuen Iu and Vanessa Man-Yi Wong. 2023. ChatGPT by OpenAI: The End of Litigation Lawyers? *SSRN Electronic Journal*. https://papers.ssrn.com/abstract=4339839.

LLM-jp. 2024. llm-jp-eval.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, pages 3111–3119.

Harsha Nori, Nicholas King, Scott Mayer Mckinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv*. https://arxiv.org/abs/2303.13375v2.

OpenAI. 2023a. ChatGPT. https://openai.com/blog/chatgpt/.

OpenAI. 2023b. GPT-4 Technical Report.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv*. https://arxiv.org/abs/2211.05100.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

StabilityAI. 2023. JP Language Model Evaluation Harness. https://github.com/Stability-AI/lm-evaluation-harness/tree/jp-stable.

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. 2023. JMedLoRA: Medical Domain Adaptation on Japanese Large Language

Models using Instruction-tuning. *arXiv*. https://arxiv.org/abs/2310.10083.

Masahiro Suzuki, Masanori Hirano, and Hiroki Sakaji. 2023. From Base to Conversational: Japanese Instruction Dataset and Tuning Large Language Models. *arXiv*. https://arxiv.org/abs/2309.03412.

Masahiro SUZUKI, Hiroki SAKAJI, Masanori HIRANO, and Kiyoshi IZUMI. 2022. Construction and Validation of a Pre-Training and Additional Pre-Training Financial Language Model [in Japanese]. In *The 28th meeting of Special Interest Group on Financial Informatics of Japanese Society for Artificial Intelligence*, pages 132–137.

Masahiro SUZUKI, Hiroki SAKAJI, Masanori HIRANO, and Kiyoshi IZUMI. 2023. Constructing and Analyzing Domain-Specific Language Model for Financial Text Mining.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. LLaMA: Open and Efficient Foundation Language Models. *arXiv*. https://arxiv.org/abs/2302.13971.

Hugo Touvron, Louis Martin, et al. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*. https://arxiv.org/abs/2307.09288v2.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5999–6009.

Vicuna. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. https://vicuna.lmsys.org/.

Pedram Babaei William Todt, Ramtin Babaei. 2023. Fin-LLAMA: Efficient Finetuning of Quantized LLMs for Finance. https://github.com/Bavest/fin-llama.

Shijie Wu, Ozan ˙ Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. *arXiv*. https://arxiv.org/abs/2303.17564v2.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-Source Financial Large Language Models. *arXiv*. https://arxiv.org/abs/2306.06031.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-FinGPT: Financial Sentiment Analysis by Instruction Tuning of General-Purpose Large Language Models. *arXiv*. https://arxiv.org/abs/2306.12659.

8. Language Resource References

Kubo, Takahiro and Nakayama, Hiroki and Kamura, Junya. 2018. *chABSA: Aspect Based Sentiment Analysis dataset in Japanese*. PID https://github.com/chakki-works/chABSA-dataset.

Tatsuki Masuda, Kei Nakagawa, and Takahiro Hoshino. 2023. Can chatgpt pass the jcpa exam?: Challenge for the short-answer method test on auditing. In *The 31st meeting of Special Interest Group on Financial Informatics of Japanese Society for Artificial Intelligence*, pages 81–88.

KRX-Bench: Automating Financial Benchmark Creation via Large Language Models

Guijin Son[1], **Hyunjun Jeon**[1], **Chami Hwang**[2], **Hanearl Jung**[1]
OneLineAI[1], Korea Exchange[2],
{guijn.son, hyunjun.jeon, earl}@onelineai.com,
hcharm2ing@krx.co.kr

Abstract

Inaccuracies or outdatedness of large language models (LLMs) in the finance domain may lead to misguided decisions and substantial financial losses, highlighting the importance of appropriate tools to evaluate and identify LLMs ready for production. In this work, we introduce **KRX-Bench**, an automated pipeline for creating financial benchmarks via GPT-4. To demonstrate the effectiveness of the pipeline, we create **KRX-Bench-POC**, a benchmark assessing the knowledge of LLMs in real-world companies. This dataset comprises 1,002 questions, each focusing on companies across the U.S., Japanese, and Korean stock markets. Our findings indicate that KRX-Bench can autonomously produce accurate benchmarks, achieving a minimal "false positive" rate of 1%. Notably, we find that despite leveraging GPT-4 as the generator, our pipeline can supplement enough knowledge to create questions beyond its limitations. Finally, we explore various applications of KRX-Bench, including generating open-ended, multilingual questions and reasoning benchmarks, showcasing its versatility in creating comprehensive evaluation tools for LLMs. We make our pipeline and dataset publicly available and integrate the evaluation code into EleutherAI's Language Model Evaluation Harness.

Keywords: Large Language Model, Benchmark,Finance

1. Introduction

With the advent of highly capable large language models (LLMs), the financial industry now faces pre-industrial adoption across diverse tasks (Son et al., 2023a; Callanan et al., 2023). However, key concerns surrounding the accuracy, reasoning skills, and safety of the content generated by LLMs raise diverse concerns (Wei et al., 2023; Bang et al., 2023; Alkaissi and McFarlane, 2023). While certain fields, such as arts or music, may tolerate or even embrace a degree of imaginative deviation (or "hallucination") in the outputs of LLMs, sectors like Medicine and Finance are notably intolerant of such inaccuracies. In the financial domain, hallucinations by LLMs can propagate misinformation, potentially leading to misguided investment decisions and consequent financial losses. However, existing research has predominantly focused on assessing financial LLMs' reasoning capabilities (Chen et al., 2021, 2022) or proficiency in singular tasks (Son et al., 2023b; Malo et al., 2014; Loukas et al., 2022), leaving a critical gap in understanding their comprehension of the real-world financial landscape.

To bridge this gap, we introduce **KRX-Bench**, a pipeline for the automated creation of financial benchmarks. The automated nature of KRX-Bench is ideally suited for generating a dynamic benchmark that can self-update, making it uniquely capable of capturing the rapidly changing financial sector. To demonstrate its effectiveness, we create **KRX-Bench-POC** a benchmark comprising 1,002 instances, each about companies across the U.S., Japanese, and Korean stock markets. Our assessment confirms that KRX-Bench can autonomously produce accurate benchmarks. We apply machine-learned techniques and verify that the benchmark is free of unwanted artifacts. Furthermore, a qualitative review highlights an exceptionally low "false positive" rate of 1%, indicating that human annotators deem the vast majority of questions reliable and answerable. We observe the best performing openly available LLMs (e.g., *Qwen1.5-72B*, and *Llama-2-70B*) to score below 80% suggesting room for improvement. Surprisingly, *GPT-4-Turbo* the most capable LLM available and the generator of the benchmark scores below 90% suggesting that the pipeline is capable of creating beyond the knowledge of the generator.

Finally, we demonstrate diverse applications of KRX-Bench, including creating open-ended, multilingual, and reasoning-focused benchmarks, with only minor modifications to the prompts or input documents. Our findings suggest that the pipeline can be readily adapted to generate more challenging questions simply by updating the input documents. Our contributions are twofold:

1. We present **KRX-Bench** an automated pipeline for creating financial benchmarks.

2. We introduce **KRX-Bench-POC**, to our knowledge, the first benchmark evaluating the knowledge of LLMs across multiple stock markets.[1]

[1] https://anonymous.4open.science/r/KRX-Bench-1FCE/

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 10–20
20 May, 2024.

2. Related Works

2.1. Financial Large Language Models

The financial industry has shown interest in adopting LLMs, demonstrated by the launch of BloombergGPT (Wu et al., 2023), a 50 billion parameter model specifically trained for Finance. An array of openly-available financial LLMs has followed the model, each focusing on reading comprehension (Cheng et al., 2023), financial task solving (Wang et al., 2023), or multimodality (Bhatia et al., 2024). Furthermore, multiple research have explored the possibility of LLMs to replace human analysts by either training open-source language models on tailored datasets (Son et al., 2023a) or prompting proprietary language models to solve CFA exams (Callanan et al., 2023). However, adopting LLMs in Finance faces hurdles, primarily due to their tendency to generate inaccurate information, known as *hallucinations* (Huang et al., 2023). This issue is critical in Finance, where incorrect data can lead to poor decision-making and significant financial losses. Furthermore, the risk of spreading false information through LLMs could be considered unethical or even fraudulent, slowing their integration into financial operations.

2.2. Evaluation of Financial LLMs

LLM evaluation tools have progressed from basic question-answering tasks (Rajpurkar et al., 2016) to complex reasoning (Cobbe et al., 2021) or large-scale knowledge benchmarks (Hendrycks et al., 2020; Son et al., 2024). The assessment of financial LLMs has followed a similar path, initially focusing on evaluating specific tasks (Chen et al., 2021, 2022; Loukas et al., 2022) to employing a comprehensive set of benchmarks (Xie et al., 2024; Shah et al., 2022) for a more thorough evaluation. However, the field lacks appropriate tools to accurately assess financial LLMs' grasp of the real-world financial environment, such as knowledge of company details, business objectives, and financial regulations. Furthermore, the financial market changes quickly over time—new companies emerge, and existing ones transform, quickly rendering benchmarks focused on real-life knowledge obsolete (Son et al., 2023b).

To this end, we introduce **KRX-Bench**, a pipeline for the automated generation of financial benchmarks, designed to adapt continuously to the dynamic financial market. Additionally, we provide a set of questions generated through the pipeline, which, to the best of our knowledge, is the first to evaluate LLMs across multiple stock markets and regulatory environments.

3. KRX-Bench

In this section, we elaborate on the **KRX-Bench** pipeline (3.1)) and conduct a proof of concept study leveraging the pipleine (Section 3.2).

3.1. KRX-Bench Pipeline

The **KRX-Bench** is an automated pipeline designed for generating financial benchmarks. It leverages *GPT-4-Turbo* to craft challenging questions from existing corpora, encompassing three main steps.

Question and Answer Generation In this step, we provide a document to *GPT-4-Turbo* and prompt it to generate Q&A pairs from the text. The document may be annual reports, documentation on financial lawsuits, or anything of the user's choice. While the model's cognitive capacity bounds the question generation, it can still craft questions extending beyond its pre-trained knowledge by leveraging the supplementary materials.

Creation of Distractors To reformat the Q&A pairs generated in the prior step to multiple-choice questions, we generate distractors (wrong answer choices). Simply choosing random answers as distractors could make them too easily distinguishable, so we employ *GPT-4-Turbo* to create distractors of high quality. For each question Q^*, we use the BM25 algorithm to find 10 similarly worded questions $[Q^1...Q^{10}]$ and then instruct *GPT-4-Turbo* to adapt the corresponding answers $[A^1...A^{10}]$ into plausible incorrect options for Q^*. To ensure the distractors' quality, we filter by two heuristic rules:

1. Exclude options mentioning companies irrelevant to the question.
2. Remove any answer option whose length significantly deviates from the average length of incorrect answers to maintain a uniform answer structure.

If the filtering process yields more than four distractors, we randomly select four from the remaining options.

Quality Control A critical condition for a fully automated pipeline for benchmark creation without a human in the loop is to minimize the inclusion of "false positives" or unanswerable questions. Accordingly, in this final step, we prompt *GPT-4-Turbo* to identify and eliminate unanswerable questions. For a comprehensive list of criteria used to determine unanswerability, see Figure 1 for the prompts used throughout the pipeline.

Question and Answer Generation

{CONTEXT}

Instruction: From the given the text generate an English question and answer pair. Do not ask for quantitative questions ask about the details about the company. Make sure to include the name of the company in the question. The name of the company is *{COMPANY}*.

Generate in Q: <str> A: <str> format
Question:

Creation of Distractors

You will given a question, gold answer and one irrelevant answer. Your job is to transform the irrelvant answer to a well-designed wrong answer. Replace mentions of different companies to the company of interest in the question. Alter the detail a bit to make good wrong answer.

Question: *{QUESTION}*
Gold Answer: *{GOLD}*
Irrelevant Answer: *{CANDIDATE}*
Good Wrong Answer:

Quality Control

You are the final sensitivity reader for a benchmark that is about to be published.
Read through a question included in the benchmark and evaluate whether the question is answerable.
A question is deemed unanswerable if:

1. The question does not include the name of a company.
2. The question is based on information that is outdated or no longer relevant.
3. The question assumes knowledge that is not commonly available or requires specialized expertise beyond the scope of the benchmark audience.
4. The question's phrasing is ambiguous or can lead to multiple equally valid interpretations.
5. The information needed to answer the question correctly is not present within the context provided in the benchmark.
6. The question contains biases or assumptions that could disadvantage certain groups of people or promote stereotypes.

Please ensure that each question in the benchmark meets these criteria to be considered answerable and appropriate for publication. Explain whether the question meets each criteria and return [[Yes]] for answersable questions and [[No]] for unanswerable questions.

Question: *{QUESTION}*
Gold Answer: *{GOLD}*
Wrong Answers:
1. *{WRONG ANSWER1}*
2. *{WRONG ANSWER2}*
3. *{WRONG ANSWER3}*
4. *{WRONG ANSWER4}*

Decision:

Figure 1: Prompts used throughout the **KRX-Bench** pipeline.

3.2. Proof of Concept

To demonstrate the KRX Bench pipeline's effectiveness in practice, we introduce **KRX-Bench-POC**, a benchmark dataset of 1,002 questions from companies of three nations: the United States, Japan, and Korea.

KRX-Bench-POC Initially, we compiled a dataset from annual reports across three nations: the United States, Japan, and Korea. For the U.S. (Loh) and Japan (chakki), we collect from existing resources, while Korean reports are from DART[2], a digital repository for company filings. The selection is not based on the latest fiscal data—U.S. reports are from 2022, and Japan's from 2018. This is because this section aims to showcase the capability of the pipeline rather than currently creating up-to-date benchmarks. We plan to release updated versions of the benchmark in the future. To ensure consistency, we randomly chose 500 annual reports each from Japan and Korea. For details on

[2]https://dart.fss.or.kr/main.do

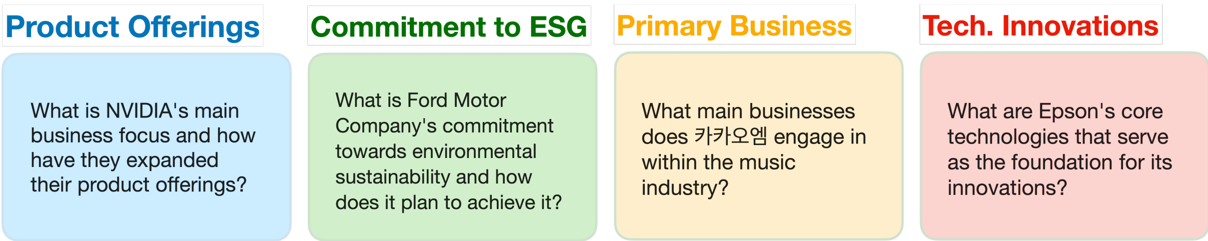


Figure 2: Selected samples of questions included in the **KRX-Bench-POC**.

the dataset composition, see Table 1.

Following this step, we execute the KRX Bench pipeline on the collected annual reports and generate multiple-choice questions. Following a quality filtering process, we retain a total of 1003 questions: 373 for the US, 319 for Korea, and 311 for Japan.

Country	# of Doc	Av. Length	Fiscal Year
United States	494	55479	2022
Korea	2896	5158	2023
Japan	3718	1339	2018

Table 1: Statistics on the collected annual reports.

Diversity We evaluate the diversity of KRX-Bench-POC, by randomly selecting 99 questions, 33 from each nation, and categorizing each by topic. This survey reveals that the pipeline yields a broad spectrum of 15 distinct categories, including Business Goals, Product Offerings, Financial Policy, and Business Strategy, with no single category predominating. Primary Business emerged as the most represented category. For a detailed breakdown of each category and sample questions, refer to Table 2 and Figure 2, respectively.

Category	Frequency
Primary Business	33
Business Goals	14
Company Mission	10
Industry	10
Business Operations	8
Product Offerings	7
ETC	17

Table 2: A survey on the category of generated questions. The ETC category includes the following: Financial Policy, Innovation, Business Strategy Commitment to ESG, Long Term Strategy, Mid Term Strategy, Global Strategy, Company History.

Quality In this section, we analyze the quality of the generated dataset. First, we test whether the benchmark includes potentially exploitable artifacts(e.g., shortcuts or patterns) that LLMs might abuse to solve the question. Specifically, we test two machine-learned features: (1) Similarity-Based Feature: We evaluate if the option most similar to the question, using BM25 and BERT[3] for similarity measurements, is likely to be correct; (2) Zero-Shot Classifier Feature: We employ a zero-shot classifier, trained on natural language inference tasks, to determine if it can accurately solve the questions without specific training (Laurer et al., 2023). Table 3 presents a performance comparison between the machine-learned features on our KRX-Bench-POC and Hellaswag (Zellers et al., 2019), a widely adopted benchmark for commonsense reasoning. Similarity-based measures on KRX-Bench-POC outperform random guessing but achieve similar or lower scores than their performance on Hellaswag. This indicates that KRX-Bench-POC maintains a comparable level of challenge and avoids introducing excessive artifacts compared to established benchmarks.

Feature	KRX-Bench-POC	Hellaswag
Random Baseline	20.0%	25.0%
Similarity (BM-25)	37.3%	**54.1%**
Similarity (BERT)	**39.8%**	32.2%
Zero-Shot Classifier	20.4%	**25.1%**

Table 3: Accuracy of machine-learned models on the **KRX-Bench-POC** and *Hellaswag*.

Furthermore, we assess the quality control step introduced in Section 3.1 through qualitative analysis, examining both answerable and unanswerable questions classified by *GPT-4-Turbo*. Two of the authors review 200 randomly sampled questions (100 deemed answerable and 100 deemed unanswerable by *GPT-4-Turbo*) without prior knowledge of *GPT-4*'s judgments. Results, shown in figure 2, reveal a remarkably low "false positive" rate of only 1%, indicating that very few unanswerable questions were incorrectly labeled as answerable. Although achieving a 0% "false positive" rate would be ideal, even human-curated datasets struggle to meet this standard. The observed 1% rate is

[3]We use *all-MiniLM-L6-v2* from the Sentence Transformers library (Reimers and Gurevych, 2019).

sufficiently low for reliable evaluation. Additionally, the "true negative" rate of 43% highlights the effectiveness of our pipeline's quality control in mirroring human judgment, ensuring the pipeline's benchmark generation abilities.

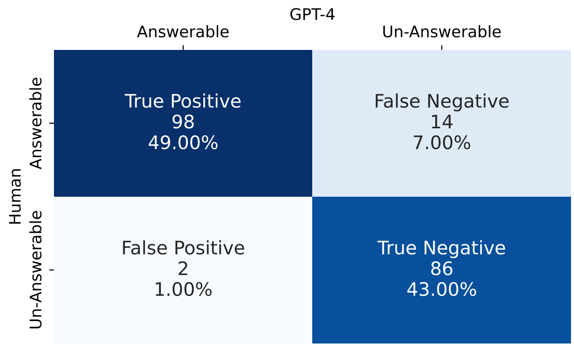


Figure 3: A confusion matrix comparing the decision of human annotators against the quality control step by *GPT-4-Turbo*.

4. Experimental Setup

In this section, we explain our experimental setup for evaluating different LLMs on the **KRX-Bench-POC**.

4.1. Models

In this work we evaluate 12 different LLMS ranging in different size for evaluation. The evaluated models include: (1) *Llama-2* (7B, 13B, 70B)) (Touvron et al., 2023) (2) *Qwen1.5* (0.5B, 1.8B, 4B, 7B, 14B, 72B) (Team) and (3) *GPT-3.5-Turbo* and *GPT-4-Turbo* (OpenAI, 2023). We also evaluate *Japanese-StableLM-Base-Beta-7B* (Lee et al.) and *Llama-2-KOEN-7B* (Junbum, 2023), which are variations of *Llama-2* each continually pre-trained on Japanese and Korean correspondingly.

4.2. Evaluation Methods

For evaluation, we prompt a model to generate the most plausible option via greedy decoding. All models are evaluated in full precision in a 3-shot setting on 8 X A100 80GB GPUs. See Figure 4 for the prompt used in our evaluation. For reproducibility, the evaluation codes used in our research are implemented via LM-Eval-Harness (Gao et al., 2023).

5. Results on KRX-Bench-POC

Model Size and Performance Table 4 presents the evaluation results for various models on the KRX-Bench-POC. Larger models consistently outperform smaller ones, indicating a linear scaling

Direct Evaluation

Question: *{QUESTION}*
Options:
A. {OPTION A}
B. {OPTION B}
C. {OPTION C}
D. {OPTION D}
E. {OPTION E}
Answer:

Figure 4: Prompt used in our Direct Evaluation.

trend. This pattern holds for both *Qwen1.5* and *Llama-2* model families, demonstrating that our benchmark aligns with typical benchmark behaviors. Notably, the top models, *Qwen1.5-72B*, and *Llama-2-70B* achieve scores under 80%, indicating room for improvement. This suggests that our pipeline successfully generates challenging benchmarks for state-of-the-art open models without any human supervision.

Regional Bias In Figure 5, we notice a regional bias in model performance; despite all questions being in English, models perform better on questions about U.S. companies than those about Japanese or Korean companies. This trend is consistent across all models, with leading models like *Qwen1.5-72B* and *Llama-2-70B* scoring around 90% for U.S. companies but only about 70% for Japanese and Korean companies. This pattern is also evident in proprietary models such as *GPT-3.5-Turbo* and *GPT-4-Turbo*. Several factors could contribute to this disparity, including the scarcity of English resources on Japanese and Korean companies, which limits the models' ability to acquire knowledge about these companies during pretraining. This implies that leveraging more difficult documents as input, internal documents, for example, could easily elevate the benchmark's difficulty.

Surprisingly, models specifically trained on additional Japanese and Korean data, such as *Japanese-StableLM-Base-Beta-7B* and *Llama-2-KOEN-7B*, show decreased performance across all subsets. Despite being trained on an extra 100B tokens of Japanese and 60B tokens of Korean, these models do not improve scores for questions related to their targeted nations; instead, their overall scores drop. This unexpected outcome may be attributed to two main reasons. Firstly, the added web-crawled tokens might not provide sufficient information about the companies featured in the benchmark. Secondly, further pretraining on dedicated national data could induce catastrophic forgetting, weakening the models' English language problem-solving abilities. This observation chal-

Models	N=3			
	US	KO	JR	Total
Pre-Trained Models				
Qwen1.5-0.5B	20.38	17.87	18.06	18.77
Qwen1.5-1.8B	39.68	24.14	20.97	28.26
Qwen1.5-4B	58.45	31.35	30.65	40.15
Qwen1.5-7B	81.77	47.34	48.06	59.06
Qwen1.5-14B	87.13	57.68	60.65	68.49
Qwen1.5-72B	87.40	72.10	72.58	77.36
Llama-2-7B	42.09	20.38	23.23	28.56
Llama-2-13B	85.52	52.98	51.94	63.48
Llama-2-70B	93.30	71.16	73.23	79.23
Continual Pretrained Models				
Japanese-StableLM-Base-Beta-7B	32.98	21.00	23.87	25.95
Llama-2-KOEN-7B	17.16	19.44	18.06	18.22
Proprietary Models				
GPT-3.5-Turbo	87.13	63.32	66.13	72.19
GPT-4-Turbo	95.44	84.33	84.84	88.20

Table 4: Average accuracy(%) calculated using the Direct method in a 3-shot setting across the entire test set. We report the macro-average accuracy across subjects within each category.

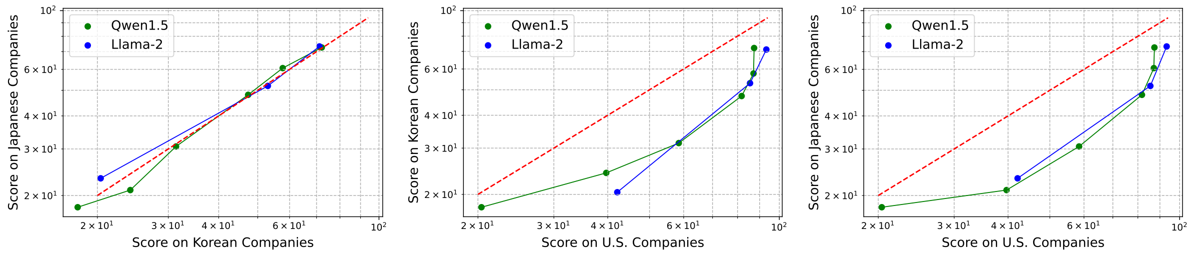


Figure 5: Comparison of Qwen1.5 and Llama-2 models based on the nationality of the companies.

lenges the assumption that language-transferred models can effectively substitute dedicated language models, revealing they may lack comprehensive knowledge.

Can GPT-4 Ace Itself? The KRX-Bench pipeline employs *GPT-4-Turbo* throughout its stages, leading to the following question: Can this pipeline effectively evaluate *GPT-4-Turbo* itself? To address this, we also assess *GPT-4-Turbo*'s performance on the benchmark. Although *GPT-4-Turbo* achieves the highest scores among tested models, it does not excel across all subsets, particularly for the Korean and Japanese questions. This outcome suggests that using supplementary materials for dataset generation enables the generator to formulate questions that extend beyond its knowledge cutoff. Consequently, the pipeline demonstrates its capacity to create benchmarks capable of challenging the model itself, including *GPT-4-Turbo*, thereby validating its utility in generating effective evaluation tools.

6. Practitioner's Guide for Implementing KRX-Bench

In this section, we extend beyond **KRX-Bench-POC** to show different use-cases of the **KRX-Bench** pipeline in generating financial Benchmarks.

6.1. Open-Ended Generation

While the KRX-Bench pipeline primarily generates multiple-choice questions, it can also assess the generative capabilities of language models by simply providing a question and prompting an LLM to generate an answer. This approach offers a more nuanced evaluation of a model's generation ability. To illustrate this, we sample 70 questions from the KRX-Bench-POC dataset to create **KRX-Bench-Gen**. The questions span various categories: Primary Business, Industry, Product Offerings, Business Strategy, Technology, Business Goals, Financial Policy, Commitment to ESG, and Risk. Primary

Question

What significant actions did HOYA Corporation take in 2017 concerning the medical devices sector?

Gold Answer

In 2017, HOYA Corporation established a joint venture company with China's Aohua for a medical flexible endoscope business and acquired ReadSpeaker, a leading company in cloud-based text-to-speech services. They also set up a new production base in Thailand for intraocular lenses used in cataract surgery, and acquired Performance Optics, LLC along with its subsidiaries VISION EASE and Daemyung Optical.

GPT-3.5-Turbo

In 2017, HOYA Corporation made a significant move in the medical devices sector by acquiring the medical unit of the US-based company, Performance Optics. This acquisition allowed HOYA to expand its presence in the ophthalmic lens market and strengthen its position in the medical devices industry. The deal included Performance Optics' manufacturing facilities …

Score:3

GPT-4-Turbo

In 2017, HOYA Corporation, a global technology company based in Tokyo, Japan, known for its innovations in optics and medical devices, took a significant step in the medical devices sector by acquiring Performance Optics, LLC, including its subsidiaries VISION EASE, a global ophthalmic lens marketer and manufacturer, and Daemyung Optical Co. Ltd., a leading optical lens manufacturer in South Korea…

Score:4

Figure 6: Example generations of *GPT-3.5-Turbo* and *GPT-4-Turbo* on open-ended questions.

Business is the largest category with 11 questions, while Risk is the smallest with four, averaging 7.8 questions per category.

Given that pre-trained models without further tuning might struggle with open-ended questions, we focuses on GPT variants. We assess *GPT-3.5-Turbo* and *GPT-4-Turbo*, employing an "LLM-as-a-Judge" approach based on Zheng et al. (2024)'s implementation. This judge model is prompted to rate answers on a scale from 1 to 5. In Table 5 we observe *GPT-4-Turbo* to score slightly higher than *GPT-3.5-Turbo*.

Models	Open-Ended Generation
GPT-3.5-Turbo	3.28
GPT-4-Turbo	3.55

Table 5: Evaluation results of GPT variants on open-ended questions.

To provide additional insight, we include an example of a question and the generated responses in Figure 6. In this example, both models accurately identify HOYA Corporation's acquisition of Performance Optics, yet *GPT-4-Turbo* provides a more detailed response by noting the inclusion of subsidiaries in the acquisition. This illustrates how our benchmark can be utilized to assess both generative capabilities and knowledge depth. The accuracy of evaluations could be further improved by employing more knowledgeable LLM judges with expertise in finance or by incorporating human evaluators.

6.2. Multilinguality

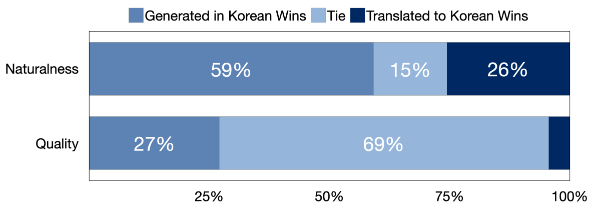


Figure 7: Win rate analysis between generating in Korean and translating a material generated in English to Korean.

OpenAI (2023) reports *GPT-4-Turbo* to have robust multilingual capabilities. Accordingly, we explore where the identical benchmark generation pipeline can be applied to generate benchmarks in languages other than English, specifically Korean. We adapt the pipeline by incorporating "Generate in Korean" into our prompts, generating 250 questions in Korean. We conduct a comparative quality analysis to assess the effectiveness of generating questions directly in Korean versus translating questions from English. We randomly select 250 questions from the KRX-Bench-POC dataset and hire two annotators for evaluation. Presented with pairs of questions—one generated in Korean and the other translated—they are tasked to identify the question that sounds more natural to native Korean speakers and is of higher quality, without knowledge of the questions' generated methodology. The annotators have the option to choose one of the options or declare a tie. Figure 7 indicates that annotators consistently find the directly generated Korean samples more natural for native speakers. We suspect that direct generation allows *GPT-4-Turbo* to leverage its in-context learning abilities to learn from the provided Korean document, thereby commanding better Korean than the translation approach. Quality-wise, annotators considered both methods to yield questions of similar quality 69% of the time, but in 29% of cases, the directly generated samples were preferred. These results demonstrate that our pipeline can be seamlessly adapted to produce high-quality multilingual benchmarks with minimal adjustments.

In Table 6, we report the evaluation results for the subset generated in Korean. Interestingly, unlike our previous experiments *Llama-2-KOEN-7B* outperforms *Llama-2-7B*. We attribute this improvement primarily to the language advantage. Unlike the assessments reported in Table 4, which involved questions about Korea in English, this experiment presented questions in the Korean language. This context likely favored *Llama-2-KOEN-7B*, benefiting from its targeted continual pretraining in Korean.

6.3. Beyond Knowledge Benchmarks

저축은행의 특수관계자가 설정한 사모펀드에 저축은행이 투자하는 것이 가능한가요?

(Can a savings bank invest in a private equity fund established by an affiliated party?)

금융회사가 암호화된 개인신용정보를 전송하기 위해 제3자의 통신회선 서비스를 이용하려고 할 때, 이는 신용정보법 제17조의 개인신용정보 처리 위탁에 해당하나요?

(Does the transmission of encrypted personal information via a third-party communication service qualify as outsourcing under Article 17 of the Credit Information Act?)

Figure 8: Examples of the generated reasoning benchmark. English translations are added for broader accessibility.

This section explores whether the KRX-Bench pipeline can be leveraged to create reasoning benchmarks. Previously, we introduced KRX-Bench-POC to showcase the pipeline's ability to generate benchmarks evaluating LLMs' knowledge of real-world companies. Alongside such knowledge benchmarks, reasoning benchmarks are crucial for a comprehensive assessment of LLMs, focusing on their capacity to apply knowledge logically to solve problems. For this purpose, we compile a set of Korean documents related to financial lawsuits and process them through the same pipeline, producing 100 questions that challenge LLMs to conduct legal reasoning on financial disputes. We choose to generate questions in Korean to preserve the intricate details crucial in legal contexts, concerned that translation might compromise these subtleties. We present an example of the generated question in Figure 8.

Table 6 presents the evaluation results for the reasoning subset, where *Llama-2-KOEN-7B* continues to outperform *Llama-2-7B*. Notably, *GPT-4-Turbo* achieves a near-perfect score on the reasoning subset. This performance could stem from various factors. Firstly, the lawsuit collection used for this subset, sourced from the internet and dating back to the 1980s, may have been part of GPT variants' pretraining data. Secondly, LLMs might struggle to generate challenging distractor options that surpass their reasoning capabilities. While supplying reference materials enables the generation of questions beyond the model's knowledge, our current pipeline might fail to guide models to create sufficiently complex distractors effectively. Future research is required to better understand these dynamics. However, despite these considerations, the benchmarks still provide a rigorous test for evaluating the capabilities of leading open LLMs.

Models	Multilingual (Kor)	Reasoning
Llama-2-KOEN-7B	38.8	58.0
Llama-2-7B	34.8	24.0
Llama-2-13B	50.4	48.0
Llama-2-70B	63.2	81.0
GPT-3.5-Turbo	58.4	92.0
GPT-4-Turbo	84.8	96.0

Table 6: Evaluation results of selected models on subsets generated in Section (6.2) and Section (6.3).

7. Conclusion

In this study, we introduce **KRX-Bench**, an automated pipeline designed for generating financial benchmarks. We validate the pipeline's effectiveness and reliability by developing **KRX-Bench-POC**, at the best of our knowledge, the first dataset aimed at evaluating LLMs' understanding of companies across diverse stock markets. Our findings confirm that the proposed pipeline can autonomously produce trustworthy benchmarks. This feature suits the fast-changing dynamics of the financial sector, enabling the generation of benchmarks that evolve in tandem with market changes. Additionally, we illustrate its broad applicability through various use cases, including creating open-ended, multilingual, and reasoning-based questions, highlighting our method's versatile utility.

8. Acknowledgements

This work was done in colaboration with OneLineAI[4] and Korea Exchange(KRX)[5].

9. Bibliographical References

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of gpt-4 level multimodal financial large language models. *arXiv preprint arXiv:2402.10986*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ethan Callanan, Amarachi Mbakwe, Antony Papadimitriou, Yulong Pei, Mathieu Sibue, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Can gpt models be financial analysts? an evaluation of chatgpt and gpt-4 on mock cfa exams. *arXiv preprint arXiv:2310.08678*.

chakki. Coarij: Corpus of annual reports in japan. `https://github.com/chakki-works/CoARiJ`. Accessed: 2024-02-27.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

[4] `https://www.onelineai.com`

[5] `https://global.krx.co.kr/main/main.jsp`

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A Smith, Iz Beltagy, et al. 2023. Camels in a changing climate: Enhancing lm adaptation with tulu 2. *arXiv preprint arXiv:2311.10702*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

L. Junbum. 2023. llama-2-ko-7b (revision 4a9993e).

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Sungho Park, et al. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *arXiv preprint arXiv:2306.02254*.

Rik Koncel-Kedziorski, Michael Krumdick, Viet Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building efficient universal classifiers with natural language inference. *arXiv preprint arXiv:2312.17543*.

Meng Lee, Fujiki Nakamura, Makoto Shing, Paul McCann, Takuya Akiba, and Naoki Orii. Japanese stablelm base beta 7b.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.

Jerry Loh. sp500-edgar-10k. `https://huggingface.co/datasets/jlohding/sp500-edgar-10k`. Accessed: 2024-02-27.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. Finer: Financial numeric entity recognition for xbrl tagging. *arXiv preprint arXiv:2203.06482*.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Tomohiro Sawada, Daniel Paleka, Alexander Havrilla, Pranav Tadepalli, Paula Vidas, Alexander Kranias, John J Nay, Kshitij Gupta, and Aran Komatsuzaki. 2023. Arb: Advanced reasoning benchmark for large language models. *arXiv preprint arXiv:2307.13692*.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

Guijin Son, Hanearl Jung, Moonjeong Hahm, Keonju Na, and Sol Jin. 2023a. Beyond classification: Financial reasoning in state-of-the-art language models. *arXiv preprint arXiv:2305.01505*.

Guijin Son, Hanwool Lee, Nahyeon Kang, and Moonjeong Hahm. 2023b. Removing non-stationary knowledge from pre-trained language models for entity-level sentiment classification in finance. *arXiv preprint arXiv:2301.03136*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multi-task language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Guijin Son, Hanwool Lee, Suwan Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2023c. Hae-rae bench: Evaluation of korean knowledge in language models. *arXiv preprint arXiv:2309.02706*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

Qwen Team. Introducing qwen1.5. https://qwenlm.github.io/blog/qwen1.5/. Accessed: 2024-02-27.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

BLU-SynTra: Distinguish Synergies and Trade-offs Between Sustainable Development Goals Using Small Language Models

Loris Bergeron, Jérôme François, Radu State, Jean Hilger
Banque de Luxembourg, University of Luxembourg - SnT SEDAN & Finnovation Hub
14 Bd. Royal L-2449 Luxembourg, 29 Av. John F. Kennedy L-1855 Luxembourg
loris.bergeron@bdl.lu, {jerome.francois, radu.state, jean.hilger}@uni.lu

Abstract

Since the United Nations defined the Sustainable Development Goals, studies have shown that these goals are interlinked in different ways. The concept of SDG interlinkages refers to the complex network of interactions existing within and between the SDGs themselves. These interactions are referred to as synergies and trade-offs. Synergies represent positive interactions where the progress of one SDG contributes positively to the progress of another. On the other hand, trade-offs are negative interactions where the progress of one SDG has a negative impact on another. However, evaluating such interlinkages is a complex task, not only because of the multidimensional nature of SDGs, but also because it is highly exposed to personal interpretation bias and technical limitations. Recent studies are mainly based on expert judgements, literature reviews, sentiment or data analysis. To remedy these limitations we propose the use of Small Language Models in addition of an advanced Retrieval Augmented Generation to distinguish synergies and trade-offs between SDGs. In order to validate our results, we have drawn on the study carried out by the European Commission's Joint Research Centre which provides a database of interlinkages labelled according to the presence of synergies or trade-offs.

Keywords: United Nations (UN), Sustainable Development Goals (SDGs), Small Language Models (SLMs), Retrieval Augmented Generation (RAG), Mistral, Orca 2, Phi-2, Generative Query Reformulation (GenQR), Context Aware Query Rewriting (CAR), Reciprocal Rank Fusion (RRF), Zero-Shot Classification

1. Introduction

In 2015, the agenda dedicated to sustainable development was adopted by all 193 member states of the United Nations (UN)(United Nations and Development, 2015). A set of 17 Sustainable Development Goals (SDGs) was defined and reported in Table 1. The establishment of these 17 SDGs, broken down into 169 targets and 232 indicators, would have us isolate all these elements as if, in theory, no interlinkages were possible between the economic, social and governance dimensions. As an example, assuming *SDG3 Good Health and Well-being* and *SDG12 Responsible Consumption and Production*, there is no clear assessment if these SDGs present synergies or trade-offs. In other words, would having a positive impact on *SDG3* also mean having a positive impact on *SDG12* and vice versa? At a first glance, having a positive impact on *SDG12* seems to have a positive impact on the health and well-being of populations. However, if we improve *SDG12* on responsible consumption and production, carbon footprint can go down also. In that case, this might lead to a trade-off with *SDG13 Climate Action* and with *SDG7 Affordable and Clean Energy*. Obviously, this reasoning is based on personal beliefs that are unique to each individual and is therefore, by definition, subject to personal bias. The complexity of these interlinkages is all the more true if we opt for a finer granularity by appealing to the SDGs targets. In this case we have a combination of potential 14196 interlinkages. It is all the more essential to obtain an overview of these interlinkages to give policy-makers all the transparency to make the right decisions to successfully implement these objectives. Understanding the range of positive and negative interlinkages among the SDGs is the key to unlocking their full potential while ensuring that progress in some dimensions does not have a negative impact on others(noa, 2017). Hence, this paper introduces a method capable to automatically distinguish synergies and trade-offs in the interlinkages of SDGs using Small Language Models (SLMs) thanks to their cognitive capacities. In particular, we are interested in reproducing results established by experts in scope of a research(European Commission. Joint Research Centre., 2023) which is part of *KnowSDGs*[1] and carried out by the European Commission's Joint Research Centre (JRC). The database provided in this study brings together a number of interlinkages at goals and targets levels. For many months now, the research on Large Language Models (LLMs) has continued to progress. Transformers architecture(Vaswani et al., 2017) were considered to be LLMs regardless of the number of training parameters included in them. We used them mainly for their cognitive capacities but also and above all for their vast knowledge since they were trained on

[1] https://knowsdgs.jrc.ec.europa.eu

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 21–33
20 May, 2024.

impressive volumes of data. However, since the research carried out by Microsoft(Eldan and Li, 2023), a distinction can be made between LLMs and SLMs. We can therefore consider as an SLM an LLM with a far smaller number of parameters, several hundred billion against a few billion. SLMs are not used for their knowledge, but rather for their impressive cognitive capacities given their small size. Recent advances in Generative AI (GenAI) have opened up new possibilities in the field of SLMs which are now used in a multitude types of tasks. In this paper, we promote the use of SLMs to replicate the results obtained in JRC's study. The obtained results shows their ability to distinguish synergies and trade-offs between SDGs targets. This type of usage can be industrialised, but is also made close through with the help of a relevant context, since such an analysis must be carried out given a specific environment (e.g. political, economic, geographical)(Le Blanc, 2015). Our contribution to scientific research in relation to these SDG themes can be broken down into four areas:

- An innovative methodology, based on the use of SLMs and an advanced RAG (Retrieval Augmented Generation) (Lewis et al., 2021) workflow, to distinguish synergies and trade-offs between SDGs targets in a set of documents
- An open architecture that can be replicated by research teams or companies while still having access to infrastructure with limited computing and hardware power and hosted internally for governance reasons
- An implementation of the aforementioned architecture using Mistral 7b (Mistral)(Jiang et al., 2023), Orca 2 7b (Orca)(Mitra et al., 2023), Phi-2 2.7b (Phi)(Javaheripi et al., 2023)
- A comparative analysis of our results based on the study carried out by the European Commission(European Commission. Joint Research Centre., 2023)

The structure of the paper is organized as follows: Section 2 provides an overview of related work. Our method, called BLU-SynTra is described in Section 3. Then, details of the validation set used to confirm our results are presented in Section 4, comparative analysis and the results we achieved are presented in Section 5. Lastly, Section 6 provides concluding remarks on the conducted research and suggests potential enhancements for future research.

2. Related Work

In 2015, when SDGs were conceptualized by the UN, research topics related to the identification of connections between the SDGs began to appear(Le Blanc, 2015). The identification of connections between SDG targets is carried out on the basis of a manual semantic analysis by determining that if two targets refer to the same global concept, they can be assumed to be interlinked. Obviously, this method is highly exposed to fluctuations in human interpretation. In 2017, the International Council for Science (ICSU)(noa, 2017) published a report to explore the nature of interlinkages between SDGs. The evaluation method is based on assigning manually a score to quantify the interlinkages. Therefore, this evaluation is based on expert opinion and a review of the literature. At European level, in 2019, the European Commission's Joint Research Centre (JRC) published a first version of a research(European Commission. Joint Research Centre., 2019) highlighting interlinkages in order to ensure policy coherence in relation to the SDGs, based on a literature review. Hereafter, more and more related research has been carried out (Bali Swain and Ranganathan, 2021; Fariña García et al., 2021; Dawes, 2022; Dawes et al., 2022; Song and Jang, 2023). Use of new methods, like analysis methods based on correlation networks or semantic analysis networks, to determine interlinkages between SDGs are being used. These research does not attempt to distinguish, from a qualitative point of view, the interlinkages type when they are actually present. In 2023, the JRC published a new research(European Commission. Joint Research Centre., 2023) to review the progress of work on the existence of synergies or trade-offs in interlinkages between SDGs in different contexts. Based on this work, a database of interlinkages is established through a literature review. This database provides the community with

SDG	Description
SDG1	No Poverty
SDG2	Zero Hunger
SDG3	Good Health and Well-being
SDG4	Quality Education
SDG5	Gender Equality
SDG6	Clean Water and Sanitation
SDG7	Affordable and Clean Energy
SDG8	Decent Work and Economic Growth
SDG9	Industry, Innovation and Infrastructure
SDG10	Reduced Inequalities
SDG11	Sustainable Cities and Communities
SDG12	Responsible Consumption and Prod.
SDG13	Climate Action
SDG14	Life Below Water
SDG15	Life on Land
SDG16	Peace, Justice and Strong Institutions
SDG17	Partnerships for the Goals

Table 1: The 17 Sustainable Development Goals

a list of 18780 interlinkages, each qualified as a synergy or a trade-off alongside the method used to assert it. As highlighted previously, past work mainly relies on experts judgments, literature review or data analysis methods to explore SDG interlinkages. As the current state of the art(Issa et al., 2024) does not refer to a methodology based on SLMs to qualitatively distinguish the type of interlinkages, our research aims to explore the potential benefit of such models.

3. Method

3.1. BLU-SynTra overview

The overall process of BLU-SynTra consists of adaptation and combination of different methods and practices as represented in Figure 1. They are also detailed in the following sub-sections. The first building block in the figure is *Optimised data indexing*. This block takes as input a set of documents related to various studies or reports, where interlinkages (i.e. synergies or trade-offs) between SDGs are explained and validated by experts. These documents are then handled by an unstructured data ingestion mechanism to extract the information they contain. Then, a series of processing steps create chunks, using a parent-child strategy and static thresholds to divide up the information. These chunks are then summarised using a SLM (Mistral, Orca and Phi) to retrieve their meaning by reducing their context size. Chunks were later incorporated into a vector database using the best performing model at the time of our research to perform Semantic and Textual Similarity (STS) operations (on the basis of information established by the Massive Text Embedding Benchmark (MTEB)[2]).

The next building block is *Information Retrieval* whose aim is to contextualize a user query given as input about interlinkages based on the previously indexed document using RAG (Retrieval Augmented Generation). Advanced RAG methods(Gao et al., 2024) like Generative Query Reformulation (GenQR)(Wang et al., 2023b), Context Aware Query Rewriting (CAR)(Anand et al., 2023), Reciprocal Rank Fusion (RRF)(Cormack et al., 2009) have been used to maximise the final results of our research. Finally, *Distinguish interlinkages* relies on the extracted context to apply common sense reasoning and understanding of language capabilities offered by SLMs to distinguish synergies and trade-offs in interlinkages available in our validation set using Zero-Shot (ZS)(Brown et al., 2020) classification.

[2]https://hf.co/spaces/mteb/leaderboard

3.2. Optimised data indexing

3.2.1. Ingest unstructured data

When processing unstructured data, the challenge is to extract the contained elements as faithfully as possible to avoid any analysis errors.

Let's define D as the set of documents used in our research:

$$D = \{d_1, d_2, \dots, d_n\}$$

Within each d_i, we assume the elements e:

$$d_i = \{e^i_1, e^i_2, \dots, e^i_{m_i}\}$$

Where i is the index of the document within the set D and m_i is the number of textual elements in document d_i.

To achieve this, we used the well known *unstructured* [3] library that includes an OCR model to segment a document and extract its content. At this stage of the process, our aim was not to extract any information that has already been chunked or organised, but only to extract $e^i_{m_i}$ as represented in the original d_i document excluding images and tables. This results in a set of elements of different element types *(e.g. title, page_break, footer, etc.)*. To focus only on information having semantic value, only *NarrativeText* typed elements are kept. They consist of text composed of at least two sentences. Assuming $narrative(d_i) \subset d_i$ is only the remaining *NarrativeText* elements of d_i, we refine D as D':

$$D' = \{narrative(d_i)\}, d_i \in D$$

3.2.2. Chunking elements in parent-child

Once narrative text is extracted, it is essential for our solution to conserve the related context and meanings. We have decomposed each element e^i_j using a parent-child strategy in which the elements can be made up of several parents p and several smaller children c implemented in the *RecursiveCharacterTextSplitter* function from *LangChain*[4]. This text splitter is suggested for general text and it uses a list of default separators (i.e. \n\n, \n, space, char), aiming to maintain paragraphs, then sentences, and finally words together as much as possible since they are viewed as the most semantically connected elements. The maximal *chunk_size* parameters for parents and children have been set to 4096 and 2048 respectively based on the maximum window context size of the SLMs we used. We can therefore establish that each element e^i_j can be represent as the set of children: $child_{e^i_j} = \bigcup_{p \in P_{i,j}} \{p_k\}$ with $P_{i,j}$ is the set

[3]https://github.com/unstructured-io/unstructured
[4]https://python.langchain.com

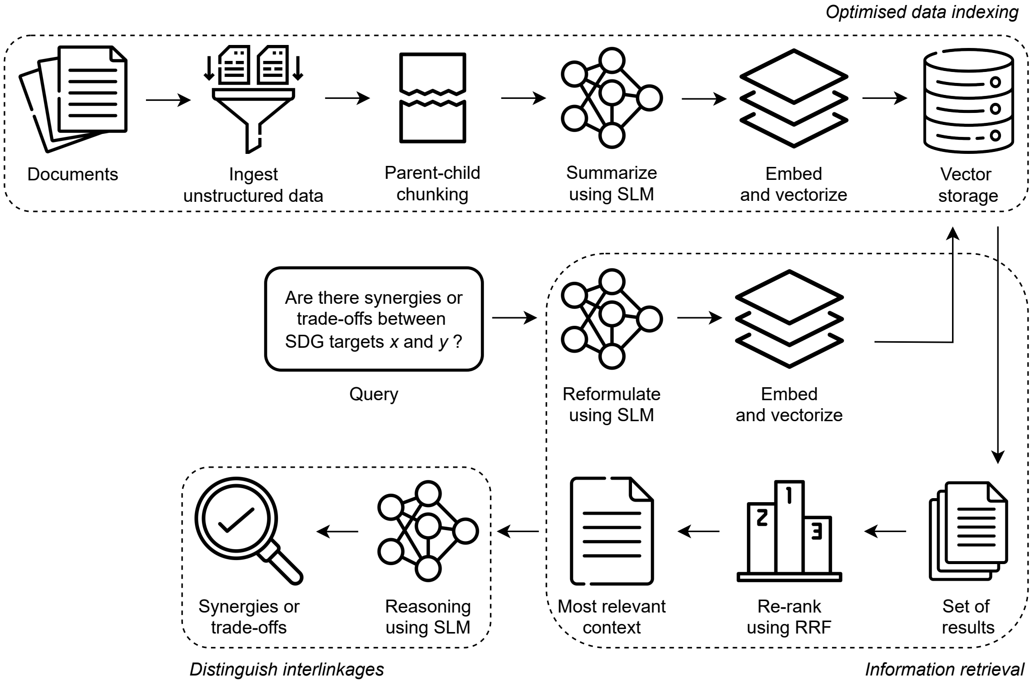


Figure 1: BLU-SynTra overview

of parents derived from e_j^i and p_k is the k-th child of the parent p.

As a result, we define the whole set of children to represent the original set of documents D as:

$$C = \bigcup_{d \in D', e_j^i \in d} child_{e_j^i}$$

3.2.3. Summarize chunks

Advanced RAG methods promotes the principle of summarization to improve the ability of LLMs to understand key information, particularly when dealing with extensive texts(Gao et al., 2024). BLU-SynTra thus includes such processing as well.

Creating informative summaries based on a longer text is a quite complex, unlike summarising smaller texts, which justifies our previous breakdown in section 3.2.2.

Each $c \in C$ is summarized using the different SLMs. sm_c^a is the generated summary for c using the SLM $a \in SLM$ with $SLM = \{Mistral, Orca, Phi\}$ resulting in the full summarization of all documents:

$$Sm_c = \bigcup_{c \in C} sm_s^a(c), a \in \{SLM\}$$

At the end of the generation process we have a set of 15009 summaries (5003 for each SLM).

The SLMs are conditioned to produce summaries as faithful and consistent as possible with our various $c \in C$ in order to minimise errors in the following way:

> *Please provide a summary of the following text. Ensure the summary is clear, coherent, and faithful to the content of the original text.*
> *Text:* $< c >$

For a clearer understanding of how this summary stage works, we have appended an example A of a randomly selected child c and the corresponding summary Sm_c produced. To remain as neutral as possible, we did not modify the parameters within the SLMs *(e.g. temperature, top_p)* and used the same prompt for each of them.

3.2.4. Embeddings creation and storage

Once summaries are created, they are stored as embeddings to enable easy comparison between them. BLU-SynTra relies on the Universal AnglE Embedding(Li and Li, 2023) in Large-V1 version(UAE-Large-V1)[5] as the embedding model. At the time of this research, this is the most advanced model to perform STS operations in English. Given the technical specificities of the model, chunks are embedded in 1024-long vector. All vectors are stored in chroma[6]. Default use of the Hierarchical Navigable Small World (HNSW)(Malkov and Yashunin, 2020) method in chroma, coupled with the use of the cosine function to perform similarity operations allows us to retrieve the appropriate information. In chroma's *documents* field, we have

[5] https://hf.co/whereisai/uae-large-v1
[6] https://docs.trychroma.com

stored all the Sm_c summaries along side the child c used to create them, their relative parent p and the source document d_i as metadata. Each Sm_c is so associated with a vector representation noted v_{Sm_c}:

$$\mathbf{v}_{Sm_c} = \begin{pmatrix} v_{Sm_c1} \\ v_{Sm_c2} \\ \vdots \\ v_{Sm_c1023} \\ v_{Sm_c1024} \end{pmatrix}$$

3.3. Information retrieval

3.3.1. Query reformulation

To retrieve information, we generate a query. Assuming an initial query q, the objective is to determine if there are synergies or trade-offs between SDG targets, as for instance:

> q: *Are there synergies or trade-offs between SDG targets 17.11 and 10.7?*

As the formulation of a query to a generative AI model can have a significant impact on the final classification result, the principle of reformulation is widespread in Information Retrieval (IR) problems and is used to counter problems linked to a more or less extensive vocabulary. To optimise our results, we used existing reformulation mechanisms(Wang et al., 2023b; Anand et al., 2023). On the one hand, a reformulation noted as $GenQ$ is solely based on the cognitive capacities of SLMs. On the other hand we also define $GenQCAR$ as a reformulation based on particular context related to the SDG targets helping SLMs in their reformulation task. While $GenQ$ reformulation simply reformulates and expands q, the $GenQCAR$ approach enriches knowledge by providing it with the definitions of synergy and trade-off as defined in the JRC study as well as the definitions of the two targets as defined by the UN[7]. In the case of $GenQCAR$, the additional information made available to the SLMs is transmitted to it when q is reformulated using a prompt B specifically written for this purpose. In the appendix C, two examples of $GenQ$ and $GenQCAR$ are given and have been derived using q mentioned earlier. A higher vocabulary richness can be observed in the case of $GenQCAR$, but also and above all the use and understanding of the terms synergy and trade-off in accordance with the definitions given by the JRC.

P denotes the reformulation process and can therefore define for each q the process:

$$P(q) = \{GenQ(q), GenQCAR(q)\}$$

[7]https://unstats.un.org/sdgs/metadata

During the IR step, we obtained a set of results for which we retrieve the 10 most similar items by query, defined as follows:

$$R_{total}(q) = R_q \cup R_{GenQ(q)} \cup R_{GenQCAR(q)}$$

3.3.2. Re-rank result sets

When retrieving information from R_{total}, the result is a set of elements associated with cosine similarity scores in R_i. In order to identify the most recurrent documents in R_{total}, we used the Reciprocal Rank Fusion (RRF)(Cormack et al., 2009) method. In contrast to individual ranking methods, the authors have shown that the RRF method is capable of consistently obtaining better results than the standard Condorcet Fuse method(Montague and Aslam, 2002). RRF weights each document in R_i with the inverse of its position on the rank. It thus gives preference to documents at the *top* of the rank and penalizes documents below the *top* of the rank. In addition, this approach is unsupervised, that is also a significant advantage to be applicable. RRF therefore sort our set R_{total} according to a scoring formula based on a set of rankings R_i:

$$\text{RRFscore}(r \in R_{total}) = \sum_{r \in R_i} \frac{1}{k + r(d)} \quad (1)$$

with $r(d)$ the rank of document d and k a parameter, set to $k = 60$ as suggested in the original paper of RRF (Cormack et al., 2009).

Finally, the document with the highest score is considered to be the most appropriate given the queries formulated in the previous step. Thanks to what we have seen in section 3.2.4, this enables us to identify the associated p and refer to as the *most relevant context* to be used by the SLM to carry out the classification step.

3.4. Distinguish interlinkages

To be able to distinguish the interlinkages type between the two targets concerned, we decided to use a Zero-Shot (ZS)(Brown et al., 2020) classification. We used ZS by augmenting the knowledge of the model with the definitions of *synergy* and *trade-off* as defined in the JRC study and also with the *most relevant context* retrieved in the previous step. Thanks to context augmentation, the knowledge of SLMs is increases and enables performing reasoning tasks and thus determine, in the given context, the type of interlinkages present between the two targets. The fact that we add in our prompt some more detailed background information (i.e. most relevant context) as well as the definitions of the two classes to be classified (i.e. synergy and trade-off) improves the accuracy of the predictions(Wang et al., 2023c). We have define the prompt detailed in appendix D to carry out this operation. In addition to the classification, we request a justified explanation for the underlying reasoning behind it. Such kind of explanation could be made available to a decision-maker to obtain all the transparency needed to understand these interlinkages. To illustrate this process, an example of output is given in appendix E resulting from the classification between targets 6.a and 10.b of our validation set.

We could have implemented methods like Zero-Shot Chain of Thought (ZS-CoT)(Kojima et al., 2023) or Clue And Reasoning Prompting (CARP)(Sun et al., 2023). These methods, used to classify texts using LLMs, compensates for the models' lack of reasoning capacity by adopting a progressive reasoning strategy to overcome these limitations in complex environments. However, the related literature highlights that the added value of such methods, based on progressive reasoning, is correlated with the size of the model used. In our case, by the limited size of the number of parameters in our SLMs, the added value in terms of reasoning is not significant and would negatively increase classification processing time.

4. Experimental setup

4.1. Selected SLMs

We chose three SLMs which considered to be the most common from the state-of-the-art at the time of our research. We also selected multiple models for comparative analysis rather than pre-selecting one. However, they are used independently and can be interchanged. In other words, BLU-SynTra used in production would rely on the use of the SLM that exhibits the most efficient summarisation behavior, as discussed in section5:

- Mistral 7b(Jiang et al., 2023) - Designed to use Grouped-Query Attention (GQA)(Ainslie et al., 2023) and Sliding Window Attention (SWA)(Beltagy et al., 2020)(Child et al., 2019). The use of GQA and SWA allows us to significantly accelerate the inference speed while reducing the memory required for the decoding phase. This choice is particularly well suited to infrastructures with limited computing power.

- Orca 2 7b(Mitra et al., 2023) - Based on the architecture of Llama-2(Touvron et al., 2023). This version 2 of Orca has the advantage of employing a varied number of reasoning techniques (e.g. step-by-step, recall then generate, recall-reason-generate, etc.) while being able to choose the right method for a given task.

- Phi-2 2.7b(Javaheripi et al., 2023) - Builds on the work of the previous version, Phi1.5(Li et al., 2023). This version currently shows similar or better cognitive performance than models with 13b parameters or less. While its parameter size is more than half that of the two previous SLMs, its main innovation lies in the use of *textbook-quality* data(Gunasekar et al., 2023) and the addition of new synthetic data. This new version uses an innovative method of knowledge transfer to accelerate its training speed while delivering superior benchmark scores compared to the previous version.

4.2. Validation set

As stated earlier, we rely on the database provided by the JRC serves. Since this database is the result of work carried out by multiple JRC experts to avoid individual bias. We thus consider this database enough accurate to serve as a validation set. At the SDG target level, there is a total of 10614 interlinkages but only 5715 are unique. There are 80.5% synergies, 10% trade-offs and 9.5% not specified resulting in a significant imbalance between classes. For the remainder of our research, only interlinkages specifically associated *synergy* or *trade-off* type are kept. In addition, we excluded interlinkages whose *clear_direction* variable was set to *no*. By applying these quality filters we obtain a set of 4682 interlinkages, of which 2956 are unique, divided into 4172 (89.1%) synergies and 510 (10.9%) trade-offs. In order to optimise our experiment, we randomly sampled this group to keep only 10% of the total. This brings our total number of classes to 468, divided into 419 (89.53%) synergy classes and 49 (10.47%) trade-off classes. Regarding to the methods of analysis used to establish the distinctions between synergy and trade-off in the database, no filter has been applied resulting in the breakdown shown in Table 2. To compare and replicate our results, we have made our final validation set available online[8].

We looked at the distribution of classes according to the targets selected in our validation set. For sake of clarity, targets are grouped by SDG they relate to. In Figure 2, the distribution of synergies and trade-offs is presented.

5. Results

5.0.1. Evaluation of summaries

This first experiment aims at assessing the quality of the summarization process which is critical for the IR process. ROUGE(Lin, 2004) metric might have been used to evaluate the quality of generated summaries. This metric measures the similarity between a summary sm_c^a in comparison to the refer-

[8] https://github.com/lrsbrgrn/blu-syntra

N	Method of analysis	Synergy	Trade-off
1	Data Analysis	8	2
2	Expert judgement	133	9
3	Literature review	48	4
4	Mixed (Expert judgement & Data analysis)	8	1
5	Mixed (Literature review & Data analysis)	25	2
6	Mixed (Literature review, Expert judgement & Data analysis)	3	0
7	Mixed (Literature review, Expert judgement & Modelling)	1	4
8	Mixed (Literature review & Expert judgement)	174	21
9	Mixed (Semantic analysis, Literature review & Expert judgement)	15	6
10	not_specified	4	0
Total		419	49

Table 2: Distribution of classes by analysis method

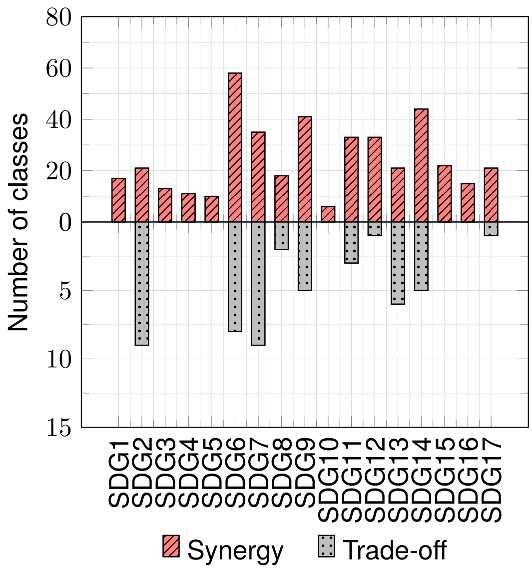


Figure 2: Distribution of classes by SDG

Metrics	Mistral	Orca	Phi
Relevance	**4.6**	**4.6**	3.9
Coherence	**4.5**	4.4	4.0
Consistency	**4.9**	4.8	4.1
Fluency	**3.0**	2.9	2.8

Table 3: Evaluation using G-Eval

ence noted as c. We could have carried out our evaluation using ROUGE-N (1 and 2) to measure the proportion of common unigrams and bi-grams. In addition, ROUGE-L metric would have enabled us to evaluate the longest sequence of words shared between our summary and its reference. A long shared sequence indicates a definite similarity between the two. However, metrics such as ROUGE, although widely used in NLP tasks, show low correlations with human judgements(Wang et al., 2023a). Based on this assessment of the state of the art we used the G-Eval(Liu et al., 2023) framework. The latter relies on an LLM, in our case GPT-4(OpenAI et al., 2023), as an evaluator to determine several metrics (i.e. relevance, coherence, consistency and fluency). Table 3 shows the different results obtained by G-Eval. We can observe very similar results between Mistral and Orca but also excellent results for Phi given its very small size. However, Mistral is superior to Orca in 3 out of the 4 metrics, and to Phi in all cases. With Mistral, we observed an average length of 222 words for each c compared with an average length of 83 words for the summaries produced. This is equivalent to dividing the size of the text by almost 3 and thus justifies the use of an advanced RAG method to reduce the text to retain only the key information.

5.0.2. Validity of classifications

This experiment evaluates to which extent our approach can automatically infer if synergies or trade-offs exist between SDG goals. For this analysis phase, only *Mistral* is used due to its highest scores on summarization as evaluated in the previous section. As a first experiment, we were interested in assessing the validity of our classifications and the underlying behaviour of the SLM according to a binary classification where the positive label $pos\_label = SYNERGY$. The results show a very good capability of BLU-SynTra to identify the synergies with $F1\_score = 0.88$, $Precision = 0.92$ and $Recall = 0.84$. However, a deeper look at the confusion matrix in the Table 4 highlights a bias in overestimating these synergies and, as a an opposite effect, a notable difficulty in identifying trade-offs. Of the 49 trade-offs available in our validation set, only 20 (40.82%) were actually correctly identified. Our validation set shows a strong asymmetry in the classes it contains, since synergy and trade-off represent 89.53% and 10.47% of the whole respectively. This result still highlights the difficulties encountered by Mistral in producing classifications for which the finesse of the language, the subtlety of the words and the intonations present challenges to their reasoning function.

		Predicted	
		Synergy	Trade-off
Actual	Synergy	354	65
	Trade-off	29	20

Table 4: Confusion matrix

Secondly, results were differentiated according to the SDG each target they relate to. We found significantly heterogeneous performance metrics for SDGs 2, 7 and 13, with F1 scores of $F1_{SDG2} = 0.70$, $F1_{SDG7} = 0.89$ and $F1_{SDG13} = 0.79$ respectively. SDGs 2 and 7, as shown in Figure 2, are among the largest contributors to trade-offs. For SDG 2, only 3 of the 9 trade-offs in our validation set were correctly classified as such. Regarding SDG 7, only 2 of the 9. Notably, SDG 6 is the third highest contributor of trade-offs in our validation set but still has an $F1_{SDG6} = 0.92$ with 5 of the 8 trade-offs correctly identified. This generally highlights a high divergence of BLU-SynTra capabilities among SDG.

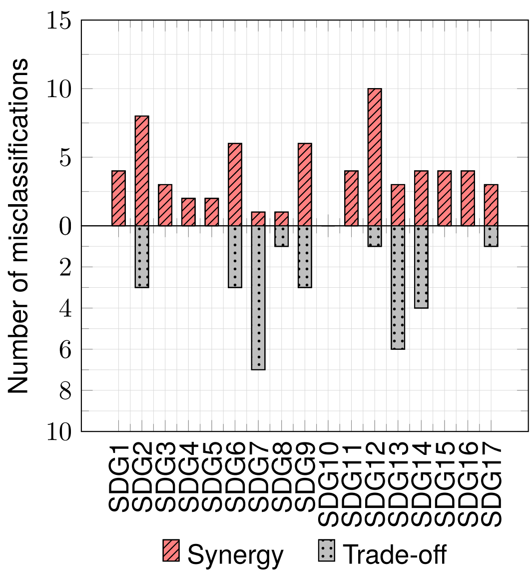


Figure 3: Misclassification by SDG

The last experiment further investigate results obtained according to the analysis methods used by the JRC to determine the presence of synergy or trade-off (see Table 2). As can be observed in Table 5, there is no consistency between the results obtained. In general, a significant deterioration in F1 scores for the *Mixed (Expert Judgement & Data analysis)* (M-EJDA) and *Data Analysis* (DA) methods can be observed. These two approaches lead to the worst performance. However, although the *Expert Judgement* (EJ) analysis method is the second method with the most number of classes, the F1 score obtained is the highest. We have also observed that the *Literature Review* (LR) method is well approximated by BLU-SynTra with all the trade-offs correctly identified.

Methods	Precision	Recall	F1-score
M-EJDA	0.80	0.50	0.62
DA	0.71	0.63	0.67
EJ	0.97	**0.92**	**0.94**
LR	**1.00**	0.73	0.84

Table 5: Metrics by analysis method

This raises questions about mixed approaches compared with single approaches (i.e. using only one analysis method). We have noted *Mixed* the analysis methods employing several sub-methods, and noted *Single* the methods employing only one analysis method. In Table 2, the *Mixed* methods are identified by the prefix *(Mixed)*, the others are consequently attached to the *Single* category. We therefore observed a slight superiority when comparing *Single* and *Mixed* approaches (see Table 6). However, the *Single* approaches were able to correctly identify 60.00% of the trade-offs, unlike the *Mixed* approaches, which were only able to obtain a score of 32.35% and therefore leads to a deterioration at the global level of the classification metrics.

Methods	Precision	Recall	F1-score
Single	**0.97**	**0.86**	**0.91**
Mixed	0.89	0.83	0.86

Table 6: *Single* and *Mixed* metrics

6. Conclusion

In this paper, we have proposed a complete solution entitled BLU-SynTra relying on SLMs to identify synergies and trade-offs between SDG targets. We have shown that traditional ZS text classification methods, enhanced by a context and definitions retrieved using several advanced RAG concepts, can make it easy to identify synergies and justify to decision-makers the underlying reasoning behind this distinction in a given environment. However, the identification of trade-offs lacks precision, and most of all with high variability according to the considered analysis method or SDG. Linguistic complexity and subtle vocabulary make it difficult for SLMs to identify trade-offs and distinguish them from synergies. Despite this, this first research work aims to open up new possibilities for using SLMs to carry out this interlinkages classification task as we have experienced and, more generally, in tasks requiring complex reasoning to be carried out in infrastructures with limited hardware resources or at lower cost that LLMs. New advances in summary generation (Zhang et al., 2023) will be the subject of future improvement to create summaries in an iterative way in order to reduce as much as possible the errors and hallucinations induced by SLMs. In terms of reasoning skills, our plan is to leverage ReAct(Yao et al., 2023) in order to compare the results obtained with those obtained in this research. ReAct seem to indicate better performance than standard approaches, even for models with very small parameter sizes.

7. Acknowledgment

We would like to thank the sustainable finance experts at Banque de Luxembourg from the Private Banking Investments team, as well as the sustainable finance experts at Banque de Luxembourg Investments for their involvement and advice in carrying out this research. We would also like to thank the teams at the Joint Research Centre (JRC) who made themselves available to answer our questions when we needed them most.

8. Bibliographical References

2017. A guide to SDG interactions: from science to implementation. Technical report, International Council for Science (ICSU).

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. ArXiv:2305.13245 [cs].

Abhijit Anand, Venktesh V, Vinay Setty, and Avishek Anand. 2023. Context Aware Query Rewriting for Text Rankers using LLM. ArXiv:2308.16753 [cs].

Ranjula Bali Swain and Shyam Ranganathan. 2021. Modeling interlinkages between sustainable development goals using network analysis. *World Development*, 138:105136. .

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165 [cs].

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. ArXiv:1904.10509 [cs, stat].

Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 758–759, Boston MA USA. ACM.

J.H.P. Dawes. 2022. SDG interlinkage networks: Analysis, robustness, sensitivities, and hierarchies. *World Development*, 149:105693. .

Jonathan H. P. Dawes, Xin Zhou, and Mustafa Moinuddin. 2022. System-level consequences of synergies and trade-offs between SDGs: quantitative analysis of interlinkage networks at country level. *Sustainability Science*, 17(4):1435–1457. .

Ronen Eldan and Yuanzhi Li. 2023. TinyStories: How Small Can Language Models Be and Still Speak Coherent English? ArXiv:2305.07759 [cs].

European Commission. Joint Research Centre. 2019. *Interlinkages and policy coherence for the sustainable development goals implementation: an operational method to identify trade offs and co benefits in a systemic way.* Publications Office, LU.

European Commission. Joint Research Centre. 2023. *Uncovering SDG Interlinkages: interconnection at the core of the 2030 Agenda : an analysis of the state of the art on SDG Interlinkages and an update of the JRC tool to foster policy coherence for sustainable development in EU policymaking.* Publications Office, LU. .

María Consuelo Fariña García, Víctor Luis De Nicolás De Nicolás, José Luis Yagüe Blanco, and Jesús Labrador Fernández. 2021. Semantic network analysis of sustainable development goals to quantitatively measure their interactions. *Environmental Development*, 37:100589. .

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. ArXiv:2312.10997 [cs].

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. ArXiv:2306.11644 [cs].

Lea Issa, Toufic Mezher, and Mutasem El Fadel. 2024. Can network analysis ascertain SDGs interlinkages towards evidence-based policy planning? A systematic critical assessment. *Environmental Impact Assessment Review*, 104:107295.

Mojan Javaheripi, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, and Caio Mendes. 2023. The Surprising Power of Small Language Models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. .

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. ArXiv:2205.11916 [cs].

David Le Blanc. 2015. Towards Integration at Last? The Sustainable Development Goals as a Network of Targets. *Sustainable Development*, 23(3):176–187.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. ArXiv:2005.11401 [cs].

Xianming Li and Jing Li. 2023. AnglE-optimized Text Embeddings. ArXiv:2309.12871 [cs].

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks Are All You Need II: phi-1.5 technical report. ArXiv:2309.05463 [cs].

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. ArXiv:2303.16634 [cs].

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, Edouard Grave, Yann LeCun, and Thomas Scialom. 2023. Augmented Language Models: a Survey. ArXiv:2302.07842 [cs].

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching Small Language Models How to Reason. ArXiv:2311.11045 [cs].

Mark Montague and Javed A. Aslam. 2002. Condorcet fusion for improved retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 538–548, McLean Virginia USA. ACM.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua

Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs. ArXiv:2312.05934 [cs].

Jaemin Song and Chang-Ho Jang. 2023. Unpacking the sustainable development goals (SDGs) interlinkages: A semantic network analysis of the SDGs targets. *Sustainable Development*, 31(4):2784–2796.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text Classification via Large Language Models. ArXiv:2305.08377 [cs].

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Shreya Pathak, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher Choquette, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clement Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma. Publisher: Kaggle.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Department of Economic United Nations and Social Affairs-Sustainable Development. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. ISBN: A/RES/70/1 Pages: 16301 Type: General Assembly.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. ArXiv:2303.04048 [cs].

Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. 2023b. Generative Query Reformulation for Effective Adhoc Search. ArXiv:2308.00415 [cs].

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023c. Large Language Models Are Zero-Shot Text Classifiers. ArXiv:2312.01044 [cs].

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. ArXiv:2109.01652 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. ArXiv:2210.03629 [cs].

Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. SummIt: Iterative Text Summarization via ChatGPT. ArXiv:2305.14835 [cs].

A. Example of a generated summary

c: in terms of meeting most of the SDGs by 2030. Fast population growth slow progress toward sustainable production and consumption and the negative impacts of the ongoing coronavirus disease 2019 COVID19 pandemic on socioeconomic development make it difficult for China to meet the SDGs. Therefore effective water pollution control in China is urgently needed. Here we analyze the interactions

Sm_c: China faces challenges in achieving most of the Sustainable Development Goals (SDGs) by 2030 due to a fast-growing population, slow progress towards sustainable production and consumption, and the negative impacts of COVID-19 on socio-economic development. Amidst these challenges, effective water pollution control in China is particularly urgent. This analysis explores the interactions between these issues.

B. Prompt to reformulate queries

$GenQ$: Rephrase and expand the user's query as best as you can in one short sentence.
Query: $<$ query $>$

$GenQCAR$: Using the definitions below, rephrase and expand the user's query as best as you can in one short sentence.
Synergy: Positive interactions where progress of an SDG promotes, contributes, or enables progress on another one.
Trade-off: Negative interactions where progress of an SDG deteriorates progress

towards another one.
Target < target_source >:
< target_source_definition >
Target < target_destination >:
< target_destination_definition >

C. Examples of query reformulation

q: *Are there synergies or trade-offs between SDG targets 17.11 and 10.7 ?*

$GenQ$: *What are the interrelations and potential benefits or costs of advancing targets 10.7 on sustainable and inclusive fiscal policies for development and 17.11 on effective public financial management for achieving the Sustainable Development Goals?*

$GenQCAR$: *How do progress on SDG target 17.11, which aims to significantly increase exports from developing countries, and target 10.7, which focuses on facilitating orderly migration, interact? Do they positively reinforce each other (synergy) or does progress on one target have a negative impact on the other (trade-off)?*

D. Augmented prompt to distinguish interlinkages

You are a text classifier. Your task is to analyse context intent and classify the context into SYNERGY or TRADE-OFF based on definitions below.
SYNERGY: Positive interactions where progress of an SDG promotes or enables progress on another one.
TRADE-OFF: Negative interactions where progress of an SDG deteriorates progress towards another one.
Analyze the given context and indicate whether it better illustrates SYNERGY or TRADE-OFF, justifying your response and explain your reasoning.
CONTEXT: < most_relevant_context >

E. Classification and underlying reasoning

$Mistral$: *Based on the context provided, it appears that SDG Water Sanitation and the target in question have a significant number of direct interlinkages, meaning that progress on this target contributes positively to the progress of SDG Water Sanitation. Additionally, the target in question is the direct driver for change for SDG Water Sanitation, as stated by "Total Direct Driver linkages deriving from Total Direct Response linkages going to 34." Furthermore, the target in question is indirectly linked with several other targets, including SDG 11, SDG 14 and SDG 15. While there may be some negative interactions or trade-offs between these targets and the target in question, the overall relationship between SDG Water Sanitation and the target in question is one of synergy. Therefore, based on the information provided, the context can be classified as SYNERGY.*

Assessing the Impact of ESG-Related News on Stock Trading in the Indonesian Market: A Text Similarity Framework Approach

Okiriza Wibisono[1], **Ali Akbar Septiandri**[2], **Reinhard Denis Najogie**[1]
[1]Independent Researcher, [2]University College London
okiriza.w@gmail.com, ali.septiandri.21@ucl.ac.uk, najomath@gmail.com

Abstract

Environmental, Social, and Governance (ESG) perspectives have become integral to corporate decision-making and investment, with global regulatory mandates for ESG disclosure. The reliability of ESG ratings, crucial for assessing corporate sustainability practices, is compromised by inconsistencies and discrepancies across and within rating agencies, casting doubt on their effectiveness in reflecting true ESG performance and impact on firm valuations. While there have been studies using ESG-related news articles to measure their effect on stock trading, none have studied the Indonesian stock market. To address this gap, we developed a text similarity framework to identify ESG-related news articles based on Sustainability Accounting Standards Board (SASB) Standards without the need for manual annotations. Using news articles from one of the prominent business media outlets in Indonesia and an event study method, we found that 17.9% out of 18,431 environment-related news are followed by increased stock trading on the firms mentioned in the news, compared to 16.0% on random-dates datasets of the same size and firm composition. This approach is intended as a simpler alternative to building an ESG-specific news labeling model or using third-party data providers, although further analyses may be required to evaluate its robustness.

Keywords: ESG, sentence embeddings, stock trading

1. Introduction

Environmental, social, and governance (ESG) perspectives have become one of the most prominent factors in corporate decisions and investment (Edmans and Kacperczyk, 2022). Many regulators worldwide have mandated public firms to disclose their ESG practices regularly. Firms are likewise concerned about the perception of their ESG practices as research has shown that it can drive firm value, e.g. through its impact on sales and market investment on firms. The number and size of investment funds which participate in the UN's Principles for Responsible Investment (PRI) also have expanded greatly, reaching more than $112 trillion in assets under management by 2023 (Principles for Responsible Investment, 2023).

Despite the increased focus on ESG, the reliability of ESG ratings, the primary source of ESG information, remains under examination. Serious inconsistencies are documented relating to differences in data sources and methodologies across rating agencies, and also internally (i.e. across time within a rating agency) (Berg and Rigobon, 2019; Johnson, 2023; Temple-West, 2023). Such issues have raised concerns about the reliability of ESG ratings in capturing firms' sustainable practices and their subsequent impact on valuations.

As an alternative, recent research suggests that monitoring ESG-related news in the media can offer valuable insights into the impact of ESG practices on firms' market prices. However, identifying ESG-related news that is relevant and also potentially market-moving for a firm requires expertise in ESG topics as well as time-consuming manual effort. Previous studies addressing this challenge have relied on third-party data providers, which often use automatic classifications based on machine learning to label news articles as positive or negative with respect to a firm's ESG performance.

In this study, we aim to quantify the effect of ESG-related news articles on the Indonesian market by using a text similarity framework and causal analysis, thus eliminating the need for manually annotating the articles. In doing so, we made two main contributions:

1. We collected a large dataset of 119k news articles pertinent to financial and economic issues from January 2016 to July 2023 with the corresponding stock ticker symbols to measure the effect of these articles on the Indonesian stock market (§3). Using Sentence-BERT (SBERT) fine-tuned on Indonesian text and the SASB Standards, we identified 11,920 environment-related articles that served as the basis for our analysis.

2. We found that 17.9% out of 18,431 news articles (adjusted by the number of firms mentioned per publication date, referred to as "firm-dates") are followed by increased stock trading on the firms mentioned in the news for five trading days after the observation period. This is higher than the 16.0% increase observed from a set of dummy datasets (§4). Furthermore, the percentage of the excess trading rose to 25.6% when we extended the number of trading days post-period to 20.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 34–39
20 May, 2024.

Building on our findings, we explore the implications of leveraging news articles to assess the impact of ESG factors on firms listed in the Indonesia Stock Exchange (§5). Additionally, we propose four avenues for further research, emphasizing enhancements in model robustness, inclusion of social and governance aspects, analysis of news sentiment, and the exploration of varied datasets.

2. Related Work

2.1. Using ESG news to predict stock market

Several research have examined the significance of ESG publication and ESG news on firms' value. Capelle-Blancard and Petit (2017) used 126 thousand ESG news, covering 100 listed firms from Dow Jones Sector Titans indexes, to study the impact on stock price cumulative average abnormal returns (CAAR). The dataset was obtained from a third-party provider, including the positive/negative labels on the ESG news. They found, among other hypotheses, that negative ESG news is associated with a price decline of about 0.1%, while positive ESG news does not lead to a gain in value.

Similar to (Capelle-Blancard and Petit, 2017), Serafeim and Yoon (2022) further studied the kinds of ESG news that result in market reaction using a third-party news dataset with positive/negative labels, comprising more than 100 thousand firm-date observations for 3,109 firms. However, they found that excess market-adjusted returns are more substantial for positive ESG news, for news related to social factors (as opposed to environmental or governance factors), and if there is more news on the same day. Pertinent to our research, the reaction is significant only when the news discusses financial material issues for a given industry – according to the standards published by the Sustainability Accounting Standards Board (SASB) (IFRS Foundation, 2023).

Our study is the first to utilize ESG news for predicting trends in the Indonesian stock market. While we identified studies confirming the impact of ESG factors on the Indonesian stock market, such as those by Nareswari et al. (2023) and Lubis and Rokhim (2021), none of these previous studies have employed Indonesian news articles as a measure of these factors.

2.2. Identifying ESG factors from news articles

One way to identify ESG factors from news articles involves training a classifier based on the articles' content. Previous studies have successfully applied this approach using articles not only in English but also in French and Chinese (Chen et al., 2023; Tseng et al., 2023; Pontes et al., 2023; Lee et al., 2023; Wang et al., 2023; Billert and Conrad, 2023; Mashkin and Chersoni, 2023; Glenn et al., 2023). Another way involves extracting ESG-related terms, as demonstrated by Sandwidi and Pallitharammal Mukkolakal (2022). However, both approaches typically require costly human expert annotations.

To bypass the need for manual annotation, one could use Sentence-BERT (SBERT) (Reimers and Gurevych, 2019) to generate embeddings for the news articles and ESG standards such as SASB, comparing the similarity values between the two sets of embeddings. This method was employed by Sen et al. (2023) to analyze the internal sustainability efforts of major US companies by comparing text from online reviews with the United Nations (UN) Sustainable Development Goals (SDGs). A more directly relevant example to our study is the work by Pontes et al. (2023), which also used SBERT, but in their case, to train a support vector machine classifier using the embeddings as inputs. Our approach differs by directly comparing the embeddings from the news articles with the SASB standards and filtering out irrelevant matches, as described in the subsequent section.

3. Methodology

3.1. Data

News We collected Indonesian news articles from Kontan.co.id[1], one of Indonesia's largest news portals specializing in financial news. In total, we collected 119k articles, spanning January 2016 to July 2023. These articles often contain stock ticker symbols to identify firms listed in the Indonesia Stock Exchange (IDX).

Stock trading We used IDX trading data in the same timespan as the news articles, i.e. January 2016 to July 2023. The data we used is the daily trading volume for each firm in IDX. In addition, we also used trading data for each sector in IDX.

SASB Standards To anchor the text embeddings to (globally-recognized) ESG concepts, we utilized sustainability standards published by the Sustainability Accounting Standards Board (SASB) (IFRS Foundation, 2023), as previously done in the literature (Taleb et al., 2020; Consolandi et al., 2022). In total, there are 77 industry standards grouped into 11 sectors; e.g. Coal Operations industry standard in the Extractives & Minerals Processing sector.

[1] https://www.kontan.co.id

Each industry standard is associated with several ESG topics[2]. We selected only environmental topics as the focus of our study as suggested by Sandwidi and Pallitharammal Mukkolakal (2022). For each environmental topic in each industry standard, we took the Topic Summary section, which describes relevant points about that topic for this industry. There are in total 244 industry-topic combinations. Since we were using an Indonesian news dataset, we translated the topic summaries into Indonesian with Google Translate API.

3.2. Identifying relevant news articles

We employed IndoSBERT (Diana and Khodra, 2023), an SBERT-based (Reimers and Gurevych, 2019) model fine-tuned on Indonesian texts, to generate embeddings of the news articles and the SASB Standards. SBERT is a framework derived from BERT (Devlin et al., 2019) designed to generate fixed-size sentence embeddings that can be compared using cosine similarity (Reimers and Gurevych, 2019). SBERT uses the siamese network architecture (Chicco, 2021) and a pooling operation on BERT's output, facilitating the fixed-size embeddings generation. This approach enables the efficient comparison of sentence pairs of varying lengths (Reimers and Gurevych, 2019).

For each article, we identified the SASB environmental topic standard with the highest article-standard similarity score (Equation 1) to represent the relevance of article a to standard s:

$$\alpha_a = \max_s sim(v_a, v_s) \tag{1}$$

where $sim(v_a, v_s)$ is the cosine similarity between the embeddings of article v_a and the embeddings of standard v_s.

We then selected the 90$^{\text{th}}$ percentile as the threshold for identifying relevant articles to ensure robustness against false positives, as suggested in a previous study (Chaturvedi et al., 2023). Given that each article was assigned a similarity value in the preceding step, this filtering is needed as the standard deemed most similar to a specific article might still be unrelated to it. Consequently, we ended up with 11,920 news articles focused on environmental topics, with an average of 1,490 articles annually (SD = 353).

3.3. Extracting firm mentions

We collected a comprehensive list of firms in the Indonesian stock market from IDX website[3]. This list, which contains 888 firms as of September 2023, serves as our reference for identifying relevant firms within news articles. We used both ticker symbols and full official firm names to search for firm mentions. We then extracted the firm mentions using a regular expression.

Additionally, we excluded irrelevant mentions by filtering lines containing phrases like "*Baca juga:*" ("Also read:"), "*Menarik dibaca:*" ("Interesting to read:"), and "*Selanjutnya:*" ("Further reading:") within the articles. These phrases are typically links to other news articles, thus extracted firms mentioned in these lines are irrelevant.

3.4. Event study

To validate the usefulness of the text embeddings in identifying environment-related news articles, we employed a causal analysis of the impact of the filtered articles on stock trading using an event study. Specifically, we tested whether a firm's stock trading volume increases for several days after environment-related news about that firm is published.

In our event study framework, we considered the publication date of a news article as the pivotal event from which we began observing changes in stock trading volumes. Given that an article may mention multiple firms, we associated each article with pairs of firms and publication dates, referred to as "firm-dates". Each firm-date then represented a time series comprising stock trading volume and several control variables, allowing us to analyze periods before (pre-periods) and after (post-periods) the event. This methodology yielded a dataset of 18,431 firm-dates (18,431 time series), using days as the unit of measurement.

For each firm-date, we used the standardized stock trading volume (setting the mean to 1) from 66 trading days before the article publication date (pre-periods), which reflects approximately 3 months or 22 trading days per month. We then evaluated the effect of the environment-related news on the stock trading volume after 5, 10, and 20 trading days (post-periods).

To validate the observed effects, we incorporated the sectoral index price from the IDX as a control variable for each firm, alongside the daily return and volatility ($=$ return^2) of the sectoral index. This comprehensive approach ensured a thorough analysis of the news articles' impact on stock trading dynamics.

Finally, we ran R `CausalImpact` library (Brodersen et al., 2015) with the previous specifications on each generated time series. From each output of `CausalImpact`, we extracted the p-value and the effect, which is the difference between the actual and predicted values of the main time series during

[2] Each topic can be associated with multiple industry standards, although the topic's summary within each industry standard can be different.

[3] https://www.idx.co.id/id/data-pasar/data-saham/daftar-saham

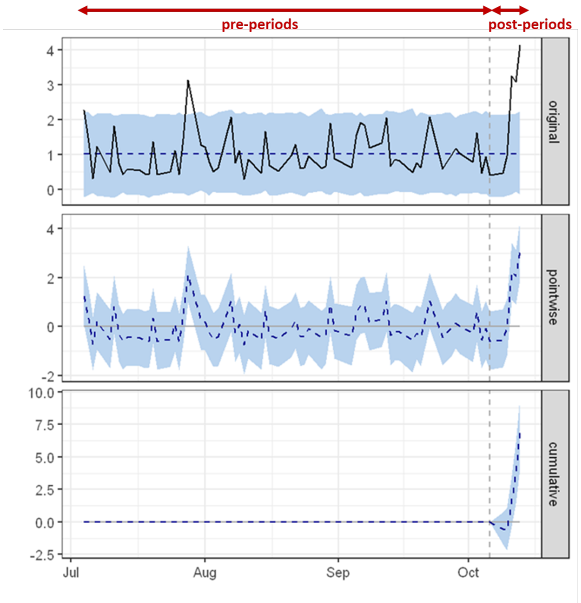


Figure 1: Example `CausalImpact` output for the trading volume of stock IDX:JSMR on 9 Oct 2017 (news similarity score to environment topic = 96$^{\text{th}}$ percentile). The dashed vertical line denotes the event/news date.

post-periods. If environment-related news significantly affects stock trading, then we would expect the resulting p-values to be low and the effects to be large and positive, i.e. since we tested on stock trading volume, a large and positive effect means that the news generates excess stock trades. An example output of the `CausalImpact` library for a given firm-date can be seen in Figure 1.

As an additional test, we ran the same analysis on 5 dummy datasets constructed by replacing dates in the original dataset with random dates from the same year, in which none of the top 10% environment-related news about the firm was published.

4. Results

The number of environment-related news (firm-dates) which result in a positive and significant effect (i.e. larger than predicted trading volume, $p < 0.05$) is tabulated in Table 1 and can be distilled into three main observations. First, the top 10% news articles most similar to environmental topics led to higher-than-predicted trading volumes for 17.9% of firm-dates over the next 5 trading days. Conversely, 9.7% of firm-dates experienced reduced trading (not detailed in Table 1), while for 72.4% of firm-dates, trading volumes did not significantly deviate from the pre-period levels. For context, control datasets (outlined in §3.4) revealed that 16.0% $\pm$ 0.4% of firm-dates were followed by excess trading (t-statistic of difference in means

Table 1: Percentage of events followed by excess trading, by the number of trading days post-periods and the proportion of news articles sorted by the similarity to environment topic by SASB standards

Number of trading days in post-periods	Top 10% news	Top 5% news	Top 1% news
5	17.9%	18.2%	18.1%
10	21.3%	21.4%	21.9%
20	25.0%	25.2%	25.6%

Table 2: Percentage of events followed by excess trading per sector

IDX sector	N. firms	N. news	% events
Technology	32	328	24.4%
Consumer Cyclical	132	1375	20.0%
Cons. Non-Cyclical	103	2814	18.4%
Energy	75	3069	18.1%
Finance	98	2254	18.0%
Basic Materials	93	2278	17.9%
Industrials	51	845	17.5%
Infrastructure	129	4279	16.9%
Transportation	29	392	16.1%
Healthcare	26	797	16.1%
Total	768	18431	17.9%

= 6.69, $p < 0.0001$, df = 25,593), and 10.6% $\pm$ 0.2% by reduced trading, indicating that our embedding methodology successfully identifies market-relevant environmental news associated with increased trading activity.

Second, the proportion of news articles followed by excess trading escalates with the length of the post-period: from 17.9% with a 5-day window to 21.3% for 10 days and 25.0% for 20 days. However, this trend may be influenced by other unaccounted events, which are more likely in longer post-periods.

Finally, increasing the topic similarity threshold by only including the top 5% or top 1% of news most similar to environmental topics does not significantly increase stock trading volume in the following days compared to the top 10%, with percentages inching from 17.9% to only 18.2% and 18.1% for 5, 10, and 20 post-periods, respectively.

In an additional analysis, we examined whether the effects vary across sectors. The results, summarized in Table 2 using IDX's firm classification, show that the significance of effects differs between sectors, ranging from 16.1% of firm-dates in the IDX Healthcare sector to 24.4% in IDX Technology.

5. Conclusion

In this paper, we have described how using SBERT to generate embeddings of sustainability standards can be used to identify ESG environment-related news. We tested our framework on an Indonesian news dataset between 2016 and 2023. The results show that the identified sustainability-related news tends to be followed by increased stock trading volumes on firms mentioned in the news. Thus, this approach can be expanded upon as a simpler alternative to building an ESG news labeling model or using third-party data providers. Firms and regulators could also use such an approach to monitor ESG news that is potentially market-moving.

We acknowledged the limitations of this study and identified four ways to extend it. First, to explore long-term effects, one could examine the impact of ESG news on firm values using regression analysis. While we opted for an event study methodology based on `CausalImpact` library, conducting regression analysis with varying numbers of pre- and post-periods could enhance the results' robustness. Second, the scope of our study is limited to environment-related news; however, the methodology should be readily applicable to social and governance pillars of ESG as well. Third, future studies could investigate whether significant price movements are associated with the sentiments of the preceding ESG-related news articles. Finally, our approach was tested solely on a dataset from one news media outlet. Thus, future studies could include datasets from various news media outlets or other stock markets.

6. Bibliographical References

Kölbel Julian F Berg, Florian and Roberto Rigobon. 2019. Aggregate confusion: The divergence of esg ratings. *Review of Finance*, 26(6):1315–1344.

Fabian Billert and Stefan Conrad. 2023. Team HHU at the FinNLP-2023 ML-ESG task: A multi-model approach to ESG-key-issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 146–150, Macao. -.

Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. COLT '92, page 144–152, New York, NY, USA. Association for Computing Machinery.

Kay H. Brodersen et al. 2015. Inferring causal impact using bayesian structural time-series models. *Annals of Applied Statistics*, 9(1):247–274.

Gunther Capelle-Blancard and Aure´lien Petit. 2017. Every little helps? esg news and stock market reaction. *Journal of Business Ethics*, 157:543–565.

Sugat Chaturvedi, Ekaterina Prytkova, Tommaso Ciarli, and Önder Nomaler. 2023. What is the future of automation? using semantic analysis to identify emerging technologies. Technical report.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115, Macao. -.

Davide Chicco. 2021. Siamese neural networks: An overview. *Artificial neural networks*, pages 73–94.

Costanza Consolandi, Robert G Eccles, and Giampaolo Gabbi. 2022. How material is a material issue? stock returns and the financial relevance and financial intensity of esg materiality. *Journal of Sustainable Finance & Investment*, 12(4):1045–1068.

François Derrien, Philipp Krueger, Augustin Landier, and Tianhao Yao. 2021. ESG News, Future Cash Flows, and Firm Value. Technical report.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kadek Denaya Rahadika Diana and Masayu Leylia Khodra. 2023. IndoSBERT: Enhancing Indonesian Sentence Embeddings with Siamese Networks Fine-tuning. In *2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE.

Alex Edmans and Marcin Kacperczyk. 2022. Sustainable finance. *Review of Finance*, 26(6):1309–1313.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

IFRS Foundation. 2023. SASB® Standards.

Steve Johnson. 2023. Hundreds of funds to be stripped of ESG rating.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. EaSyGuide : ESG issue identification framework leveraging abilities of generative large language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 127–132, Macao. -.

MFF Lubis and R Rokhim. 2021. The effect of environmental, social, and governance (ESG) disclosure and competitive advantage on companies performance as an implementation of sustainable economic growth in Indonesia for period of 2015-2019. In *IOP Conference Series: Earth and Environmental Science*, volume 940, page 012059. IOP Publishing.

Ivan Mashkin and Emmanuele Chersoni. 2023. HKESG at the ML-ESG task: Exploring transformer representations for multilingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 140–145, Macao. -.

Ninditya Nareswari, Małgorzata Tarczyńska-Łuniewska, and Rizqi Umar Al Hashfi. 2023. Analysis of Environmental, Social, and Governance Performance in Indonesia: Role of ESG on Corporate Performance. *Procedia Computer Science*, 225:1748–1756.

Hilal Pataci, Yunyao Li, Yannis Katsis, Yada Zhu, and Lucian Popa. 2022. Stock price volatility prediction: A case study with AutoML. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 48–57, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging BERT language models for multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 121–126, Macao. -.

Principles for Responsible Investment. 2023. Data Portal.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Blaise Sandwidi and Suneer Pallitharammal Mukkolakal. 2022. Transformers-based approach for a sustainability term-based sentiment analysis (STBSA). In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 157–170, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Indira Sen, Daniele Quercia, Licia Capra, Matteo Montecchi, and Sanja Šćepanović. 2023. Insider stories: analyzing internal sustainability efforts of major US companies from online reviews. *Humanities and Social Sciences Communications*, 10(1).

George Serafeim and Aaron Yoon. 2022. Which Corporate ESG News Does the Market React To? *Financial Analysts Journal*, 78(1):59–78.

Walid Taleb, Théo Le Guenedal, Fréderic Lepetit, Vincent Mortier, Takaya Sekine, and Lauren Stagnol. 2020. Corporate ESG news and the stock market. *Available at SSRN 3723799*.

Patrick Temple-West. 2023. Sp drops esg scores from debt ratings amid scrutiny.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging contrastive learning with BERT for ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 116–120, Macao. -.

Development and Evaluation of a German Language Model for the Financial Domain

Nata Kozaeva, Serhii Hamotskyi, Christian Hänig
Anhalt University of Applied Sciences
Bernburger Str. 55, 06366 Köthen, Germany
nata.kozaeva@hs-anhalt.de, serhii.hamotskyi@hs-anhalt.de, christian.haenig@hs-anhalt.de

Abstract

Recent advancements in self-supervised pre-training of Language Models (LMs) have significantly improved their performance across a wide range of Natural Language Processing (NLP) tasks. Yet, the adaptation of these models to specialized domains remains a critical endeavor, as it enables the models to grasp domain-specific nuances, terminology, and patterns more effectively, thereby enhancing their utility in specialized contexts. This paper presents an in-depth investigation into the training and fine-tuning of German language models specifically for the financial sector. We construct various datasets for training and fine-tuning to examine the impact of different data construction strategies on the models' performance. Our study provides detailed insights into essential pre-processing steps, including text extraction from PDF documents and language identification, to evaluate their influence on the performance of the language models. Addressing the scarcity of resources in the German financial domain, we also introduce a German Text Classification benchmark dataset, aimed at fostering further research and development in this area. The performance of the trained models is evaluated on two domain-specific tasks, demonstrating that fine-tuning with domain-specific data improves model outcomes, even with limited amounts of domain-specific data.

Keywords: Language Modelling, Financial NLP, German

1. Introduction

In the rapidly evolving financial sector, where precision and accuracy of information dissemination are paramount, the development of specialized Language Models (LMs) becomes not just beneficial but essential. The financial domain is characterized by its dynamic nature, requiring the processing of vast quantities of data that include market reports, regulatory filings, and financial news. Each of these data types is imbued with complex jargon, numerical information, and nuanced expressions specific to the financial industry. The application of specialized language models in this sector enables several promising use cases, including automatic checking for eligibility criteria (Hänig et al., 2023), facilitating automatic financial reporting (Oyewole et al., 2024), and ensuring automatic consistency checking (Ali et al., 2023).

The predominance of English in the global financial literature has led to a wealth of text data in English, ranging from publicly accessible 10-K forms[1] and earnings call transcripts to comprehensive resources like Seeking Alpha[2] and the System for Electronic Document Analysis and Retrieval[3]. In stark contrast, the German financial sector faces a significant challenge due to the scarcity of equivalent resources in the German language, highlighting a critical gap in both financial text corpora and annotated datasets within this domain.

This research addresses this gap by development and evaluation of a German LM fine-tuned for the financial sector. We compare its performance on downstream tasks against a general-purpose German LM (referred to as vanilla LM). Our goal is to ascertain whether a domain-specific LM can surpass the vanilla model in the nuanced task of processing German financial texts. Through a series of experiments involving both the further pre-training of existing LMs and the training of new models from scratch using various dataset configurations, we explore this question in depth.

Research, such as that by Hänig et al. (2023), demonstrates that an English FinBERT model (Yang et al., 2020) fine-tuned for the financial domain falls short in performance when applied to German financial data, compared to a general German LM, which, in turn, outperforms an English model on out-of-domain German tasks.

Considering the features of financial language, including complex sentence structures, formal tone, specialized vocabulary, and legal terminology, the development of a dedicated German LM for the financial domain is imperative. To facilitate the development of German financial LMs, we perform thorough analyses of financial text corpus compilation and study the effect of various pre-processing steps. Furthermore, we create and publish a new German benchmark dataset for evaluation language models in the financial domain.

Our research utilizes the BERT architecture (Devlin et al., 2019), specifically German BERT (Chan

[1] https://www.sec.gov/
[2] https://seekingalpha.com/
[3] https://www.sedarplus.ca/

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 40–49

et al., 2020), drawing inspiration from its application in related fields, including FinBERT (Yang et al., 2020), SciBERT (Beltagy et al., 2019), ClinicalBERT (Huang et al., 2020), and BioBERT (Lee et al., 2019).

1.1. Related Work

The same approach was used to develop models for other domains: ClinicalBERT pretrained on clinical notes (Huang et al., 2020), SciBERT pretrained on scientific papers (Beltagy et al., 2019).

There is a significant shortage of publicly available text corpora and labeled datasets related to financial topics in the German language. The CODE ALLTAG corpus (Krieg-Holz et al., 2016) is a text dataset comprised of emails in the German language. Within this corpus, there is a "FINANCE" collection, which includes 174,375 emails, containing nearly 2.5 million sentences. The Bundesstelle for Open Data has released deutschland[4] and handelsregister[5] to enable the retrieval and download of data from the Bundesanzeiger and Handelsregister, respectively. Data extracted from the Bundesanzeiger has been used in academic research, serving various purposes, such as company name recognition (Loster et al., 2017) and the training of language models on text resembling financial content (Biesner et al., 2022). However, these datasets were not made publicly available.

Jørgensen et al. (2023) conducted a comprehensive analysis of labeled datasets in the financial domain revealing that the vast majority of resources is in English. Only few non-English datasets exist with just one multilingual dataset containing the German language: SIXX-Corpora (Gaillat et al., 2018) for sentiment analysis (non-open dataset).

1.2. Contribution

Our first contribution involves the creation of a German financial dataset suitable for multiclass and multilabel classification tasks. For this we used the MultiFin dataset and translated it in German.

Our second contribution includes development and evaluation of domain-specific LMs for German financial language and thorough analysis of the impact of decisions made during dataset construction and pre-processing on the models' performances.

2. Financial Data for Language Model Training

Delimiting the scope of financial language is challenging, covering diverse subdomains like capital markets, banks, and insurance, with data from varied sources including financial documents, laws, and news. These sources, while thematically aligned, differ in vocabulary and complexity—news articles are generally more accessible, while documents like prospectuses feature domain-specific jargon. Some texts, such as annual reports, follow strict standards, contributing to their uniformity.

Given this linguistic diversity and the specific characteristics of various document types, we opted to construct a dataset that encompasses multiple categories of documents. This approach aims to maximize the dataset's diversity, thereby providing a comprehensive foundation for training and evaluating our language models.

2.1. Financial Document Collection

In this study, we utilize FinCorpus-DE10k (Anonymous, 2024), a domain-specific dataset composed of various document types, as a foundation for our analysis. It features the following document types:

Base and Final Terms Prospectuses Financial prospectuses that provide terms and conditions of the issuance of financial securities. The structure, content, release procedure are regulated by Article 8 and 10 of REGULATION (EU) 2017/1129 ("Prospectus Regulation").

Annual Reports of the Bundesbank Documents providing information about economic and financial issues, monetary policy, risks of financial stability etc. Annual reports usually contain a larger number of data visualizations and images.

International Financial Reporting Standards EU International Financial Reporting Standards (IFRS)[6] from the years 2017–2023. These documents define standards as accounting rules that facilitate understanding and comparability of financial statements across borders to ensure corporate transparency.

Law Documents containing German laws in the financial domain. The core regulations applicable to the financial sector in Germany are laid down in the Banking Act (KWG)[7]; the Securities Institutions Act (WpIG)[8], the Securities Trading Act (WpHG)[9] etc. as well as EU Directives implemented into German law.

Informational Materials Brochures and advertisements in the area of finance, description of financial products and general terms and conditions. Most documents of this collection have a wider variety of fonts, photos, colors, and are mostly aimed at a more general audience.

[4]https://github.com/bundesAPI/deutschland
[5]https://github.com/bundesAPI/handelsregister
[6]https://www.ifrs.org/
[7]https://www.gesetze-im-internet.de/kredwg/
[8]https://www.gesetze-im-internet.de/wpig/
[9]https://www.gesetze-im-internet.de/wphg/

	num txt doc	num tokens	num numeric tok	num sent	mean length tok.	mean length sent.
Final terms	10,986	112,344,212	5,307,180	4,026,251	6	26
Base prospectuses	731	49,353,187	1,996,865	1,435,924	5	28
Annual reports	88	7,406,590	731,624	318,683	6	21
Informational materials	139	2,200,884	68,976	94,071	6	20
Law	138	4,062,628	373,439	95288	6	28
IFRS	7	3,726,002	135,215	107,577	6	30
BBK monthly	412	48,182,195	21,720,392	1,750,691	3	25
News	20	2,144,970	52,497	94,888	6	19
Wikipedia	1	9,181,311	331,821	457,495	6	17
Total	12,516	238,601,979	30,718,009	8,380,868	-	-

Table 1: Document statistics in TXT files

Bundesbank Monthly Reports The initial collection (PDF documents) contains 866 monthly reports of the German Bundesbank from the years 1949–2022.

Statistics of the dataset are provided in Table 1.

2.2. Layout and Text Extraction

The PDF documents contain files with very diverse layouts. Financial information is oft presented using tables and charts, incorporating a large number of figures compared to regular language. Another layout features are columns and table-like structures. The presence of columns and tables can disrupt the linear extraction (corresponds to the natural reading order) of text. In context of pre-training a LM this is important, because otherwise, the attention mechanism will be applied on a sequence with an incorrect token order.

For the experiments pdfplumber [10] library was employed to extract the text from PDF files. Given the uncertainty in document layouts, our initial experiment used a text extraction library without adjustments for specific structures.

Next we conducted text extraction taking into account possible layout differences. Assuming that the document collections likely contain columns and tables due to their financial nature, the impact of an alternative text extraction method on the Model's performance was assessed.

PyMuPDF[11] was used for layout-specific extraction. Upon comparing the results with those obtained using pdfplumber, this solution demonstrated accuracy within the randomly selected documents chosen for comparison. The extracted text was observed in its natural reading order.

2.3. Language Detection

To train a German language model, a critical step is to analyze the linguistic composition of our dataset to ascertain the prevalence and distribution of languages within it. This analysis leverages insights from Anonymous (2024), wherein the authors utilized the automatic language identification tool lingua-py[12] to quantify the language proportions across the document collection.

Within the dataset, the predominant language is German, succeeded by English, while the presence of other languages is comparatively minimal. It is presumed that the detection of other languages originates from language detection inaccuracy. Figure 1 illustrates a histogram of language distribution within the dataset, denoting German, English, and other languages.

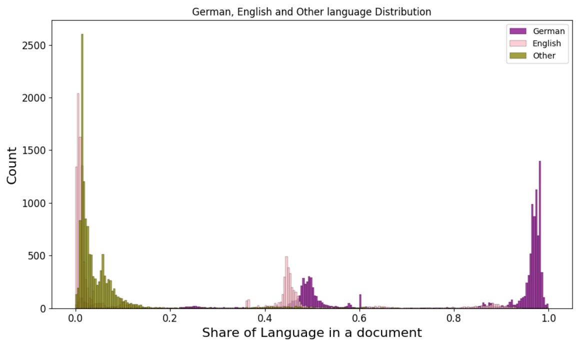


Figure 1: Language distribution for each document of the corpus

The dataset predominantly features documents exclusively in German. There is also a significant subset of bilingual documents, with German and English content, primarily between 40-60%. A smaller fraction of documents includes *Other* languages paired with either English or German. Trilingual documents, which are relatively scarce, are likely artifacts of language identification errors and are considered noise. To refine the dataset for German language specificity, the language detection algorithm from SpaCy[13] was employed to segregate and remove English language texts, thereby curating a corpus composed solely of German language.

[10] https://github.com/jsvine/pdfplumber
[11] https://github.com/pymupdf/PyMuPDF
[12] https://github.com/pemistahl/lingua-py
[13] https://spacy.io/

2.4. Corpus Compilation

For the experiments, the financial data was augmented with common language data, utilizing the *Wortschatz* collection (Goldhahn et al., 2012) from Leipzig University[14] to create a corpus of common German language. This corpus consists of separate sentences of varying length. In the process of training a LM, each sentence serves as an individual instance or training example. Given that the financial corpus is composed of documents, it inherently contains more contextual information compared to isolated sentences. Consequently, at this point, the aim was to incorporate an additional common language corpus that comprises full texts rather than discrete sentences. The German colossal, cleaned Common Crawl corpus[15] was employed, comprising texts of varying lengths.

Further the results of LM performance for mixed datasets (financial corpus mixed with common language sentences and financial corpus mixed with common language texts) will be compared. The token count in both corpora of common German language is approximately equivalent to that of the financial corpus so that the token count in the mixed corpora is approximately double that of the financial corpus.

At this point term frequency was calculated and sorted in the financial and common language corpus to check to which extent the domain specific dataset vocabulary varies from the common language. There was a considerable contrast between the two corpora, emphasizing the financial corpus's domain-centric nature.

2.5. Corpus Configurations

By employing two text extraction methods, language detection and mixed corpus, we analyzed an array of data combinations.

From a *data* perspective, the following preprocessing configurations were explored:

none Text is extracted as it is.
language detection German-only language extraction (leveraging language detection).
layout detection Extraction accounts for document layout (applying columns and tables detection).
layout & language detection Extraction considering both layout detection and German-only language extraction.

From a *domain-focused* perspective, examination encompassed:

fin Financial data is used.

[14]https://corpora.uni-leipzig.de
[15]https://german-nlp-group.github.io/projects/gc4-corpus.html

No.	Topic	Examples
1	Technology	1,088
2	Industry	1,239
3	Tax& Accounting	3,371
4	Finance	1,447
5	Government& Controls	912
6	Business& Management	1,991
Total		10,048

Table 2: Overview of High-Level tags across the 6 classes used in the multiclass classification task (Jørgensen et al., 2023)

mixed A combination of financial and general language data is used.

Regarding *mixed* data, the following data is added to the financial corpus:

sentence General German language sentences.
text General German language text providing a larger context.
text and sentence Both sentence and text data.

3. Financial Datasets for Downstream Evaluation

In the context of financial language processing, the evaluation of language models on domain-specific tasks is crucial for assessing their practical utility and effectiveness. This section delves into the use of two pivotal downstream tasks: Text Classification (TC) and Named Entity Recognition (NER), which serve as benchmarks for evaluating the performance of our fine-tuned German financial language models.

3.1. Financial Text Classification Dataset

Text Classification in the financial domain involves categorizing text into predefined categories, an essential function for organizing and interpreting vast amounts of financial data. Our new benchmark dataset is based on the MultiFin dataset (Jørgensen et al., 2023), a rich collection of real-world financial article headlines annotated with both high-level and low-level topics. The original MultiFin dataset consists of 10,048 real-world financial article headlines in 15 languages. The dataset is annotated with 6 high-level topics and 23 low-level topics for multi-class and multi-label classification, respectively (see Table 2, Figure 3). For the multi-label classification task, there are up to 3 annotations per example, which sums up to 14,230 annotations with an average of 1.4 annotations per example.

OpenAI API gpt-3.5-turbo[16] was used to translate the dataset examples from the source languages to German. Each example was accompanied by a specific prompt that included the source language

[16]https://platform.openai.com/docs/models/gpt-3-5

from the dataset. This guided the model more effectively, eliminating the need for language detection as the source language was explicitly provided.

Given the dataset's multilingual nature and time constraints, exhaustive manual verification of each translation was impractical, making it impossible to guarantee translation perfection. To evaluate translation quality, we selectively reviewed 100-150 examples per class across English, Italian, and Russian, focusing primarily on semantic accuracy. Translations were classified as either *semantically correct* or *semantically incorrect*, with the latter category excluded from further grammatical evaluation due to their failure in meaning transmission. This methodology confirmed that the translations maintain a quality level adequate for their intended analytical use, as evidenced by the outcomes illustrated in Figure 2:

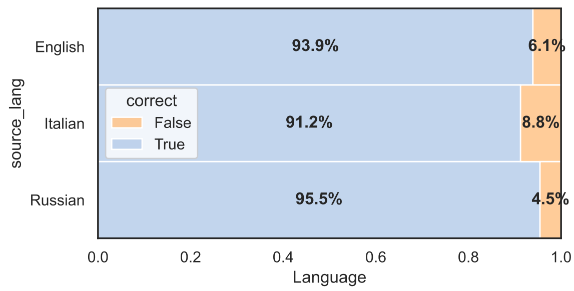


Figure 2: Language distribution for each document of the corpus

The original MultiFin dataset comprises three subsets: train, dev, and test, containing *6430*, *1608*, and *2010* examples, respectively. The German MultiFin dataset features the same number of instances per split as the original MultiFin dataset, as all instances have been translated to German.

Given the problem of imbalanced classes (Kubat, 2000), instances for each class in each subset were counted. This was done to ensure that each subset (train, val, test) contains a proportional number of examples for each class (see Figure 3).

The created German MultiFin Dataset is available on HuggingFace[17].

3.2. Financial Named Entity Recognition Dataset

Named Entity Recognition (NER) in the financial domain seeks to identify and classify key information pieces from unstructured text, such as financial instruments, criteria, and terms. For this task, we use a dataset for examination of eligibility criteria from securities prospectuses (Hänig et al., 2023)

[17] https://huggingface.co/datasets/anhaltai/german-financial-dataset

Target type	Train	Test
coupon fixed	431	375
coupon variable index	56	84
coupon variable margin	38	42
coupon variable operator	37	43
coupon variable tenor	45	75
currency	514	577
early redemption amount	64	52
early redemption	177	108
isin	421	417
principal amount	784	800
redemption at maturity amount	26	42
redemption at maturity	370	347
special termination	96	109
special termination amount	61	63
status non preferred	56	47
status senior non preferred	488	333
type of instrument	431	422

Table 3: Number of annotations per target type in the dataset splits (Hänig et al., 2023).

which is meticulously annotated across 17 distinct entity classes.

Being able to detect this array of classes empowers models to advance the automation process for determining the eligibility criteria of securities prospectuses issued by central banks, addressing eight intricately varied criteria essential for evaluating an issuance's eligibility. The criteria encompass a broad spectrum, including:

- Coupon
- Currency
- Early redemption amount
- Principal amount
- Redemption (amount) at maturity
- Special termination right
- Liquidation Status (Senior/Subordinated)
- Type of instrument

The documents were annotated manually, to assess consistency of the manual annotation process the authors measured inter-annotator agreement (IAA) using Intersection over Union (IoU). The resulting IAA scores range from 0.731 to 0.932 (Hänig et al., 2023). The total number of annotations per type are shown in Table 3. The annotated data was converted and transformed into a dataset for token classification, namely into BIO-encoded sequences. The labels were aligned to the tokenization of the BERT model.

4. Language Model Training

32 distinct training experiments were conducted, categorized based on various factors, which were to be explored.

The factors encompassed aspects described in subsection 2.5. Each data aspect was used for different model *weight initialization*:

pre-trained model training uses pre-trained weights,

from scratch model training uses randomly initialized weights.

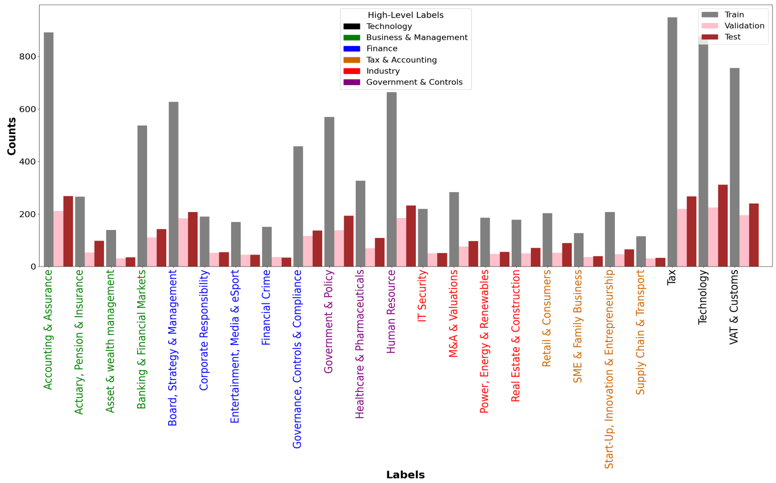


Figure 3: Distribution of Labels Across Training, Validation, and Test Sets. (The bars represent the distribution of low-level labels, with colors corresponding to high-level labels.)

4.1. Training Results of Language Models

For language model training, we report loss scores which directly correspond to the commonly used intrinsic language model evaluation metric Perplexity. The following regularities can be observed (see Table 4):

Comparing models based on *weight initialization*, pre-trained models consistently outperform the models initialized from scratch in all experiments.

From a *data* perspective, language detection improves the results for four models , but slightly lowers the performance in the other four compared to models without language detection. *Layout detection* consistently contributes to the model performance.

The results obtained when *layout detection* has been applied outperformed all other models except for the model trained from scratch using a mixed dataset comprising text and sentence examples.

When comparing different example compositions within *mixed* datasets, an evident pattern can be observed. Pre-trained models leveraging text examples tend to outperform other variants (sentences combined with texts or solely sentences). Conversely, models trained from scratch perform better when trained on mixed datasets with sentence compositions.

In the comparison between text & sentences and solely sentences compositions for pre-trained models, the text & sentences approach is worse for the model without layout and language detection, it fares better for the other three models.

The best model performance is achieved when fine-tuning a pre-trained model on financial data using layout detection. This configuration achieves a loss of *0.72*.

5. Evaluation

5.1. Text Classification Task

In *multiclass* classification task two models outperform the baseline (vanilla German BERT-base) model: the model *pre-trained on financial corpus with language detection* and the model *pre-trained with layout and language detection on a mixed dataset with text and sentence composition*. The LM model, that exhibited best result based on intrinsic metrics (cross-entropy loss and perplexity) did not achieve the best score for the downstream task. Conversely, the poorest-performing LM, trained from scratch on financial data without language and layout detection, similarly demonstrated worst performance for the downstream task.

While LM results indicate that models with both layout and language detection consistently achieved inferior results compared to those with layout detection, the downstream task results present a more nuanced picture. Five models incorporating language and layout detection show better performance on multi-class classification and four on multi-label classification compared to those employing layout detection, only.

Among from scratch-trained models, one model stood out with notably lower loss compared to other from-scratch counterparts. This model demonstrates a slightly better performance in this downstream task (0.8537) compared to most scratch-trained models, except for the model featuring layout and language detection trained on mixed data with text example composition (0.8811).

The analysis of the downstream task suggests, that a small loss (or a small perplexity) does not guarantee great performance on this downstream task.

Multi-label classification results are shown in

	fin pretr	fin from scr	mixed data text&sent pretr	mixed data text&sent from scratch	mixed data sent pretr	mixed data sent from scratch	mixed data texts pretr	mixed data texts from scr
none	1.11	7.33	1.44	**2.78**	1.32	4.51	0.91	**5.48**
language detection	0.92	5.33	1.37	5.03	1.43	4.06	1.02	5.58
layout detection	**0.72**	**5.28**	**1.22**	4.30	**1.28**	**3.89**	**0.83**	**5.48**
layout & language detection	0.79	5.29	1.30	4.90	1.35	4.05	0.91	5.54

Table 4: Comparison of Language Model Training Results (loss values)

	none	language detection	layout detection	layout & language detection
fin pretrained	0.8829	0.8849	0.8915	**0.8957**
fin from scratch	0.0	0.8151	0.8261	0.8201
mixed text & sent pretrained	0.8819	**0.8940**	**0.8934**	0.8842
mixed text & sent from scratch	0.8547	0.8379	0.8901	0.8359
mixed sent pretrained	0.8834	**0.8923**	0.8905	0.8890
mixed sent from scratch	0.8282	0.8361	0.8269	0.8333
mixed text pretrained	**0.8944**	**0.8967**	0.8890	**0.8957**
mixed text from scratch	0.8175	0.8165	0.8329	0.8380

Table 5: Multi-class / multi-label TC results on downstream dataset (macro-averaged F-score)

Table 5. For this task, *seven* models outperform the baseline model, additionally one model (pre-trained with financial data with layout detection) achieves the same results as the baseline.

The baseline model exhibited comparable results for both multi-class and multi-label tasks, with performance metrics of 0.891 and 0.8915, respectively. In contrast, the trained models displayed varying degrees of performance across these tasks.

Among the models that surpassed the baseline in this task, two belong to the *layout and language detection* category, while three models were trained without layout and language detection enabled.

Concerning mixed data, text example composition seems to have a positive impact, as three pre-trained models of this category outperformed the baseline model.

The from scratch trained model using financial data without language and layout detection ceased training after just 2 epochs due to its inability to improve results, yielding a 0.0 F-Score.

The most successful from scratch trained model for this task was the model with layout detection and sent & text example composition (*0.8901*).

Models utilizing *layout detection* generally outperformed those lacking this feature, with one exception observed in the case of a model trained on mixed data using text example compositions. This could be attributed to the substantial dataset providing more contextual information, countering the negative impact of the absence of layout detection.

On the other hand, among models with language and layout detection compared to those with language detection only, three models of the first category outperformed the language detection.

5.2. Named Entity Recognition Task

Results for the NER downstream task are shown in Table 6. The F1-score is calculated separately for every class in the dataset. Additionally, macro-averaged F1-scores are reported to provide a single performance indicator.

Seven models (highlighted with bold font) outperformed the vanilla German BERT base (0.738). All of them belong to the category of pre-trained models while in every pre-trained model category there is at least one that outperformed the vanilla model. Three of them are pre-trained on data with *layout and language detection*. For this downstream task the model *pre-trained only on financial corpus with layout and language detection* achieved the best results. This might be explained by a a strong domain-focus in the data of the NER task.

As outlined in 3.2, the dataset comprises securities prospectuses annotated according to a predefined set of eligibility criteria. The nature of the dataset's content and the specificity of its labels demonstrate a closer alignment with the financial domain than observed in datasets utilized for other downstream tasks. Such alignment enables a comprehensive evaluation of LMs on this dataset to effectively assess their domain-specific performance capabilities.

Augmenting the dataset with common language data with different example composition contribute to the model performance, however the results are slightly worse than of the model pre-trained on the financial data only. Similar to other experiments, models trained from scratch achieve inferior results compared to models using pre-trained weights. Additionally, in contrast to multi-class and multi-label classification, there is a more pronounced disparity in performance between pre-trained models and those built from scratch.

6. Discussion

Our analysis reveals that certain models consistently surpass the baseline German BERT model across all downstream tasks, suggesting that the observed performance gains are systematic rather

	none	language detection	layout detection	layout & language detection
fin pretrained	0.711	0.73	0.732	**0.748**
fin from scratch	0.0	0.387	0.365	0.385
mixed text & sent pretrained	**0.74**	0.695	0.733	**0.74**
mixed text & sent from scratch	0.546	0.501	0.563	0.548
mixed sent pretrained	0.734	0.706	**0.745**	**0.744**
mixed sent from scratch	0.566	0.551	0.59	0.561
mixed text pretrained	0.732	**0.739**	**0.742**	0.724
mixed text from scratch	0.407	0.39	0.4	0.386

Table 6: NER results on downstream dataset (macro-averaged F-score)

than coincidental. This opens up potential for further refinements at both data and model levels.

6.1. Data-Driven Improvements

Models pre-trained on extensive corpora have shown better performance, potentially due to larger data sizes which are critical for models trained from scratch to exhibit comparable results to pre-trained models. For instance, the FinBERT model (Yang et al., 2020) was trained from scratch on sizable corpora exceeding one billion tokens. Similarly, the training dataset for BloombergGPT encompassed a significant token count from financial domains (Wu et al., 2023). This raises the question about the data volume threshold at which models trained from scratch for a specified domain begin to perform on par with pre-trained models.

Deduplication of financial documents presents another research direction, considering the frequent occurrence of redundant text which can affect both training efficiency and cost. Lee et al. (2022)'s work on deduplication indicates potential benefits in training efficiency. However, the impact of deduplication on model perplexity and the balance between content removal and retention of document context has yet to be fully understood. Investigating deduplication at the document level could shed light on its effects.

6.2. Refining Language Detection

Models incorporating both layout and language detection underperformed compared to those utilizing layout detection alone. This discrepancy might be due to the language detection method's word-by-word operation, which can misidentify language transitions in bilingual documents. A sentence-based language detection approach, filtering out sentences with insufficient German content, could preserve context better and improve performance. Assessing this method's impact on both intrinsic metrics and downstream task efficacy is a promising area for exploration.

6.3. Data Filtering Techniques

In datasets like the Bundesbank Monthly Reports, prevalent layout elements such as tables and checkboxes could introduce noise due to a higher ratio of numeric tokens and shorter mean token lengths. Investigating advanced filtering methods or document understanding techniques could be beneficial in addressing these challenges.

6.4. Model and Training Enhancements

Improvements in the training process could include utilizing both Masked Language Modelling and Next Sentence Prediction tasks of BERT for text examples. Further research could explore the impact of training models on additional tasks such as Sentiment Analysis or Named Entity Disambiguation, drawing comparisons with models like BloombergGPT.

7. Conclusions and Future Research

The central aim of this research was to develop a language model specialized for the German financial domain.

A financial corpus was meticulously assembled and two domain-specific datasets were assembled and used for downstream evaluation. The corpus compilation was subject to a series of preprocessing steps and was enriched with a general language data. Furthermore, we created and published the new German dataset *German MultiFin* useful for multi-class multi-label classification in the financial domain.

Across three downstream tasks – multi-class classification, multi-class multi-label classification and Named Entity Recognition – several models displayed enhanced performance relative to the baseline. Particularly, the model pre-trained on the financial corpus incorporating both layout and language detection emerged as superior, yielding the highest average scores across tasks. The strategic inclusion of layout detection, both in conjunction with and independent of language detection, significantly bolstered the performance of pre-trained models in downstream applications. The expansion of financial data with general language content was advantageous for models trained from scratch.

Future research could delve into further refinements, potentially examining alternative language filtering techniques, data deduplication approaches, and the procurement of more domain-specific data.

8. Bibliographical References

Syed Musharraf Ali, Tobias Deußer, Sebastian Houben, Lars Hillebrand, Tim Metzler, and Rafet Sifa. 2023. Automatic Consistency Checking of Table and Text in Financial Documents. *Proceedings of the Northern Lights Deep Learning Workshop*, 4.

Anonymous Anonymous. 2024. FinCorpus-DE10k: A Corpus for the German Financial Domain. In *Accepted for the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italy.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. ArXiv:1903.10676 [cs].

David Biesner, Rajkumar Ramamurthy, Max Lübberinf, Lars Hillebrand, Anna Ladi, Maren Pielka, Rüdiger Lotz, Christian Bauckhage, and Rafet Sifa. 2022. Anonymization of German financial documents using neural network-based language models with contextual word representations. *International Journal of Data Science and Analytics*, 13(2):151–61.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's Next Language Model. ArXiv:2010.10906 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understandng*. Association for Computational Linguistics.

Thomas Gaillat, Manel Zarrouk, André Freitas, and Brian Davis. 2018. The SSIX Corpora: Three Gold Standard Corpora for Sentiment Analysis in English, Spanish and German Financial Microblogs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. ArXiv:1904.05342 [cs].

Christian Hänig, Markus Schlösser, Serhii Hamotskyi, Gent Zambaku, and Janek Blankenburg. 2023. NLP-based Decision Support System for Examination of Eligibility Criteria from Securities Prospectuses at the German Central Bank. In *Proceedings of AAAI23 Bridge 8: AI for Financial Institutions*, Washington, D. C., USA.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A Dataset for Multilingual Financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 894–909, Dubrovnik, Croatia. Association for Computational Linguistics.

Ulrike Krieg-Holz, Christian Schuschnig, Franz Matthies, Benjamin Redling, and Udo Hahn. 2016. CodE Alltag: A German-Language E-Mail Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2543–2550, Portorož, Slovenia. European Language Resources Association (ELRA).

Miroslav Kubat. 2000. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. *Fourteenth International Conference on Machine Learning*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. ArXiv:1901.08746 [cs].

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating Training Data Makes Language Models Better. ArXiv:2107.06499 [cs].

Michael Loster, Zhe Zuo, Felix Naumann, Oliver Maspfuhl, and Thomas Dirk. 2017. Improving Company Recognition from Unstructured Text by using Dictionaries.

Adedoyin Tolulope Oyewole, Omotayo Bukola Adeoye, Wilhelmina Afua Addy, Chinwe Chinazo Okoye, Onyeka Chrisanctus Ofodile, and hinonye Esther Ugochukwu. 2024. Automating financial reporting with natural language processing: A review and case analysis. *World Journal of Advanced Research and Reviews*, 21(3):575–589.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. BloombergGPT: A Large Language Model for Finance. ArXiv:2303.17564 [cs, q-fin].

Yi Yang, Mark Christopher Siy UY, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. ArXiv:2006.08097 [cs] version: 2.

Evaluating Multilingual Language Models for Cross-Lingual ESG Issue Identification

Wing Yan Li[1,2], Emmanuele Chersoni[2], Cindy Sing Bik Ngai[2]
University of Sussex[1], The Hong Kong Polytechnic University[2]
justina.li.106@gmail.com {emmanuele.chersoni,cindy.sb.ngai}@polyu.edu.hk

Abstract

The automation of information extraction from ESG reports has recently become a topic of increasing interest in the Natural Language Processing community. While such information is highly relevant for socially responsible investments, identifying the specific issues discussed in a corporate social responsibility report is one of the first steps in an information extraction pipeline. In this paper, we evaluate methods for tackling the Multilingual Environmental, Social and Governance (ESG) Issue Identification Task. Our experiments use existing datasets in English, French and Chinese with a unified label set. Leveraging multilingual language models, we compare two approaches that are commonly adopted for the given task: off-the-shelf and fine-tuning. We show that fine-tuning models end-to-end is more robust than off-the-shelf methods. Additionally, translating text into the same language has negligible performance benefits.

Keywords: ESG Reports, Pre-trained Language Models, Cross-lingual Transfer, Text Classification, Multilingual NLP

1. Introduction

Yearly releases of Environmental, Social and Governance (ESG) reports represent an important part of a financial company's life cycle. Such reports are used to guide the decisions of responsible investors, by guaranteeing that the company satisfies measurable and objective criteria that have a positive impact on society (Van Marrewijk, 2003; Sheehy and Farneti, 2021; Serafeim and Yoon, 2022). Complying with ESG practices is a requirement for corporations, for example, SEC filings in the United States have to follow the standard for Climate Change and Human Governance, and every European company providing investment products must disclose how its economic activity aligns with sustainability norms (Kang et al., 2022).

ESG reports address various issues and correspond to labels in internationally-defined standards[1]. Modern language models (LMs) can potentially play an important role in processing such reports by extracting ESG-relevant sections and automating the analysis of sustainability aspects emphasised by a company.

In this work, we propose a comprehensive evaluation of the task of multilingual ESG issue identification, using existing datasets that are written in English (EN), French (FR) and Chinese (ZH) (Chen et al., 2023a) but unifying them in a single label space and treating the labels as non-mutually exclusive (*multi-label* classification). We evaluate two commonly-used approaches to the task, namely off-the-shelf (embedding-based classification) and fine-tuning. In the former, a conventional classifier such as a Support Vector Machine (SVM), is trained on representations encoded by a pre-trained LM; in the latter, a pre-trained LM is fine-tuned end-to-end on the given task.

Using multilingual LMs, we compare the two aforementioned methods. Additionally, we test a translation-based approach by translating the FR and ZH datasets into English, the most resource-rich language. Our evaluation shows that fine-tuning is more robust than training with off-the-shelf representations, and that translation has a limited effect on model performance. We will also release our code and data, in order to allow other researchers to evaluate ESG issue identification systems in a unified multilingual setting[2].

2. Related Work

Recently, the Natural Language Processing (NLP) community has increased interest in automating the identification of issues in ESG reports where these issues are organised into taxonomies.

A dedicated workshop has been organized in conjunction with LREC 2022 (Wan and Huang, 2022), and related shared tasks are regular events in financial NLP workshops such as the FinNLP workshop series (Kang et al., 2022; Chen et al., 2023a,c). In particular, the organisers of the shared task co-located with the FinNLP IJCAI workshop 2023 have made available a multilingual dataset for English, French and Chinese. They

[1] https://www.msci.com/esg-and-climate-methodologies.

[2] https://github.com/justinaL/ML-ESG-Eval

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 50–58
20 May, 2024.

were annotated with labels defined on the basis of the MSCI ESG standard rating guidelines.

While the English and French datasets are fully comparable, the Chinese dataset includes additional labels and exhibits variations in the naming of the common label set. Moreover, in the Chinese dataset, the labels are not mutually exclusive, making it difficult to experiment with Chinese in a multilingual setting.

In the shared task, given the relatively limited size of the data, augmentation approaches relying on ChatGPT (OpenAI, 2022) to generate new instances were the most successful (Glenn et al., 2023), together with methods combining traditional classifiers (e.g. SVMs) and multilingual sentence representations (Linhares Pontes et al., 2023).

A recent and highly-relevant research trend in NLP involves the *domain adaptation* of LMs to specific domains. For instance, FinBERT models (Araci, 2019; Yang et al., 2020a) trained specifically on the financial domain and ESG-BERT models (Mukherjee, 2020; Mehra et al., 2022) trained on ESG reports in the sustainability investing field.

Essentially, these models further pre-train a general-purpose LM such as BERT (Devlin et al., 2019), on an in-domain corpus (e.g. ESG reports). Then, the domain-adapted LM is fine-tuned on the given issue identification task. Some of these models “inherit” a general-domain vocabulary from the original architecture, while others create a new in-domain vocabulary from scratch. This choice has been shown to significantly affect performance on several tasks (Peng et al., 2021, 2022).

However, current ESG-adapted models are limited to the English language. While translation is a common approach to re-adapt monolingual models to other languages, Mashkin and Chersoni (2023) showed that this approach is not significantly better than simpler classifier baselines.

3. Experiments

In this paper, we frame the Multilingual ESG Issue Identification (ML-ESG) task as a multi-label classification task. The task assigns instances (ESG-related news articles) to non-exclusive labels (ESG key issues categories), while not constraining the number of categories per instance. Given the multilinguality (EN, FR and ZH) of the datasets, the investigation is conducted with multilingual encoders where representations of various languages are mapped into a shared semantic space.

3.1. Dataset

The dataset is obtained from the ML-ESG task of FinNLP-2023, containing ESG-related news articles. According to Chen et al. (2023a), the articles were sourced from ESGToday[3] (EN), RSE-DATANEWS (FR)[4], Novethic (FR)[5] and ESG-BusinessToday (ZH)[6]. ESG-BusinessToday is a Taiwanese website, where every article is written in traditional Chinese. The articles are annotated by human experts following the MSCI ESG rating guidelines and are categorised into 35 pre-defined ESG key issues across three main topics: Environment, Social and Corporate Governance. Table 1 shows the sample size of each dataset split.

Language	Train	Test
EN	1199	300
FR	1200	300
ZH	653	131

Table 1: Sample size of dataset splits.

From the table, the ZH dataset is shown to possess the smallest sample size compared to EN and FR. During the annotation process of ZH articles, the SASB Standard[7] are merged with the original MSCI guidelines. As a result, there are extra labels in the original ZH dataset. We identify similarities between the ZH labels and those of the other two languages. Additionally, we re-analyse the set of labels of the ZH dataset, mapping missing labels to the corresponding ones in the shared set. For the labels without close correspondences, we discard the corresponding instances. Details on the label mappings are provided in Appendix B. Given that many of the ZH governance labels cannot be mapped to the shared set, the final dataset has unfortunately a limited number of governance-related labels. This is a problem that will have to be addressed by future studies, as governance labels are particularly relevant for Chinese ESG reports.

In the original task, the labels for the EN and FR datasets are mutually exclusive, while multiple labels can be assigned to the ZH instances. To facilitate cross-lingual learning, we unify the label space of all languages during data pre-processing. We treat each task as a multi-label classification by binarising the labels in every dataset. That is, given a dataset instance, the model has to carry out a binary classification for every possible label.

We focus on the actual multilingual identification of ESG issues, by unifying the task and dataset across languages. Our results are not directly comparable to Chen et al. (2023a) due to: i)

[3]https://www.esgtoday.com/category/esg-news/companies/.

[4]https://www.rsedatanews.net/.

[5]https://www.novethic.fr/actualite/environnement.html.

[6]5https://esg.businesstoday.com.tw/.

[7]https://sasb.org/standards/materiality-finder/?lang=en-us

after mapping the labels and filtering the instances, the ZH dataset is no longer the same; ii) the task on EN and FR is different. We do not assume the labels of the CSR reports to be mutually exclusive for the purpose of uniformity with the ZH data.

3.2. Implementation Details

Leveraging multilingual large LMs, we compare two popular approaches in tackling the 2023 ML-ESG shared task: off-the-shelf and fine-tuning.

Off-the-shelf. Representations are derived from the encoder and passed to a classifier for the issue identification task. We use Support Vector Machine (SVM) as the classifier. Since SVM is designed for binary classification, we utilise the `MultiOutputClassifier` from scikit-learn that fits one SVM per target, extending SVM to support multi-label classification.

Hyper-parameters of the SVM are optimised with Bayesian optimisation[8]. For the optimisation process, we apply a 5-fold stratified sampling and constrain the search space to the following hyper-parameters: `C, gamma, degree` and `kernel`.

Fine-tuning. While off-the-shelf approaches require less training data and parameters for optimisation, they often underutilise the model capacity of the encoders. To address this, we also fine-tune the encoder on the given task. Encoders are fine-tuned end-to-end with a classification layer stacked on top. The weights of encoder and stacked classifier are updated during training. Given the small dataset size, we utilize dropout to prevent over-fitting (Srivastava et al., 2014). Further training details are provided in Appendix A.

Translation. Given that LMs are typically trained on a larger share of English data, it may be advantageous to translate other languages to English before fitting the data. To analyse the impact of translation, we re-run our models with the two aforementioned approaches after translating the FR and ZH datasets to English using Google Translate[9] and DeepL Translate[10].

3.2.1. Encoders

We leverage the following encoders: Sentence-BERT (Reimers and Gurevych, 2019) with distilled multilingual Universal Sentence Encoder (Yang et al., 2020b) (SBERT-DUSE) as the base model, a pre-trained multilingual BERT (mBERT) (Devlin et al., 2019) and a multilingual E5 model (mE5) (Wang et al., 2024b).

SBERT-DUSE follows the SBERT framework by training DUSE on the Stanford Natural Language Inference Corpus (SNLI) (Bowman et al., 2015) and Multi-Genre Natural Language Inference Corpus (MultiNLI) (Williams et al., 2018) for better sentence representations. The sentence encodings are obtained by mean pooling of all the vectors of the final layer.

mBERT shares the same structure as BERT with 12 transformer encoder layers in the base version. The model is pre-trained on Wikipedia pages of 104 languages instead of monolingual English data only. mBERT uses the same pre-training objectives as BERT, namely masked language modelling (MLM) and next sentence prediction (NSP).

mE5 is a variation of the E5 model with XLM-R (Conneau et al., 2020) as the base model. E5 is a general-purpose encoder that aims to yield robust off-the-shelf representations in both zero-shot or fine-tuned settings (Wang et al., 2022). Following the training recipe of the English E5, mE5 is trained using a contrastive loss with weak supervision, leveraging data from Wang et al. (2024a).

For the off-the-shelf approach, sentence representations of mBERT and mE5 are mean pooled vectors of the final layer as done by SBERT.

3.2.2. Evaluation Metric

Macro-F1 scores are used as the performance metric. Given the highly imbalanced classes, the macro score treats each class equally regardless of the number of samples. Thus, the model has to perform well in both majority and minority classes. The class distribution is plotted in Appendix B.

4. Results

Table 2 and 3 are the results with and without translation applied. *EU Lang.* refers to training data including EN and FR only; *All Lang.* indicates that training data from all languages are used. Compared to the results of the original shared task (Chen et al., 2023a), the scores for EN and FR are lower, but this is not surprising, since we are working in a multi-label classification setting.

In Table 2, a first noticeable trend is that models using *All Lang.* have significant improvements on ZH compared to those using *EU Lang.* only. While performance on EN and FR drop observably for SBERT-DUSE and mE5, this is not the case for mBERT, which shows more robust performance. This suggests that using multilingual data is, as expected, very helpful for languages in a low-resource setting for this task. However, the per-

[8] https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html

[9] https://translate.google.com/

[10] https://www.deepl.com/en/docs-api

Encoder	Lang.	EN	FR	ZH
SBERT-DUSE	EU	**0.52**	**0.67**	0.06
	All	0.45	0.59	**0.18**
mBERT	EU	**0.48**	**0.47**	0.04
	All	**0.48**	0.46	**0.22**
mE5	EU	**0.53**	**0.60**	0.09
	All	0.48	0.54	**0.20**

(a) Off-the-shelf approach. SVM as the classifier with Bayesian optimisation.

Encoder	Lang.	EN	FR	ZH
SBERT-DUSE	EU	**0.49**	**0.59**	0.01
	All	0.44	0.53	**0.19**
mBERT	EU	**0.55**	0.64	0.05
	All	**0.55**	**0.66**	**0.23**
mE5	EU	0.56	**0.67**	0.09
	All	**0.58**	**0.67**	**0.28**

(b) Fine-tuning approach. Attention dropout for regularisation.

Table 2: Macro-F1 on ML-ESG per language (average across 3 seeds). **No translation** is applied. Best performance per model is highlighted in **bold**.

Encoder	Translator	EN	FR	ZH
SBERT-DUSE	Google	0.48	0.57	**0.16**
	DeepL	**0.51**	**0.61**	**0.16**
mBERT	Google	0.44	**0.47**	**0.24**
	DeepL	**0.45**	**0.47**	0.23
mE5	Google	0.45	**0.50**	0.18
	DeepL	**0.46**	**0.50**	**0.21**

(a) Off-the-shelf approach on translated text. SVM as the classifier with Bayesian optimisation.

Encoder	Translator	EN	FR	ZH
SBERT-DUSE	Google	**0.49**	**0.58**	0.17
	DeepL	0.47	0.55	**0.20**
mBERT	Google	**0.55**	**0.65**	0.22
	DeepL	**0.55**	0.60	**0.23**
mE5	Google	0.55	0.65	**0.26**
	DeepL	**0.58**	**0.68**	**0.26**

(b) Fine-tuning approach on translated text. Attention dropout for regularisation.

Table 3: Macro-F1 on ML-ESG per language (average across 3 seeds). All inputs are **translated** to English, all models are trained with ***All Lang.***. Best performance per model is highlighted in **bold**.

formance can be detrimental for higher-resource ones, especially in cases where the training data mix languages that have deep typological differences, as in the case of Chinese and the two European languages.

While mBERT performs more stably across the board, the overall performance is slightly lower than the other models. Plausibly, this is due to both SBERT-DUSE and mE5 having taken advantage of their extensive training and their exposure to more training data compared to a standard pre-trained mBERT.

Table 2b also shows that, despite the limited size of our training data, fine-tuning models end-to-end tends to yield better performance than the off-the-shelf approach. Fine-tuning modifies representations to be more task-specific, in contrast to the off-the-shelf approach where the encoder representation space remains static throughout the training process. Finally, it can be noticed that mE5 achieves the top overall performances after fine-tuning, with a marked improvement on ZH compared to the competitors.

Table 3 shows the results using the translated text of *All Lang.* for training. One would hypothesize that the task gets easier after translation as the models have to handle a single language only. Yet, this step often exhibits insignificant or even detrimental effects.

Also with translation, performance remains generally higher for the fine-tuning approach. This highlights the robustness of the feature learning with this technique. Once again, mE5 is the model achieving the overall highest scores for all the three languages as shown in Table 3b. Google Translate and DeepL Translator demonstrate comparable performance, regardless of the encoder utilised. Despite the slight bias towards DeepL translations in the off-the-shelf setting, the choice of the translator should be subject to the specific task and target language.

5. Conclusion

In this work, we evaluate methods for tackling the Multilingual ESG Issue Identification. To facilitate cross-lingual learning, we have modified the ML-ESG dataset (Chen et al., 2023a) and unified the sets of labels across languages. Moreover, the evaluation is carried out in a multi-label, non-exclusive classification setting, in order to make the task in English and French similar to Chinese. In our view, the multi-label setting allows for a more natural evaluation of this task, since in real-world ESG reports often cover more than one issue.

We have also studied the differences between the off-the-shelf and fine-tuning approaches. The latter consistently outperformed the former on multilingual and translated datasets, demonstrating its advantage of learning task-specific features. Furthermore, translation has minimal impact on both methods, suggesting that it may be an optional step for the given task.

Acknowledgements

The authors acknowledge the support from the project "Analyzing the semantics of Transformers representations for financial natural language processing" (ZVYU), sponsored by the Faculty of Humanities of the Hong Kong Polytechnic University.

The authors would like to thank Ivan Mashkin for his help in data preparation, and the two anonymous reviewers for their constructive feedback. Also, we would like to thank Leonidas Gee for proofreading the paper.

6. Bibliographical References

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *arXiv preprint arXiv:1908.10063*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *Proceedings of EMNLP*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. Multi-Lingual ESG Issue Identification. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023c. Multi-Lingual ESG Impact Type Identification. In *Proceedings of the IJCNLP-AACL Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using Synthetic Data and Transfer Learning for Multilingual ESG Issue Classification. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Juyeon Kang, Mehdi Kchouk, Sandra Bellato, Mei Gan, and Ismail El Maarouf. 2022. FinSim4-ESG Shared Task: Learning Semantic Similarities for the Financial Domain. Extended Edition to ESG insights. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing*.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging BERT Language Models for Multi-lingual ESG Issue Identification. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Lu Lu, Jinghang Gu, and Chu-Ren Huang. 2022. Inclusion in CSR Reports: The Lens from a Data-driven Machine Learning Model. In *Proceedings of the LREC Workshop on Computing Social Responsibility*.

Ivan Mashkin and Emmanuele Chersoni. 2023. HKESG at the ML-ESG Task: Exploring Transformer Representations for Multilingual ESG Issue Identification. In *Proceedings of the IJCAI Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. *arXiv preprint arXiv:2203.16788*.

Mukut Mukherjee. 2020. ESG-BERT: NLP Meets Sustainable Investing. *Towards Data Science Blog*.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is Domain Adaptation Worth your Investment? Comparing BERT and FinBERT on Financial Tasks. In *Proceedings of the EMNLP Workshop on Economics and Natural Language Processing*.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2022. Discovering Financial Hypernyms by Prompting Masked Language Models. In *Proceedings of the LREC Workshop on Financial Narrative Processing Workshop*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings Using Siamese BERT-networks. In *Proceedings of EMNLP*.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of EMNLP*.

George Serafeim and Aaron Yoon. 2022. Stock Price Reactions to ESG News: The Role of ESG Ratings and Disagreement. *Review of Accounting Studies*, pages 1–31.

Benedict Sheehy and Federica Farneti. 2021. Corporate Social Responsibility, Sustainability, Sustainable Development and Corporate Sustainability: What Is the Difference, and Does It Matter? *Sustainability*, 13(11):5965.

US SIF. 2020. Sustainable Investing Basics. *Available on https://www. ussif. org/sribasics (last accessed on 20 May 2022)*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*.

Marcel Van Marrewijk. 2003. Concepts and Definitions of CSR and Corporate Sustainability: Between Agency and Communion. *Journal of Business Ethics*, 44(2-3):95–105.

Mingyu Wan and Chu-Ren Huang. 2022. Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference. In *Proceedings of the LREC Workshop on Computing Social Responsibility*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving Text Embeddings with Large Language Models. *arXiv preprint arXiv:2401.00368*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *Proceedings of NAACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of EMNLP: System Demonstrations*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020a. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2020b. Multilingual Universal Sentence Encoder for Semantic Retrieval. In *Proceedings of ACL: System Demonstrations*.

A. Training Details

We implement the models using the Hugging Face `transformers` library (Wolf et al., 2020) for fine-tuning models end-to-end. We use a batch size of 16 and learning rate of $2e-5$. We carry out a grid search on the probability value for attention dropout, $p \in$ {0.1, 0.2, 0.3}. In Table 4, the best attention dropout value per model is highlighted in bold, these values are found using *All Lang.* and are applied to the experiments on the translated datasets. Results reported are from the corresponding `best_model`, where we define the `best_model_metric` as the macro-F1 score.

Encoder	p	EN	FR	ZH
SBERT-DUSE	**0.1**	0.44	0.53	0.19
	0.2	0.43	0.53	0.18
	0.3	0.42	0.50	0.18
mBERT	0.1	0.54	0.66	0.23
	0.2	0.55	0.66	0.23
	0.3	0.55	0.64	0.23
mE5	0.1	0.58	0.66	0.28
	0.2	0.57	0.67	0.27
	0.3	0.58	0.67	0.28

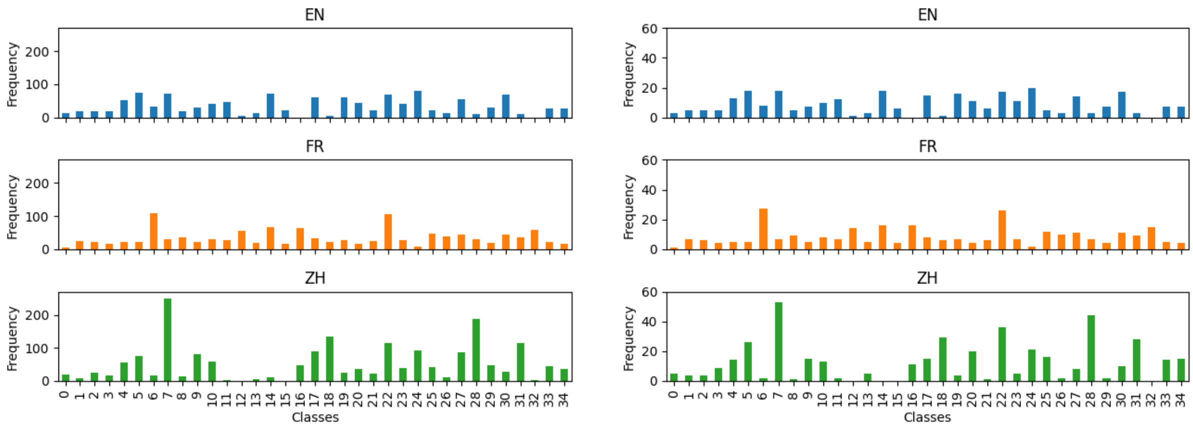
Table 4: Fine-tune models end-to-end with different probability values (p) for attention dropout. 0.1 is the default value. Models are trained with *All Lang.*. The best attention dropout value per model is highlighted in **bold**.

B. Dataset Details

Figure 1 shows the plots of the class distribution of the training and test sets per language (EN, FR and ZH). As ZH instances have multiple labels, the total number of counts is higher than EN and FR. Table 5 provides the labels of ESG key issues. Table 6 list the mappings of original ZH labels to the unified label space across the languages.

Index	Label
0	Access to Communications
1	Access to Finance
2	Access to Health Care
3	Accounting
4	Biodiversity & Land Use
5	Board
6	Business Ethics
7	Carbon Emissions
8	Chemical Safety
9	Climate Change Vulnerability
10	Community Relations
11	Consumer Financial Protection
12	Controversial Sourcing
13	Electronic Waste
14	Financing Environmental Impact
15	Health & Demographic Risk
16	Health & Safety
17	Human Capital Development
18	Labor Management
19	Opportunities in Clean Tech
20	Opportunities in Green Building
21	Opportunities in Nutrition & Health
22	Opportunities in Renewable Energy
23	Ownership & Control
24	Packaging Material & Waste
15	Pay
26	Privacy & Data Security
27	Product Carbon Footprint
28	Product Safety & Quality
29	Raw Material Sourcing
30	Responsible Investment
31	Supply Chain Labor Standards
32	Tax Transparency
33	Toxic Emissions & Waste
34	Water Stress

Table 5: Labels of ESG key issues.



(a) Training set. (b) Test set.

Figure 1: Class distribution of the EN, FR and ZH datasets.

	Original ZH Label	New Label
S12	產品責任 | 銷售模式和產品標示 (Selling Practices & Product Labeling)	Product Safety & Quality
S13	產品責任 | 產品設計與生命週期管 (Product Design & Lifecycle Management)	Product Safety & Quality
S14	產品責任 | 供應鏈管理 (Supply Chain Management)	Supply Chain Labor Standards
G11	公司行為 | 競爭行為 (Competitive Behavior)	None
G05	公司治理 | 重大事件風險管理 (Critical Incident Risk Management)	None
G08	公司治理 | 商業模式靈活度 (Business Model Resilience)	None
G06	公司治理 | 風險管理系統 (Systemic Risk Management)	None
G06	公司治理 | 風險管理系統 (Systemic Risk Management)	None
S05	人力資源 | 人權與社區關係 (Human Rights & Community Relations)	Community Relations
S11	產品責任 | 健康與人口風險 (Insuring Health & Demographic Risk)	Access to Health Care
E06	自然資源 | 原材料採購 (Raw Material Sourcing)	Raw Material Sourcing
S03	人力資源 | 人力資本發展 (Human Capital Development)	Human Capital Development
S19	社會機會 | 衛生保健管道 (Access to Health Care)	Access to Health Care
E11	環境機會 | 可再生能源的機會 (Opportunities in Renewable Energy)	Opportunities in Renewable Energy
S17	社會機會 | 溝通管道 (Access to Communication)	Access to Communication
E13	環境機會 | 綠色建造的機會 (Opportunities in Green Building)	Opportunities in Green Building
S08	產品責任 | 責任投資 (Responsible Investment)	Responsible Investment
G10	公司行為 | 納稅透明度 (Tax Transparency)	Tax Transparency
G07	公司治理 | 法律和法規環境的管理 (Management of the Legal & Regulatory Environment)	Management of the Legal & Regulatory Environment
S10	產品責任 | 金融產品安全性 (Consumer Financial Protection)	Consumer Financial Protection
S15	股東否決權 | 有爭議的採購 (Controversial Sourcing	Controversial Sourcing
S20	社會機會 | 營養與健康的機會 (Opportunities in Nutrition & Health)	Opportunities in Nutrition & Health

Table 6: Mapping of labels from the original ZH dataset to the unified label space with EN and FR.

	Original ZH Label	New Label
E12	環境機會 | 清潔技術的機會 (Opportunities in Clean Tech)	Opportunities in Clean Tech
G01	公司治理 | 董事會 (Board)	Board
E10	汙染與浪費 | 用於包裝的材料及浪費 (Packaging Material & Waste)	Packaging Material & Waste
S01	人力資源 | 人力資源管理 (Labor Management)	Labor Management
E02	氣候變化 | 產品碳足跡 (Product Carbon Footprint)	Product Carbon Footprint
G04	公司治理 | 會計 (Accounting)	Accounting
E07	自然資源 | 生物多樣性與土地利用 (Biodiversity & Land Use)	Biodiversity & Land Use
S02	人力資源 | 員工健康和安全 (Health & Safety)	Health & Safety
S04	人力資源 | 供應鏈勞動標準 (Supply Chain Labor Standards)	Supply Chain Labor Standards
E01	氣候變化 | 碳排放量 (Carbon Emissions)	Carbon Emissions
S09	產品責任 | 產品安全與品質 (Product Safety & Quality)	Product Safety & Quality
G03	公司治理 | 所有權 (Ownership & Control)	Ownership & Control
E04	氣候變化 | 氣候變化脆弱性 (Climate Change Vulnerability)	Climate Change Vulnerability
E03	氣候變化 | 融資環境影響 (Financing Environment Impact)	Financing Environment Impact
S18	社會機會 | 融資管道 (Access to Finance)	Access to Finance
E09	汙染與浪費 | 電子廢物 (Electronic Waste)	Electronic Waste
G09	公司行為 | 商業道德 (Business Ethics)	Business Ethics
S16	股東否決權 | 社區關係 (Community Relations)	Community Relations
E08	汙染與浪費 | 有毒物排放及浪費 (Toxic Emissions & Waste)	Toxic Emissions & Waste
S06	產品責任 | 化學物質安全性 (Chemical Safety)	Chemical Safety
S07	產品責任 | 隱私和數據安全 (Privacy & Data Security)	Privacy & Data Security
E05	自然資源 | 水資源壓力 (Water Stress)	Water Stress
G02	公司治理 | 薪酬 (Pay)	Pay
	Not related to ESG	None

Table 6: Mapping of labels from the original ZH dataset to the unified label space with EN and FR (continued).

Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM

Kun Ouyang[1], **Yi Liu**[1], **Shicheng Li**[1], **Ruihan Bao**[2], **Keiko Harimoto**[2], **Xu Sun**[1]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2]Mizuho Securities Co., Ltd.
kunouyang10@gmail.com, {imliuyi, lisc99}@pku.edu.cn,
{ruihan.bao, keiko.harimoto}@mizuho-sc.com, xusun@pku.edu.cn

Abstract

Financial prediction from Monetary Policy Conference (MPC) calls is a new yet challenging task, which targets at predicting the price movement and volatility for specific financial assets by analyzing multimodal information including text, video, and audio. Although the existing work has achieved great success using cross-modal transformer blocks, it overlooks the potential external financial knowledge, the varying contributions of different modalities to financial prediction, as well as the innate relations among different financial assets. To tackle these limitations, we propose a novel **M**odal-**A**daptive k**N**owledge-enh**A**nced **G**raph-bas**E**d financial p**R**ediction scheme, named MANAGER. Specifically, MANAGER resorts to FinDKG to obtain the external related knowledge for the input text. Meanwhile, MANAGER adopts BEiT-3 and Hidden-unit BERT (HuBERT) to extract the video and audio features, respectively. Thereafter, MANAGER introduces a novel knowledge-enhanced cross-modal graph that fully characterizes the semantic relations among text, external knowledge, video and audio, to adaptively utilize the information in different modalities, with ChatGLM2 as the backbone. Extensive experiments on a publicly available dataset Monopoly verify the superiority of our model over cutting-edge methods.

Keywords: Financial Prediction, LLM, Multimodal Learning

1. Introduction

Forecasting the fluctuation of prices for a financial asset over a specific period is a crucial task in financial analysis, essential for both investors and policymakers (Lewellen, 2003). Accurate prediction results can assist investors in making informed decisions regarding investment returns, while policymakers can implement prudent monetary measures to uphold a robust economy (Cai et al., 2021; Shapiro and Wilson, 2019). In early work, researchers made efforts to solve financial prediction for textual financial data, such as BloombergGPT (Wu et al., 2023b) and FinGPT (Wang et al., 2023b).

Despite their promising performance, the above models can only solve text-based financial tasks. With unprecedented advances in multimodal learning, investors now have access to a vast amount of unstructured data for financial prediction (Jiang, 2020). Moreover, the non-verbal information involved in the visual and acoustical modalities (e.g., vocal tone and facial expressions) can be indicative and correlated with trading activities in the financial market. One such abundant source of multimodal information is the Monetary Policy Conference (MPC's) call. Previous work (Boukus and Rosenberg, 2006) has underscored the influence of MPC calls on financial stock markets. Therefore, Mathur et al. (2022) curated a public Monetary Policy Call Dataset named Monopoly and proposed to predict the price movement and volatility for six principal financial assets (i.e., Stock In-

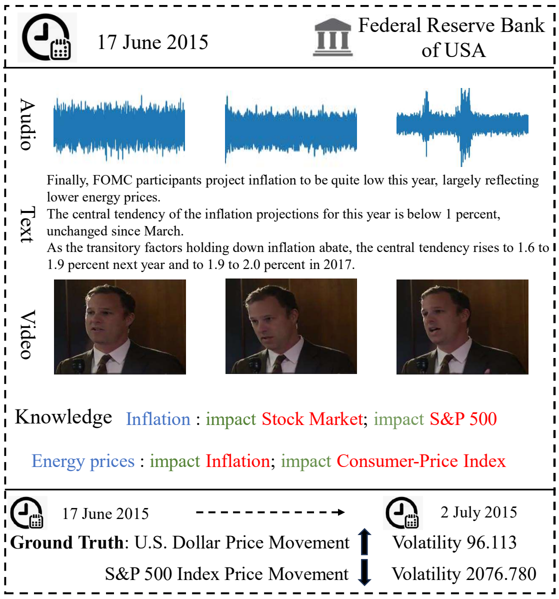


Figure 1: An example of the financial prediction from MPC calls. We also present the external knowledge inferred by FinDKG for the given text. Notably, the words in blue are the anchor entities while those in green are the relations and those in red are the related entities.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 59–69
20 May, 2024.

dex (Small), Stock Index (Large), Gold Price, Currency Exchange Rate, Long-term bond yield (10-years), Short-term bond yield (3-months)) based on MPC calls. The authors adopted cross-modal transformer blocks and modality-specific attention fusion to conduct price movement and volatility prediction. Although this pioneering study has achieved promising performance, it still suffers from three key limitations.

1) **Overlook the potential external knowledge in the financial domain.** The pioneering study fails to utilize the related knowledge contained in the external public knowledge base in the financial domain. As shown in Figure 1, the related knowledge obtained from FinDKG (Li, 2023) can strengthen the context comprehension (e.g., "impact S&P 500") and promote the financial prediction.

2) **Overlook the varying contributions of different modalities to financial prediction.** The existing work equally feeds the multimodal features (i.e., text, video and audio) into the model, and treats them as the equal source of information to conduct multimodal information fusion with the same weights. In fact, the content of given text is the prime cue for the financial prediction, while the non-verbal cues such as facial expressions and vocal tone involved in the video and audio play a minor role in comprehending the context. How to adaptively utilize the information residing in the multiple modalities merits our attention.

3) **Overlook the innate relations among different financial assets.** The former method predicts the price movement and volatility of six financial assets independently, ignoring the potential relationships among different financial assets. Actually, the price changes of a financial asset may provide useful information to predict price trend of the other financial assets.

To tackle these limitations, we propose a novel **M**odal-**A**daptive k**N**owledge-enh**A**nced **G**raph-bas**E**d financial p**R**ediction scheme, MANAGER for short. In detail, MANAGER consists of four components: external financial knowledge acquisition, video-audio feature extraction, knowledge-enhanced modal-adaptive context comprehension and task-specific instruction tuning for financial prediction, as shown in Figure 2. In the first module, we focus on acquiring the external related knowledge for the given text, where a large-scale financial knowledge base FinDKG (Li, 2023) is used. In the second module, we utilize BEiT-3 (Wang et al., 2022) and Hidden-Unit BERT (HuBERT) (Hsu et al., 2021) to extract the video and audio representations, respectively. In the third module, we construct the knowledge-enhanced cross-modal graph to aggregate the given text, input video, audio and inferred external knowledge through two types of relations (i.e., intra-modal and inter-modal semantic relations). We then employ the commonly used graph neural networks (GCNs) (Kipf and Welling, 2017), which have shown great performance in NLP tasks (Jing et al., 2023; Ouyang et al., 2024), to adaptively utilize the different modalities for cross-modal context comprehension. In the last module, considering that up-to-date Large Language Models (LLMs) have shown promising performance in multimodal context learning (Zhang and Li, 2023; Wu et al., 2023a), the potential of LLMs in the multimodal financial prediction task is increasingly evident. Therefore, we adopt ChatGLM2 (Du et al., 2022) as the backbone and feed the cross-modal representation into ChatGLM2 with a task-specific instruction devised for the certain financial asset to predict the price movement or volatility, respectively. Unlike previous work, we do not conduct prediction for different financial assets independently, but utilize ChatGLM2 to capture the innate relation among the financial assets. Finally, we conduct extensive experiments on a publicly available Monopoly dataset, on which our method outperforms the best baseline across all the metrics for both price movement and volatility prediction. Our contributions can be concluded as follows.

- We propose a novel modal-adaptive knowledge-enhanced graph-based financial prediction scheme, where the text, external knowledge, video and audio are aggregated for cross-modal context comprehension.
- As far as we know, we are the first to introduce an up-to-date LLM named ChatGLM2 to solve the financial prediction task for Monetary Policy Conference (MPC) calls, which contain multiple modalities (i.e., text, video and audio).
- The results of extensive experiments on the Monopoly dataset demonstrate the superiority of our MANAGER over other cutting-edge methods, and prove the effectiveness of each component of MANAGER. As a byproduct, we release our code and parameters[1] to facilitate the research community.

2. Related Work

2.1. Large Language Models (LLMs) in Finance

In early work about the application of LLMs in Finance, researchers resorted to BERT (Devlin et al., 2019) to conduct financial tasks, such as FinBert (Liu et al., 2020), which is dedicated for financial sentiment analysis with under one billion

[1] https://github.com/OuyangKun10/MANAGER.

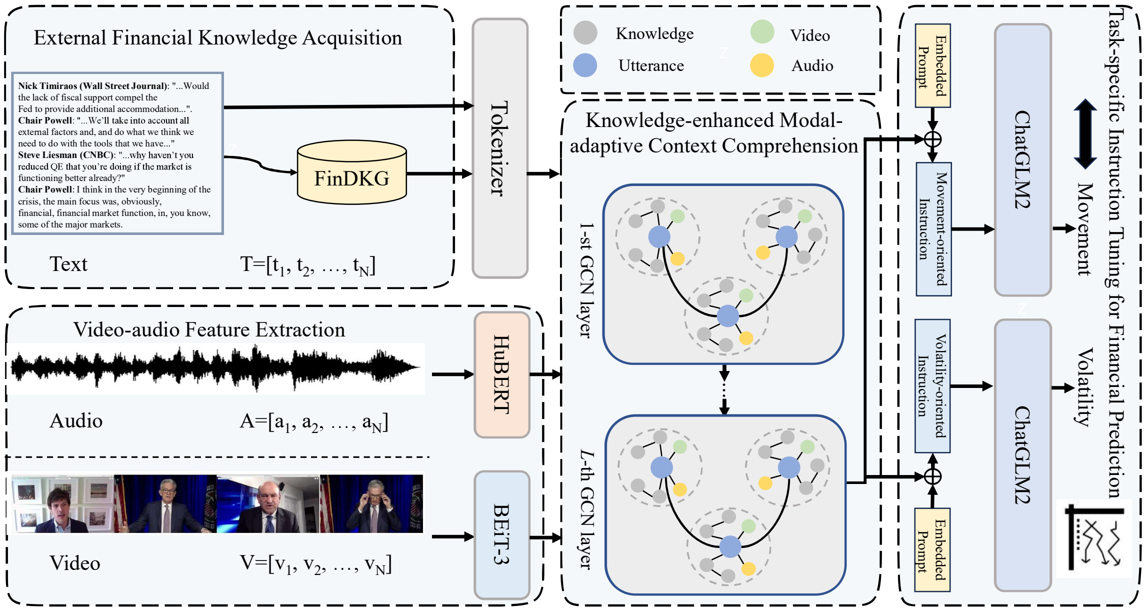


Figure 2: The architecture of MANAGER, which consists of four key components including External Financial Knowledge Acquisition, Video-audio Feature Extraction, Knowledge-enhanced Modal-adaptive Context Comprehension and Task-specific Instruction Tuning for Financial Prediction.

parameters, fine-tuned on a rich financial corpus to excel in finance-specific tasks. Although it achieves promising performance, it falls short of comprehending the long and complex financial text. In recent years, there has been a surge in research dedicated to integrating financial datasets with GPT-based models (Brown et al., 2020), aimed at enhancing Natural Language Processing (NLP) applications. For example, BloombergGPT (Wu et al., 2023b) is a closed-source model, trained extensively on diverse financial datasets, thereby encapsulating a broad spectrum of the financial domain. FinGPT (Wang et al., 2023b) is an open-source LLM, fine-tuned from a general LLM using low-rank adaptation method (Hu et al., 2021), fostering accessibility for the broader community.

2.2. Multimodal Financial Prediction

Existing work in the financial realm utilize vocal and textual cues from earnings conference calls (Qin and Yang, 2019; Sawhney et al., 2020), and mergers and acquisitions calls (Sawhney et al., 2021) for stock volatility prediction. Multimodal architectures that use these cues for financial predictions have seen significant improvements in their performances (Sawhney et al., 2020; Yang et al., 2020). However, the vision modality, which may offer important cues that correlate with the performance of financial markets (Cao, 2021) remains underexplored. Therefore, Mathur et al. (2022) first introduced video modality in the financial prediction task and released a dataset named Monopoly. They adopted cross-modal transformer blocks and modality-specific attention fusion to forecast the financial risk and price movement. Despite its promising performance on financial prediction, it overlooks the potential external knowledge, the varying contributions of different modalities, the innate relations among different financial assets, which are the major concerns of our model.

3. Task Formulation

Suppose we have a training dataset $\mathcal{D}$ composed of N_d samples, i.e., $\mathcal{D} = \{d_1, d_2, \cdots, d_{N_d}\}$. Each sample $d_i = \{T_i, V_i, A_i, Y_i\}$, where $T_i = \{u_1^i, u_2^i, \cdots u_N^i\}$ denotes the input text containing N utterances, $V_i = \{v_1^i, v_2^i, \cdots v_N^i\}$ and $A_i = \{a_1^i, a_2^i, \cdots a_N^i\}$ are the set of the input video and audio clips, respectively. Each utterance u_i contains N_{u_i} tokens. i.e., $u_i = \{t_1^i, t_2^i, \cdots, t_{N_{u_i}}^i\}$ And $Y_i^\tau = \{p_i^\tau, o_i^\tau\}$ denotes the target labels over a period of τ days, where p_i^τ is the price movement and o_i^τ is the volatility, respectively. Our target is to learn a multimodal financial prediction model $\mathcal{F}$ that is able to predict the price movement and volatility for six principal financial assets (i.e., Stock Index (Small), Stock Index (Large), Gold Price, Currency Exchange Rate, Long-term bond yield (10-years), Short-term bond yield (3-months)), based on the given multimodal input as follows,

$$\hat{Y}_i = \mathcal{F}(T_i, V_i, A_i|\Theta) \tag{1}$$

where Θ is a set of learnable parameters of the model $\mathcal{F}$. $\hat{Y}_i = \{\hat{p}_i^\tau, \hat{o}_i^\tau\}$ is the labels (i.e., price movement and volatility) predicted by $\mathcal{F}$. For simplicity, we omit the subscript i that indexes the training samples.

4. Method

In this section, we detail the four components of our proposed MANAGER, as shown in Figure 2.

4.1. External Financial Knowledge Acquisition

As aforementioned, the external financial knowledge inferred by the input text can assist the financial prediction since it can introduce corresponding financial entities as well as the relations, and provide some external factors to analyze the financial environment, leading to more informed predictions. Specifically, we resort to FinDKG (Li, 2023), which provides dynamic knowledge graph data in the financial domain, as the source of external knowledge. Notably, FinDKG changes dynamically over time. In detail, it contains $13,645$ financial entities and 15 types of relations. Given the input text, we adopt the period-specific FinDKG[2] that only contains the knowledge before the date of the input text, to prevent our model from obtaining the information beyond the date. The ration is that the information beyond the date can influence the prediction.

To acquire the related external knowledge for the given text, i.e., $T = \{u_1, u_2, \cdots u_N\}$, we first identify all the entities in FinDKG that exist in the input text. Let $\{e_1, \cdots, e_{N_e}\}$ be the set of identified entities, where N_e is the total number of the identified entities. We then use these identified entities as the anchors to obtain the related entities and corresponding relations as the external knowledge for the input text. Notably, for each anchor entity e, we retrieve all its one-hop neighboring entities, as well as the corresponding relations that are treated as the edges, from FinDKG and deem them as the external knowledge for e. Mathematically, let $\mathcal{N}(e) = \mathcal{N}^1(e) \cup \mathcal{N}^2(r)$ be the set of external knowledge (i.e., $\mathcal{N}^1(e)$ is the set of neighboring entities and $\mathcal{N}^2(r)$ is the set of corresponding relations between each neighboring entity and the anchor entity, respectively.) of the entity e in FinDKG. Then the related external knowledge for the input text can be represented as $\{\mathcal{N}^1_{e_1}, \mathcal{N}^1_{e_2}, \cdots, \mathcal{N}^1_{e_{N_e}}\} \cup \{\mathcal{N}^2_{e_1}, \mathcal{N}^2_{e_2}, \cdots, \mathcal{N}^2_{e_{N_e}}\}$. N_e is the number of the neighboring entities as well as the number of the relations.

[2] https://xiaohui-victor-li.github.io/FinDKG/.

4.2. Video-audio Feature Extraction

To obtain the global feature of video and audio clips, we choose BEiT-3 (Wang et al., 2022) and Hidden-Unit BERT (HuBERT) (Hsu et al., 2021) as the visual and acoustical encoder, respectively.

Video Encoding, to encode the video clips, we resort to BEiT-3, which is an advanced general-purpose multimodal foundation model pre-training for all vision and vision-language tasks and shows great performance in visual modality encoding (Wang et al., 2022). Specifically, we embed each frame v_j^k in the video clip v_j as the arithmetic mean of visual tokens representations of that frame. We then average over all the frames to obtain the aggregated encoding feature $x_V^j \in \mathbb{R}^D$, where D is the feature dimension. Mathematically, we have

$$x_V^j = \frac{1}{N_f} \sum_{k=1}^{N_f} \mathrm{BEiT\text{-}3}(v_j^k), \forall j \in [1, N], \qquad (2)$$

where N_f is the number of frames in the clips v_j. And we represent the sequence of video features as $X_V = [x_v^1, x_v^2, \cdots, x_v^N]$.

Audio Encoding, we extract the feature of the audio clips via the self-supervised speech representation model named HuBERT, which has shown significant power for extracting audio features for speech language understanding tasks (Yoon et al., 2022). We embed each audio utterance a_j^k in the audio clip a_j as the arithmetic mean of the representation derived by HuBERT and obtain the encoded acoustical feature $x_A^j \in \mathbb{R}^D$. Formally,

$$x_A^j = \mathrm{HuBERT}(a_j^k), \forall j \in [1, N], \qquad (3)$$

we represent the sequence of audio features as $X_A = [x_a^1, x_a^2, \cdots, x_a^N]$.

4.3. Knowledge-enhanced Modal-adaptive Context Comprehension

In this module, we aim to enhance the cross-modal context comprehension utilizing the inferred external knowledge in the financial domain. In fact, there are rich relations (i.e., intra-modal semantic relation and inter-modal semantic relation) existing in the multiple input including the text, video, audio and external knowledge. Therefore, to adaptively utilize different modalities via these semantic relations for boosting cross-modal context comprehension, we resort to the widely used graph neural networks (GCNs). Specifically, we first build a knowledge-enhanced cross-modal graph $\mathcal{G}$.

4.3.1. Nodes Initialization

In particular, the nodes in the knowledge-enhanced cross-modal graph $\mathcal{G}$ come from four kinds of sources, the given text T, input video clips V, input audio clips A and inferred external knowledge $\mathcal{N}(e)$. We define all the nodes as $\{n_1, \cdots, n_{N_c}\} = \{T, \mathcal{N}(e), V, A\}$, where N_c is the total number of nodes. To initialize the nodes, we feed the textual input $\{T, \mathcal{N}(e)\}$ into the encoder of ChatGLM2 (Du et al., 2022) to extract their features. Specifically, we first concatenate them into a sequence of tokens, denoted as $X_T = \{T, \mathcal{N}(e)\}$, and then feed X into the encoder $\mathcal{E}$ as follows,

$$\mathbf{H} = \mathcal{E}(X_T), \tag{4}$$

where $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_{N_t+2\times N_e}] \in \mathbb{R}^{(N_t+2\times N_e)\times D}$ is the encoded representation matrix, N_t is the tokens number of the whole utterances and each column of which corresponds to a token. Accordingly, nodes of the textual part (utterances and external knowledge) in the knowledge-enhanced cross-modal graph $\mathcal{G}$ can be initialized by $\mathbf{H}$, where the j-th token node is initialized with $\mathbf{h}_j$. In addition, the other nodes are initialized by the extracted video feature sequence X_V and the extracted audio feature sequence X_A, respectively.

4.3.2. Semantic Relation Construction

To enhance the cross-modal context comprehension with related external knowledge, we consider two kinds of semantic relations: intra-modal semantic relation and inter-modal semantic relation. The former captures the basic information flow of the multiple modalities input, also incorporates the related external knowledge into the text. The latter enables the injection of non-textual information from video and audio into the context and achieves cross-modal information fusion.

Intra-modal Semantic Relation. To capture the information flow in the specific modality, we design three types of intra-modal semantic edges. a) *Token-token edges*. We introduce an edge between each pair of adjacent nodes in given text to represent the neighboring relations among the tokens of text. b) *Token-knowledge edge*. We connect the tokens that act as an anchor entity in the aforementioned external knowledge retrieval process, relation token and the related entity token sequentially. c) *Video-video edge* and d) *Audio-audio edge*. We link each pair of adjacent video nodes and connect each pair of adjacent audio nodes to represent the adjacency relations of the video and audio modalities, respectively. The above edges are characteristics of the information flow, and weighted by 1. Formally, we introduce the corresponding adjacency matrix $\mathbf{A}^1$ for representing these edges as follows,

$$\mathbf{A}^1_{i,j} = \begin{cases} 1, & if\ D_1(n_i, n_j), \\ 0, & otherwise, \end{cases} \tag{5}$$

where N_c denotes the total number of nodes in $\mathcal{G}$ and $i, j \in [1, N_c]$. $D_1(n_i, n_j)$ denotes that the nodes n_i and n_j have the certain above defined intra-modal semantic relation.

Inter-modal Semantic Relation. To comprehensively utilize the multiple modalities to promote cross-modal context comprehension, we devise two types of inter-modal semantic edges. a) *Token-video edges*. For each video node, we connect it to each token in the corresponding utterance. The ration is to inject the visual information (e.g., facial expressions and hand gestures) that can help semantics understanding and hence improve financial analysis, into the context. b) *Token-audio edges*. For each audio node, we link it with each token in the corresponding utterance. In this way, we can incorporate the acoustic information (e.g., vocal tone) that is also useful for context comprehension, into the context. The weight of all the above edges is set to 1. Accordingly, the adjacency matrix $\mathbf{A}^2 \in \mathbb{R}^{N_c\times N_c}$ for capturing the above inter-modal semantic relations can be constructed as follows,

$$\mathbf{A}^2_{i,j} = \begin{cases} 1, & if\ D_2(n_i, n_j), \\ 0, & otherwise, \end{cases} \tag{6}$$

where $D_2(n_i, n_j)$ indicates that nodes n_i and n_j have certain above inter-modal semantic semantic relation, $i \in [1, N_t]$ and $j \in [N_t + 2 \times N_e + 1, N_c]$.

Ultimately, by combing the adjacency matrices for intra-modal and inter-modal semantic relations, i.e., $\mathbf{A}^1$ and $\mathbf{A}^2$, we can derive the final adjacency matrix $\mathbf{A}$ for the knowledge-enhanced cross-modal graph.

4.3.3. Graph Convolution Network

Towards the final cross-modal context comprehension, we adopt L layers of GCN to extract the multimodal fusion feature of the cross-modal context. Then the node representations are iteratively updated as follows,

$$\mathbf{G}_l = ReLU(\tilde{\mathbf{A}}\mathbf{G}_{l-1}\mathbf{W}_l), l \in [1, L], \tag{7}$$

where $\tilde{\mathbf{A}} = (\mathbf{D})^{-\frac{1}{2}}\mathbf{A}(\mathbf{D})^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, and $\mathbf{D}$ is the degree matrix of the adjacency matrix $\mathbf{A}$. In addition, $\mathbf{W}_l \in \mathbb{R}^{D\times D}$ is a trainable parameter of the l-th GCN layer. $\mathbf{G}_l$ are the node representations obtained by the l-th GCN layer, where $\mathbf{G}_0 = \mathbf{H}$ is the initial node representation.

4.4. Task-specific Instruction Tuning for Financial Prediction

The final nodes representation $\mathbf{G}_L$ obtained by the L-th layer GCNs absorb rich semantic information from their correlated nodes and can be used as the

Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15
HistPrice	0.390	0.470	0.400	0.420	0.430	0.430	0.410	0.420	0.190	0.260	0.210	0.230
P-SVM	0.400	0.480	0.340	0.530	0.433	0.490	0.338	0.500	0.190	0.270	0.190	0.370
P-LSTM	0.410	0.473	0.291	0.546	0.399	0.391	0.421	0.442	0.123	0.232	0.165	0.341
MLP	0.349	0.435	0.209	0.539	0.267	0.319	0.331	0.351	0.101	0.201	0.124	0.311
LSTM	0.449	0.435	0.269	0.527	0.414	0.596	0.371	0.432	0.137	0.229	0.199	0.369
MMIM	0.435	0.653	0.302	0.605	0.392	0.631	0.329	0.601	0.296	0.217	0.142	0.385
MDRM	0.449	0.419	0.462	0.355	0.409	0.392	0.494	0.324	0.177	0.161	0.379	0.152
HTML	0.490	0.645	0.458	0.541	0.431	0.504	0.557	0.482	0.484	0.531	0.298	0.626
MULT	0.491	0.630	0.536	0.629	0.443	0.625	0.572	0.612	0.499	0.547	0.473	0.521
MPCNet	0.501	0.590	0.565	0.638	0.460	0.590	0.559	0.620	0.520	0.570	0.329	0.450
MANAGER	**0.548***	**0.694***	**0.610***	**0.659***	**0.517***	**0.652***	**0.589***	**0.646***	**0.564***	**0.608***	**0.511***	**0.681***

(a) Stock Indices and Currency Exchange Rate

Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15
HistPrice	0.360	0.390	0.350	0.400	0.310	0.290	0.220	0.390	0.220	0.160	0.340	0.330
P-SVM	0.390	0.420	0.370	0.380	0.340	0.310	0.330	0.330	0.370	0.220	0.310	0.390
P-LSTM	0.365	0.352	0.371	0.346	0.320	0.291	0.342	0.258	0.377	0.234	0.332	0.314
MLP	0.243	0.215	0.288	0.315	0.244	0.299	0.234	0.174	0.332	0.157	0.248	0.394
LSTM	0.361	0.337	0.304	0.345	0.364	0.311	0.255	0.394	0.381	0.168	0.382	0.444
MMIM	0.209	0.508	0.412	0.318	0.411	0.318	0.345	0.138	0.417	0.306	0.417	0.379
MDRM	0.434	0.383	0.214	0.317	0.287	0.242	0.314	0.149	0.346	0.198	0.478	0.505
HTML	0.441	0.654	0.379	0.526	0.529	0.278	0.466	0.389	0.424	0.314	0.397	0.450
MULT	0.329	0.590	0.454	0.533	0.534	0.364	0.485	0.400	0.428	0.171	0.466	0.493
MPCNet	0.444	0.668	0.413	0.637	0.386	0.327	0.560	0.625	0.493	0.556	0.374	0.537
MANAGER	**0.486***	**0.696***	**0.507***	**0.672***	**0.612***	**0.391***	**0.587***	**0.649***	**0.521***	**0.583***	**0.519***	**0.574***

(b) Gold Prices, Long-term (10-Years) and Short-term (3-Months) Bonds

Table 1: Performance comparison with baselines for movement prediction in terms of F1 score τ-days after the call ($\tau \in \{1, 3, 7, 15\}$). The best results are in boldface, while the second best are underlined. $\star$ denotes that the p-value of the significant test between our result and the best baseline result is less than 0.01.

Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
HistPrice	2.486	2.234	1.880	1.664	3.397	3.316	2.934	2.972	2.709	3.187	3.127	3.291
P-SVM	2.489	2.220	1.915	1.753	2.568	2.921	1.971	2.012	2.104	2.534	1.921	2.231
P-LSTM	2.421	2.217	1.845	1.731	2.128	2.194	2.108	1.456	1.424	1.867	1.015	1.569
MLP	2.524	2.214	1.899	1.680	1.469	1.597	0.937	0.981	1.060	1.441	0.802	1.159
LSTM	2.290	2.210	1.750	1.680	1.346	1.304	0.724	0.779	1.219	1.296	0.762	0.558
MMIM	2.290	2.092	1.779	1.598	1.287	1.133	0.718	0.622	0.975	1.081	0.500	0.510
MDRM	2.065	2.511	1.748	1.597	1.281	1.578	0.683	0.612	1.183	1.627	0.769	0.512
HTML	2.296	2.133	1.771	1.611	1.302	1.127	0.766	0.609	0.988	1.118	0.588	0.498
MULT	2.073	2.179	1.768	1.605	1.288	1.133	0.672	0.742	1.022	1.018	0.549	0.497
MPCNet	2.233	2.089	1.732	1.594	1.269	1.046	0.806	0.607	1.176	1.001	0.469	0.470
MANAGER	**1.819***	**1.725***	**1.608***	**1.471***	**1.126***	**0.813***	**0.584***	**0.572***	**0.906***	**0.957***	**0.416***	**0.402***

(a) Stock Indices and Currency Exchange Rate

Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
HistPrice	3.193	3.039	2.675	2.683	4.132	4.020	3.472	3.334	3.899	3.665	3.063	2.913
P-SVM	2.568	2.543	1.967	2.104	3.212	3.589	2.986	3.141	3.235	3.143	2.922	2.874
P-LSTM	1.965	1.998	1.043	1.764	2.212	1.699	2.340	1.453	3.433	2.909	2.678	2.477
MLP	1.431	1.654	0.904	0.955	1.811	1.743	1.288	1.382	2.582	2.523	2.239	2.231
LSTM	1.472	1.484	0.703	0.508	1.735	1.801	1.169	1.235	2.421	2.439	2.044	2.013
MMIM	1.292	1.292	0.565	0.486	1.698	1.604	1.080	1.053	2.345	2.392	1.977	1.902
MDRM	1.436	1.843	0.710	0.483	1.729	1.699	1.126	1.223	2.406	2.622	2.096	1.993
HTML	1.277	1.291	0.589	0.524	1.685	1.612	1.103	1.149	2.342	2.356	1.962	1.998
MULT	1.314	1.335	0.579	0.503	2.122	1.837	1.104	1.037	1.174	2.515	1.973	1.903
MPCNet	1.342	1.275	0.562	0.477	1.767	1.602	0.979	1.142	2.431	2.319	1.948	1.901
MANAGER	**1.106***	**1.144***	**0.527***	**0.419***	**1.452***	**1.574***	**0.825***	**0.917***	**1.049***	**2.076***	**1.804***	**1.276***

(b) Gold Prices, Long-term (10-Years) and Short-term (3-Months) Bonds

Table 2: Performance comparison with baselines for volatility prediction in terms of MSE τ-days after the call ($\tau \in \{1, 3, 7, 15\}$). The best results are in boldface, while the second best are underlined. $\star$ denotes that the p-value of the significant test between our result and the best baseline result is less than 0.01.

input for the following financial prediction. Considering that we need to solve a couple of tasks (i.e., prediction of the price movement and volatility), we resort to the advanced large language model ChatGLM2, which shows great performance in context comprehension (Du et al., 2022), and fine-tune it for each task independently. In addition, constructing proper instructions is pivotal for task-specific tuning, with each task being guided by a unique instruction prompt. Therefore, we adopt the instruction template (Wang et al., 2023a) structured as follows:

`Instruction`: $[prompt]$ Input: $[input]$ Answer: $[output]$

This template provides a standardized format, facilitating consistency across different tasks. Notably, we utilize the aforementioned final nodes representation $\mathbf{G}_L$ as input. Next, we design prompt for specific task.

Movement Prediction. In this task, we aim to predict the price movement for the financial assets. Therefore, the movement-oriented prompt is designed to guide ChatGLM2 to judge the price movement (e.g., "increase" or "decrease") of the given asset based on the multimodal input. The prompt template is "Please predict the price movement of O in τ days after the $date$ according to the input", where O is the to-be-predicted financial asset, $\tau \in \{1, 3, 7, 15\}$ and $date$ is formatted as `YYYY-MM-DD`.

Volatility Prediction. In this task, we aim to predict the volatility, a float number that measures the instability of an asset. Therefore, the volatility-oriented prompt is designed to guide ChatGLM2 to output the volatility of the given financial asset based on the multimodal input. Similar to the above prompt template, we just replace "price movement" with "volatility".

We then utilize encoder of ChatGLM2 to embed the prompt and concatenate it with the input. $[output]$ is the prediction result that is answered after we feed the instruction into ChatGLM2. And we can obtain the task-specific instruction (i.e., Movement-oriented instruction I_p and volatility-oriented instruction I_v). Finally, we feed I_p and I_v into ChatGLM2 independently to guide it to conduct the two financial prediction tasks. For optimizing our MANAGER, we adopt Binary Cross-Entropy (BCE) loss and Mean Squared Error (MSE) loss to train the output for price movement prediction and volatility prediction, respectively.

5. Experiment

5.1. Dataset

In this work, we conducted extensive experiments on Monopoly (Mathur et al., 2022) dataset for financial prediction. It is a collection of public monetary conference call videos along with their corresponding audio recordings and text transcripts

Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15
w/o-Text	0.479	0.590	0.513	0.544	0.437	0.556	0.436	0.549	0.483	0.510	0.433	0.609
w/o-Knowledge	0.530	0.677	0.592	0.631	0.485	0.630	0.574	0.628	0.547	0.591	0.497	0.670
w/o-Video	0.509	0.664	0.581	0.627	0.477	0.606	0.562	0.611	0.530	0.573	0.484	0.657
w/o-Audio	0.527	0.671	0.586	0.612	0.493	0.639	0.570	0.620	0.546	0.588	0.494	0.660
w/o-Graph	0.533	0.679	0.491	0.608	0.492	0.611	0.560	0.627	0.532	0.570	0.483	0.654
w/-FullGraph	0.499	0.582	0.548	0.601	0.429	0.627	0.451	0.595	0.510	0.521	0.448	0.639
MANAGER	**0.548**	**0.694**	**0.610**	**0.659**	**0.517**	**0.652**	**0.589**	**0.646**	**0.564**	**0.608**	**0.511**	**0.681**

(a) Stock Indices and Currency Exchange Rate

Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15	F1_1	F1_3	F1_7	F1_15
w/o-Text	0.369	0.571	0.448	0.525	0.498	0.328	0.467	0.583	0.460	0.516	0.428	0.502
w/o-Knowledge	0.470	0.659	0.488	0.656	0.593	0.370	0.562	0.627	0.504	0.576	0.498	0.557
w/o-Video	0.451	0.639	0.471	0.646	0.589	0.352	0.550	0.617	0.501	0.537	0.500	0.545
w/o-Audio	0.456	0.662	0.479	0.631	0.573	0.362	0.554	0.619	0.496	0.553	0.481	0.521
w/o-Graph	0.457	0.681	0.457	0.650	0.593	0.359	0.559	0.624	0.503	0.551	0.487	0.540
w/-FullGraph	0.397	0.592	0.461	0.620	0.519	0.317	0.512	0.607	0.495	0.532	0.413	0.518
MANAGER	**0.486**	**0.696**	**0.507**	**0.672**	**0.612**	**0.391**	**0.587**	**0.649**	**0.521**	**0.583**	**0.519**	**0.574**

(b) Gold Prices, Long-term (10-Years) and Short-term (3-Months) Bonds

Table 3: Ablation study results of our proposed MANAGER for movement prediction. The best results are highlighted in boldface.

Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
w/o-Text	2.146	2.048	1.795	1.613	1.379	1.016	0.879	0.926	1.198	1.348	0.702	0.617
w/o-Knowledge	1.930	1.908	1.784	1.536	1.291	0.905	0.670	0.713	0.977	1.019	0.474	0.566
w/o-Video	2.101	1.997	1.891	1.741	1.324	1.089	0.851	0.854	1.122	1.141	0.668	0.609
w/o-Audio	1.959	1.893	1.748	1.609	1.237	1.003	0.776	0.790	1.035	1.124	0.603	0.593
w/o-Graph	1.941	1.929	1.754	1.712	1.331	1.004	0.750	0.755	1.136	1.180	0.603	0.449
w/-FullGraph	2.144	1.962	1.919	1.791	1.495	1.022	0.898	0.904	1.175	1.306	0.681	0.640
MANAGER	**1.819**	**1.725**	**1.608**	**1.471**	**1.126**	**0.813**	**0.584**	**0.572**	**0.906**	**0.957**	**0.416**	**0.402**

(a) Stock Indices and Currency Exchange Rate

Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
w/o-Text	1.451	1.400	0.796	0.794	1.812	1.910	1.110	1.173	1.400	2.403	2.090	1.498
w/o-Knowledge	1.222	1.215	0.593	0.427	1.594	1.708	0.886	1.092	1.237	2.211	1.900	1.467
w/o-Video	2.184	1.351	0.576	0.704	1.641	1.850	1.070	1.101	1.344	2.280	1.979	1.510
w/o-Audio	1.307	1.358	0.639	0.549	1.588	1.758	1.024	1.127	1.264	2.213	1.943	1.419
w/o-Graph	1.347	1.278	0.742	0.540	0.674	1.861	1.013	1.164	1.275	2.226	1.937	1.494
w/-FullGraph	1.365	1.367	0.795	0.680	1.816	1.834	1.073	1.168	1.234	2.373	2.075	1.463
MANAGER	**1.106**	**1.144**	**0.527**	**0.419**	**1.452**	**1.574**	**0.825**	**0.917**	**1.049**	**2.076**	**1.804**	**1.276**

(b) Gold Prices, Long-term (10-Years) and Short-term (3-Months) Bonds

Table 4: Ablation study results of our proposed MANAGER for volatility prediction. The best results are highlighted in boldface.

released by six international banks between 2009 and 2022. Overall, it consists of $24,180$ samples, and each sample includes the corresponding text, video and audio clips with the annotated price movement and volatility. We adopted the original dataset split setting, the ratio of data split for training/validation/testing sets is $7:1:2$.

5.2. Experimental Setup

We adopted ChatGLM2[3] as the backbone of our model. The total number of tokens in each sample, i.e., N_t is unified to 768. The feature dimension D is set to 768. We used AdamW (Loshchilov and Hutter, 2017) as the optimizer and set the learning rate of GCN layers to 1e-3. Following Mathur et al. (2022), we use a learning rate of 1e-4 for movement prediction and 1e-3 for volatility prediction, respectively. The batch size is set to 1 due to GPU limitation, and the maximum number of epochs for model training is set to 10. Following the previous work, we employed mean squared error (MSE) to evaluate the predicted volatility and used F1 score to measure the predicted price movement, respectively, for $\tau \in \{1, 3, 7, 15\}$.

5.3. Baseline methods

5.3.1. Text-only baselines

- **HistPrice** (Du and Budescu, 2007). It utilizes the ARIMA model to perform regression/classification.
- **P-SVM** (Chatzis et al., 2018). This model applies Support Vector Regression (SVR) and Classifiers (SVC) for volatility and price movement prediction, respectively.
- **P-LSTM** (Yu and Li, 2018). It uses LSTM to extract forecast patterns from 30-day historical price time-series.

5.3.2. Multimodal baselines

- **MLP** (Tolstikhin et al., 2021). It is a simple multi-layer perceptron where multimodal features are aggregated across a time series and concatenated for prediction.
- **LSTM** (Poria et al., 2017). It feeds the multimodal time series to individual LSTMs and averages them before the final prediction.
- **MMIM** (Han et al., 2021). In this model, LSTMs are employed to encode the video and audio sequences, while BERT is utilized for text processing. Subsequently, the encoded features are fused for prediction.
- **MDRM** (Qin and Yang, 2019). It adopts BiLSTM layers to encode unimodal sequences, which are then fused together for prediction.
- **HTML** (Yang et al., 2020). HTML utilizes fused multimodal feature representations before passing through Transformer layers for final prediction.
- **MULT** (Tsai et al., 2019). It employs transformer encoders to align text, video, and audio sequences for prediction.

[3]`https://huggingface.co/THUDM/chatglm2-6b`.

- **MPCNet** (Mathur et al., 2022). It adopts cross-modal transformer blocks and modality-specific attention fusion for prediction.

5.4. Experimental results

We reported the experiment results in Table 1 and Table 2. From the above tables, we have the following observations. 1) Our model **MANAGER** consistently exceeds all the baselines in terms of all the metrics for both price movement and volatility prediction, which thoroughly demonstrates the superiority of our model. 2) Overall, the second best model is always multimodal baseline which verifies that the video and audio modalities can provide useful information for the financial prediction. 3) Notably, multimodal models not always outperform text-only models. For example, **HistPrice**, **P-SVM** and **P-LSTM** exceed **MLP** in the movement prediction of Stock Index (Large). It imples that improper use of non-verbal information in video and audio may lead to worse performance.

6. Analyses

6.1. Ablation Study

We introduced the following variants to explore the contribution of each component.

- **w/o-Text**, **w/o-Knowledge**, **w/o-Video w/o-Audio** and **w/o-Graph**. To prove the effectiveness of the input text, inferred knowledge, video, audio and constructed knowledge-enhanced cross-modal graph, we eliminated the text, external financial knowledge, video, audio and graph in these variants, respectively.
- **w/-FullGraph**. To further investigate the semantic relations of our knowledge-enhanced cross-modal graph, we erased all the semantic relations and transformed the semantic graph to a full connection graph.

The ablation study results are shown in Table 3 and Table 4. From this table, we have the following observations. 1) w/o-Text performs terribly compared with MANAGER. This is reasonable since the caption is the main source for delivering information to predict the price movement or volatility. 2) MANAGER exceeds w/o-Knowledge. It proves that external knowledge in the financial domain can assist in comprehending the context. 3) MANAGER consistently outperforms w/o-Video and w/o-audio across different evaluation metrics. It demonstrates the non-verbal information residing in the video and audio can improve context comprehension and hence boost financial prediction. 4)w/o-Text performs worse than both w/o-Video and w/o-Audio. It implies that the given text contributes more to the financial prediction than video and audio. 5) MANAGER outperforms w/o-Graph, denoting that the graphs are essential to capture the semantic relations among text, knowledge, video and audio, which help comprehend the cross-modal context. And 6) w/-FullGraph performs worse than MANAGER, which verifies the utility of proposed semantic relations.

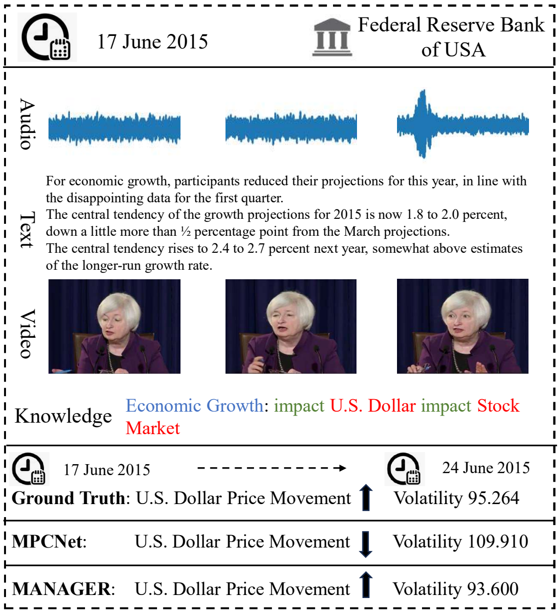


Figure 3: Comparison between the results predicted by MANAAGER and the best baseline MPCNet on one testing sample.

6.2. Case Study

To get an intuitive understanding of how our model works on financial prediction from MPC calls, we showed one testing sample in Figure 3 due to the limited space. For comparison, we also displayed the prediction results of the best baseline MPCNet.

As you can see, our MANAGER predicted the price movement of U.S. Dollar correctly, while MPCNet failed. In addition, the volatility 93.600 forecasted by MANAGER is closer to the ground truth 95.264 than 109.910 predicted by MPCNet. This may be attributed to the fact that the external knowledge (e.g., relation: “impact”, entity: “U.S. Dollar” and “Stock Market”) inferred by the entity “Economic Growth” may guide our model to pay attention to “Economic Growth” existed in the text, since it may provide some useful information for the price movement or volatility of U.S. Dollar. Overall, this case shows the benefit of incorporating external knowledge into the context of financial prediction from MPC calls.

7. Conclusion and Future Work

In this work, we propose a novel modal-adaptive knowledge-enhanced graph-based financial prediction scheme. Experimental results on a public dataset demonstrate the superiority of our model over existing cutting-edge methods, and validate the advantage of utilizing external knowledge in the financial domain, as well as the benefit of constructing the knowledge-enhanced cross-modal graph to characterize the intra-modal and inter-modal relations among the multiple input (i.e., text, external knowledge, video and audio). In the future, we plan to explore the Multimodal Large Language Models, such as VisualGLM, in financial prediction.

8. Acknowledgement

We thank all the anonymous reviewers for their valuable suggestions. This work is supported by a Research Grant from Mizuho Securities Co., Ltd. We sincerely thank Mizuho Securities for valuable domain expert suggestions. Ruihan Bao and Xu Sun are the corresponding authors.

9. Bibliographical References

Ellyn R. Boukus and Joshua V. Rosenberg. 2006. The information content of fomc minutes. *Monetary Economics*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*.

Yongbin Cai, Santiago Camara, and Nicholas Capel. 2021. It's not always about the money, sometimes it's about sending a message: Evidence of informational content in monetary policy announcements.

Longbing Cao. 2021. Ai in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*.

Sotirios P. Chatzis, Vasilis Siakoulis, Anastasios Petropoulos, Evangelos Stavroulakis, and Nikos E. Vlachogiannakis. 2018. Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Syst. Appl.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. ACL.

Ning Du and David V. Budescu. 2007. Does past volatility affect investors' price forecasts and confidence judgements? *International Journal of Forecasting*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 320–335.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *ArXiv*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*.

Weiwei Jiang. 2020. Applications of deep learning in stock market prediction: recent progress. *Expert Syst. Appl.*

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 11349–11361. ACL.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*. OpenReview.net.

Katharina Lewellen. 2003. Financing decisions when managers are risk averse. *MIT Sloan School of Management Working Paper Series*.

Xiaohui Victor Li. 2023. Findkg: Dynamic knowledge graph with large language models for global finance. pages 1–64. SSRN.

Vivian Liu and Lydia B. Chilton. 2021. Design guidelines for prompt engineering text-to-image generative models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Puneet Mathur, Atula Tejaswi Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. 2022. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. *Proceedings of the 30th ACM International Conference on Multimedia*.

Kun Ouyang, Liqiang Jing, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Sentiment-enhanced graph-based sarcasm explanation in dialogue. *CoRR*, abs/2402.03658.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*.

Soujanya Poria, E. Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Annual Meeting of the Association for Computational Linguistics*.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Annual Meeting of the Association for Computational Linguistics*.

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. 2021. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Annual Meeting of the Association for Computational Linguistics*.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. *Proceedings of the 28th ACM International Conference on Multimedia*.

Adam Hale Shapiro and Daniel J. Wilson. 2019. Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. *Federal Reserve Bank of San Francisco, Working Paper Series*.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. Mlp-mixer: An all-mlp architecture for vision. In *Neural Information Processing Systems*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. *Proceedings of the conference. Association for Computational Linguistics. Meeting*.

Neng Wang, Hongyang Yang, and Chris Wang. 2023a. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *ArXiv*.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023b. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *NeurIPS Workshop on Instruction Tuning and Instruction Following*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *ArXiv*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *ArXiv*.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. *Proceedings of The Web Conference 2020*.

J. Yoon, Beom Jun Woo, and Nam Soo Kim. 2022. Hubert-ee: Early exiting hubert for efficient speech recognition. *ArXiv*.

Shui-Ling Yu and Zhe Li. 2018. Forecasting stock price index volatility with lstm deep neural network.

Yixuan Zhang and Haonan Li. 2023. Can large langauge model comprehend ancient chinese? a preliminary test on aclue. In *International Conference on Algebraic and Logic Programming*.

NetZeroFacts: Two-Stage Emission Information Extraction from Company Reports

Marco Wrzalik[1], **Florian Faust**[2], **Simon Sieber**[2], **Adrian Ulges**[1]
[1]RheinMain University of Applied Sciences, Wiesbaden / Germany
[2]Sociovestix Labs, Kaiserslautern / Germany
{marco.wrzalik,adrian.ulges}@hs-rm.de, {florian.faust,simon.sieber}@sociovestix.com

Abstract

We address the challenge of efficiently extracting structured emission information, specifically emission goals, from company reports. Leveraging the potential of Large Language Models (LLMs), we propose a two-stage pipeline that first filters and retrieves potentially relevant passages and then extracts structured information from them using a generative model. We contribute an annotated dataset covering over 14,000 text passages, from which we extracted 739 expert annotated facts. On this dataset, we investigate the accuracy, efficiency and limitations of LLM-based emission information extraction, evaluate different retrieval techniques, and estimate efficiency gains for human analysts by using the proposed pipeline. Our research demonstrates the promise of LLM technology in addressing the intricate task of sustainable emission data extraction from company reports.

1. Introduction

To address the climate crisis - probably the most central and difficult challenge of our time - economies have to cope with massive industrial restructuring. The focus is on "Net Zero", i.e. achieving at least a neutral net balance of emitted greenhouse gases (especially CO_2) as quickly as possible. To reach this goal, policies such as the EU's *green taxonomy* are targeted at redirecting budget flows into sustainability-oriented businesses. To qualitatively capture the environmental impact of organizations, also referred to as "carbon accounting", analysts have to examine business reports and manually locate and extract the relevant data.

However, a complete and efficient screening remains an open issue: Companies often publish a variety of reports, such as quarterly, annual, sustainability and stewardship reports, which all may include aspects of emission reporting. To have a comprehensive understanding of a companies' net zero commitment, it is important to consider all of these reports. Analysts faced with the task of gathering net zero data need to browse large amounts of text for relevant information and extract it in a structured way. The expenses incurred are significant, in particular due to the need to carry out the screening process (1) for multiple companies, (2) in different industrial sectors, and (3) at regular intervals.

To increase the efficiency of this process, we address the question whether AI technology can support analysts with extracting structured greenhouse gas emission information from business reports. Specifically, the focus of this work is on extracting *emission goals*, i.e. the intention by companies and investment portfolios to reduce carbon emissions relatively or absolutely in certain timeframes and across certain sectors/scopes. For example, given the statement "*From 2019-2025, we will reduce the carbon footprints of our investments by 29 per cent*", we extract a fact with formal attributes (reduction-percentage=*29%*, target year=*2025*, reference year=*2019*, sector=*None*). Note that we are interested in extracting information in a *structured*, pre-defined format to store into a knowledge base. Such information extraction tasks have been studied intensely (Yang et al., 2022). However, what makes our particular task challenging is that substantial, binding goals must be discriminated from 'soft,' vague statements and self-praise, and that external requirements induced upon a company must be distinguished from self-imposed goals. For example, the statement "In order to achieve carbon neutrality by 2050, the Energy-Climate law provides for the reduction of fossil fuels consumption by 40% by 2030 " does not contain a self-imposed goal.

To address the above challenges, Large Language Models (LLMs) have recently appeared as a promising technology. These models show the remarkable ability to generate human-like text and perform a wide range of natural language understanding and generation tasks, serving as domain-agnostic problem solvers. The focus of our work is to investigate LLM technology for the auto-extraction of emission goals. We suggest a two-stage pipeline following the well-known retrieval-augmented generation (RAG paradigm) (Lewis et al., 2020), in which first potentially relevant passages are identified (*filtering/retrieval*) and then structured information is extracted by reasoning over these passages (referred to as *reading/extraction*). Our contributions are the following:

1. Although prior work on text classification for climate statements and emission targets exists

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 70–84
20 May, 2024.

(see Section 2), there is – to our knowledge – no public datasets covering end-to-end climate goal extraction. Therefore, we have conducted our own annotation effort, resulting in a dataset we coin *NetZeroFacts*. The dataset is based on $>$ 14,000 passages from climate-related business reports. We make it available for research purposes upon request.

2. We evaluate our pipeline through a case study on the *NetZeroFacts* dataset, and assess the overall accuracy of LLMs in three experiments: (1) We study the extraction stage, including – besides quantitative results – an analysis of error cases, (2) regarding the retrieval stage, we benchmark LLM-based retrieval against various keyword baselines and challenge the necessity of LLMs (given their high computational cost), and (3) we conduct an end-to-end evaluation, in which we identify accuracy bottlenecks and assess overall the extent to which LLMs can increase analysts' efficiency in practice.

2. Related Work

Emission Screening: The basis for emission ratings are annual reports and sustainability reports, codes of conduct, or controversial publications by the press and NGOs. These sources contain facts both in plain text and tabular form, from which analysists extract KPIs or other assessable statements (Is there a code of conduct? Has the company set emission goals? Are emissions even reported? etc.). Manual fact extraction comes with considerable manual effort, and tool support in practice has so far been limited to a coarse-grained document classification of report types (in order to filter out irrelevant documents) and keyword search, which could be problematic since reports from different sources differ in form and vocabulary (e.g. "CO2 emission" vs. "carbon release"). More advanced commercial tools such as *Intelligent Tagging*[1] can identify entities and indicators, but do not use Large Language Model (LLM) support yet.

Large Language Models (LLMs): The latest generation of large ($>$1 billion parameters), instruction-tuned LLMs – such as OpenAI's GPTs (OpenAI, 2023) or open-source alternatives such as Llama (Touvron et al., 2023) – learn to generate text on large-scale datasets. Since the quality of results has been shown to improve drastically with model and training data scale (Wei et al., 2022), a variety of large-scale models has been trained on increasing datasets recently – see Zhao et al. (Zhao et al., 2023) for a recent in-depth overview of the model landscape. Many models have been fine-tuned to follow instructions by a human conversation partner using reinforcement learning techniques (Ouyang et al., 2022), and can thus serve as general-purpose task solvers.

Since LLMs have led to significant progress across virtually any text understanding task, they can be useful for *both* stages of our pipeline, namely *retrieving* potentially relevant passages and *extracting* facts from them. We outline research in both areas in the following.

Passage Retrieval: To identify passages containing relevant facts, the predominant industry solution remains keyword search employing word occurrence statistics such as BM25 (Robertson and Zaragoza, 2009), which has proven an effective, cheap strategy for many use cases. However, more recently, LM-based models have been shown to yield improved results. These can be trained on labeled data (e.g., (Karpukhin et al., 2020)) or in a self-supervised fashion, with adjacent text segments treated as positive sample pairs (Neelakantan et al., 2022)), and encode both queries and passages into vector representations called *embeddings*. By comparing queries' and passages' embeddings via nearest neighbor search, this dense representation-based retrieval becomes a powerful alternative to traditional retrieval methods. In our study, we will compare and discuss both fundamental approaches – keyword search and embeddings search – for identifying emission goals.

Information Extraction: For more than five years, LMs have been the go-to approach for the extraction of facts from sentences and short documents. Early LM-based approaches add a so-called *head* component on top of a pre-trained LM, and fine-tune the resulting model to the targeted extraction task given a limited number of annotated training sentences. This way, models can be tailored to specific extraction tasks (e.g., (Gao et al., 2019; Eberts and Ulges, 2019)). With the aforementioned development of instruction-tuned LLMs as general-purpose problem solvers, it seems that a quality comparable to specialized extraction heads can be reached by prompting a system and requesting it to yield a structured output (Jiao et al., 2023; Zhang et al.; Gao et al., 2023). Since this comes without the need for fine-tuning, prompting instruction-tuned models appears to be the predominant approach today, and we follow this line of work.

Climate Fact Extraction: While information extraction has been applied to various domains (such as medical texts (Rasmy et al., 2020)) and target structures (such as arguments graphs (Lawrence and Reed, 2019)), the extraction of *climate-related* information has been studied rather scarcely. Stammbach et al. (Stammbach et al., 2023) formu-

[1] https://www.lseg.com/en/data-analytics/products/intelligent-tagging-text-analytics

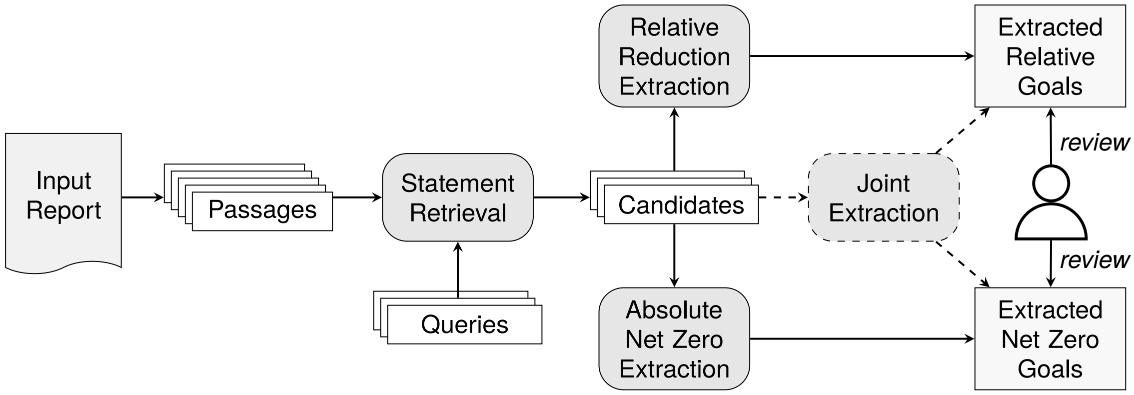


Figure 1: Approach Overview. A given report is broken down into passages, which are then filtered by statement retrieval. The resulting candidates are processed by extractors, which identify relative and net zero targets either in a joint processing step or separately.

late the detection of broader environmental statements as a binary classification problem (classifying a high vs low priotitization of environmental issues). ClimateBERT (Webersinke et al., 2021) follows the seminal BERT model (Devlin et al., 2019), combining a self-supervised masked LM pretraining on domain-specific text with a supervised fine-tuning of dedicated head models. On three climate-related text classification tasks, improvements over domain-agnostic pretraining are demonstrated. ClimateBERT-netzero (Schimanski et al., 2023) contributes a classification model and dataset for emission goal extraction. All these works primarily address text classification tasks with fine-tuned LMs, while we target a complete extraction pipeline (including retrieval *and* the extraction of structured information) and employ large-scale instruction-tuned LMs. The only other work we are aware of investigating these models specifically for climate-related text is ChatClimate (Ashraf Vaghefi et al., 2023), which – similar to our approach – investigates LLMs coupled with a climate-related text corpus. This work, however, adresses the answering of broad, climate-related questions, and not the bulk extraction of structured facts.

3. Approach

Our proposed method for extracting climate goals from given reports is targeted at two types of goals:

- A *net zero goal* expresses that a company wants to reach (at least) carbon neutrality. It comes with a target year, and optionally a subdivision of the company or the company operations.
- A *relative goal* expresses that a company wants to reduce its emissions by a certain percentage. It comes with a target year, reduction rate, reference year, and optionally a subdivision of the company or the company operations.

Goals are expressed in passages of text inside a report, consisting of at least one sentence up to a paragraph. Each report can contain multiple relevant passages, and each passage can state multiple different goals. For example, the text *"We commit to a target of carbon neutrality in own operations and own scope 1 and 2 GHG emissions reduced by at least 80% by 2030 compared with baseline year 2019."* contains

1. a *net zero goal* (`target year=2030, subdivision=own operations`)
2. and a *relative goal* (`target=80%, target year=2030, reference year=2019, subdivision=scope 1 and 2`)

Our approach towards extracting these goals is divided into two stages: First, a retrieval stage acts as a filter, limiting the amount of text to be processed and reducing false positives. Second, given the retrieved passages, we extract goals of both types. The result of the extraction is a list of structured facts, each with the aforementioned set of information fields. Figure 1 gives an overview of the approach.

3.1. Pre-processing

Our approach operates on the basis of plain text passages. Starting with PDF reports, we first extract the textual contents of each page using *Apache Tika*[2]. Next, we split the textual content of each page into sentences using the Python library *SoMaJo* (Proisl and Uhrig, 2016). The resulting sequence of sentences is used to generate overlapping passages: Each passage consists of three sentences, with subsequent passages sharing one

[2] https://tika.apache.org/

sentence. In other words, we use a sliding window of three sentences and shift this window by two positions to take the next passage. The resulting overlap reduces the risk that a passage is split in such a way that some information is missing from the target goal.

3.2. Statement Retrieval

Our statement retrieval (see Figure 2) serves as a filter for passages that contain climate goals. We use an information retrieval approach, i.e. queries are defined to express the information need for emission statements, and passages are ranked according to the relevance to these queries. Specifically, we explore two query types:

1. **Search by Question**: These are hand-crafted natural language queries that specifically ask for details to climate goals such as "*By what year do they expect to be carbon neutral?*". For this query modality, we have created a set of 14 questions.

2. **Search by Example**: These are example sentences or short passages that express one or more climate goals such as "*We are committed to carbon neutrality by 2050 with our investments*". We have collected 131 examples from held-out reports for this query modality.

Note that both methods use a pool of *multiple* queries $q_1, ..., q_n$ and that these pools can be refined iteratively with feedback.

Given a query q_i and a corpus of passages $d_1, ..., d_m$, a retriever model computes scores $s_{i,j} = score(q_i, d_j)$ which estimate the relevance of the passage. We explore two retriever models:

1. **Keyword Search** relies on *Elasticsearch*[3], a renowned industry standard search engine built on *Apache Lucene*[4]. Precisely, BM25 (Robertson and Zaragoza, 2009) is employed, a common relevance scoring technique based on keyword matching that adjusts each match based on the uniqueness of the word.

2. **Embedding Search** uses nearest neighbor search on LLM embeddings. We specifically adopt the OpenAI embedding model `text-embedding-ada-002`, which, according to the BEIR retrieval benchmark (Thakur et al., 2021), is the highest performing model currently available from OpenAI. Note that embedding search is more costly compared to keyword search, since it requires an LLM forward pass for each passage in the corpus.

[3] https://www.elastic.co/
[4] https://lucene.apache.org/

Both retriever models – given a query q_i – yield a ranked list of top results with scores s_{ij}. Given a passage d_j, these scores are fused across the queries using *score fusion* to obtain a single relevance score s_j^* indicating whether the passage contains a relevant fact (as illustrated in Figure 2). Given the passage's scores resulting from n different queries as $s_{1j}, s_{2j}, \ldots, s_{nj}$, we explore three score fusion techniques:

- **Max-Pooling**: Adopts the maximum score for a passage across all queries:

$$s_j^* = \max(s_{1j}, s_{2j}, \ldots, s_{nj}) \tag{1}$$

- **Sum**: The fused score for a passage is the sum of the scores across all queries:

$$s_j^* = \sum_{i=1}^{n} s_{ij} \tag{2}$$

- **Sum with Min-Max Normalization**: Each score is min-max normalized within its query's ranking: Let $s_i^{\min} := min_j \, s_{ij}$ and $s_i^{\max} := max_j \, s_{ij}$ be the minimum and maximum scores calculated for the i-th query, respectively. The fused score is calculated as:

$$s_j^* = \sum_{i=1}^{n} \frac{s_{ij} - s_i^{\min}}{s_i^{\max} - s_i^{\min}} \tag{3}$$

If a passage d_j is not retrieved by a query q_i, we set $s_{ij} = 0$.

3.3. Information Extraction

We feed all passages (ranked by the retriever) up to a certain cut-off rank to the extraction model. As described in the beginning of Section 3, we are interested in extracting two types of emission goals from passages, namely *net zero* goals vs *relative* goals. Both goals come with several attributes, such as a target year and (in case of relative goals) a reduction rate.

We tackle the extraction of emission targets in a two-stage process that relies heavily on few-shot prompting, using an instruction-tuned LLM (Ouyang et al., 2022). Specifically, we use the OpenAI model `gpt-3.5-turbo`. In this context, "few-shot" refers to the inclusion of a limited set of examples with correct answers, which serve as a pseudo-history accessible to the LLM.

1. In the first *filtering* stage, the LLM is asked *whether* the input passage describes at least one goal. The prompt instructs the model to respond with either "true" or "false", which is demonstrated in few-shot examples.

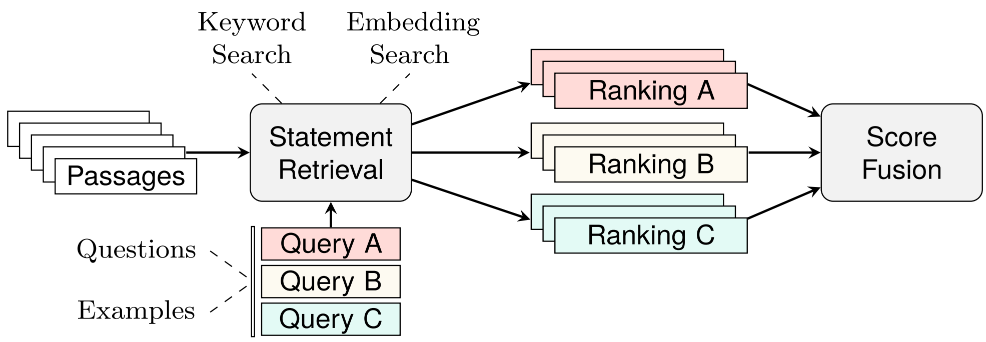


Figure 2: Retrieval is performed separately for each query or example. The resulting rankings are fused afterwards using conventional score fusing techniques.

2. If the LLM's response is positive (i.e., it starts with "true" or "yes"), the system proceeds to the actual extraction phase, which utilizes another prompt with three manually defined few-shot examples. Both the prompt and the examples instruct the LLM to produce output in the form of JSON objects containing only the fields relevant to the particular extractor. After successful parsing, these JSON objects represent our final extraction results.

In prior experiments, we found this two-stage process to outperform a single-stage extraction. This is also in line with OpenAI's public prompt engineering guidelines (Ope, 2024), which recommend to split tasks into simpler subtasks.

Note that the above two-stage process can either be executed for both types of goals at once, or separately. We explore both options:

- **Separate** prompting runs two separate two-stage processes, one for netzero goals and one for relative goals, resulting in four prompts per text passage.
- **Joint** prompting runs a single two-stage process: the filtering prompt responds positively if *either* type of goal is declared, and correspondingly the *extraction* prompt covers both goal types, resulting in two prompts per passage.

Each of the above six prompts was optimized independently from the others in a manual process of about 10 iterations, each including a small-scale inspection of a few responses but no quantitative benchmarking. Public prompt engineering guidelines were consulted in the process.

4. The NetZeroFacts Dataset

In this section, we introduce the *NetZeroFacts* dataset, which is based on real-world business reports known to contain emission statements. These were chunked into passages following our preprocessing as described in Section 3.1, and annotated by domain experts according to the criteria layed out in Appendix A. To evaluate not only end-to-end performance of our pipeline but also the individual steps, namely retrieval and extraction, the dataset consists of three partitions. We share our dataset, including all its partitions, upon request for research purposes.

NetZeroFacts-Small is based on 222 reports by different asset owner companies reporting sustainability and financial aspects (sustainability, annual, and integrated reports). The dataset's passages have been annotated by climate rating analysts during their daily sustainability rating activities, resulting in 270 passages annotated with a total of 317 climate goals. The purpose of the dataset is to evaluate the extraction step in-depth on a small-scale set of relevant passages.

NetZeroFacts-Big serves to evaluate the extraction of facts on a dense corpus of (widely irrelevant) text. It contains $13,950$ passages covering the complete content of 16 reports disjoint from the reports used for NetZeroFacts-Small.

To annotate the dataset, we applied extraction (using *separate* prompts, see Section 3.3) *densely* to all passages, resulting in 1250 climate goal facts belonging to 619 passages. The extracted facts were manually validated by an expert, resulting in a set of 422 positively validated facts in 289 passages. This dataset includes all passages, the automatically extracted facts, and the expert validation for each fact.

NetZeroFacts-Retrieval To evaluate the retrieval step, what matters is *whether* a passage contains at least one climate goal. Thus, we extend *NetZeroFacts-Big* to contain relevance labels. Relevant passages include those labeled positively by the expert *NetZeroFacts-Big*. However, since these include only passages for which LLM extraction was successful, and since our extractor may miss some climate goal facts in other passages, we also annotate additional passages for relevance using a top-15 pooled annotation of our best-performing keyword and embedding retrievals, focusing on those passages for which no facts have

Table 1: Extraction Recall and Precision on *NetZeroFacts-Small*. LLM-based extraction discovers 74.8% of known facts (left) and also yields new, unknown facts, at a precision of 71.3% (right). P is the number of positives, TP true positives, FP false positives.

Goal Type	**Annotated**	**Extracted**	**Recall**	**P**	**TP**	**FP**	**Precision**
Net Zero	295	221	0.749	358	269	89	0.751
Relative	22	16	0.727	136	83	53	0.610
Total	317	237	0.748	494	352	142	0.713

been extracted. In the resulting pool, 21 more passages were annotated as relevant. Note that such pooling is common practice in corpora for which dense annotation of relevance ratings is infeasible.

5. Experiments

In our experiment on the NetZeroFacts dataset, we investigate the individual system components and the overall end-to-end pipeline, and assess the level to which LLMs can improve the process of extracting CO_2 reduction targets from heterogeneous corporate documents:

- **Extraction Evaluation**: We first focus on the extraction step, and conduct two experiments: (a) a detailed evaluation on a small-scale set of relevant passages (*NetZeroFacts-Small*), and (b) a precision-oriented evaluation in which extraction is applied densely over all reports in *NetZeroFacts-Big*.
- **Retrieval Evaluation**: Retrieval as a pre-filtering is a key step to avoid a dense extraction – which would come with substantial computational cost and response delay in application. Therefore, we explore the different retrieval models proposed in Section 3 and assess their quality based on recall measures.
- **End-to-end Evaluation**: Finally, we assess the performance of our end-to-end pipeline, which includes the best-performing retrieval setting and the two variants of the extraction component.

5.1. Extraction Evaluation

We evaluate the extraction component in two experiments: First, we apply extraction on passages known to contain emission targets (NetZeroFacts-Small), second on the large but sparse dataset (NetZeroFacts-Big). This subsection's experiments focus on *separate* prompts (we will present a comparison of both prompting variants in the end-to-end evaluation in Section 5.3).

Detailed Evaluation (NetZeroFacts-Small): Our first evaluation on *NetZeroFacts-Small* gives us an assessment of the recall and discovery capabilities of extraction, and allows us to inspect challenge cases and common errors in-depth.

We ran extraction on *NetZeroFacts-Small*'s 270 passages, after which the correctness of the extracted facts was revised manually by an expert. Thereby, an extraction only counts as correct if all its fields are extracted correctly. Extractions that did not satisfy this strict criterion are counted as false positives. On the dataset, 237 extracted facts had previously been extracted by analysts in daily operations. Out of these, 74.8% have been extracted by the LLM (Table 1, left). Also, our LLM-based extraction managed to yield new facts undiscovered in the daily operations, which were again revised by the analyst. Table 1 (right) shows that 352 correct facts were discovered in total (including 115 *new* facts), at a precision of 71.3%. This indicates our pipeline's potential to increase the coverage of extraction.

An in-depth inspection revealed that most extraction mistakes fall into the following categories (ordered by descending frequency):

- relative goals and net zero goal are misclassified (69$\times$)
- the fact is missing altogether (74$\times$)
- the fact is incorrect (46$\times$)
- one goal refers to a target year of another goal in the same passage (27$\times$).

Dense Evaluation (NetZeroFacts-Big): While the passages in the last experiment were prefiltered to contain known emission targets, in a real-world scenario, the extractor is also faced with many irrelevant passages. Therefore, we performed the extraction densely for *all* 13,950 passages passages in *NetZeroFacts-Big*, resulting in 1,198 extracted fact candidates expressed in 657 passages.

These were manually revised, and the precision of the facts is reported in Table 2. We observe a significant drop in precision ($<$ 40%) compared to the previous experiment, which indicates that the LLM extracts a substantial amount of false positives from non-relevant passages. This is another motivation for pre-filtering candidate passages with a retrieval step, which will be investigated in the next section.

Table 2: Dense extraction results on *NetZeroFacts-Big* indicate a lower precision, showing that extraction tends to produce false positives on irrelevant facts.

Goal Type	Extracted	Precision
Net Zero	453	0.411
Relative	745	0.317
Total	1198	0.352

5.2. Retrieval Evaluation

We evaluate retrieval on the *NetZeroFacts-Retrieval* dataset, which features the $13,950$ passages from *NetZeroFacts-Big* with 310 positive relevance ratings. Thereby, we test the four retrieval variants outlined in Section 3.2: Using either keyword or embedding-based scoring, and using either questions or examples as queries. For each report, we rank all its passages and employ the Recall@k metric, which indicates how many of the relevant passages the analyst would discover when inspecting the top-k passages. These metrics are averaged over all reports.

First, we discuss the effectiveness of the statement retrieval using hand-crafted questions. Results are presented in the upper part of Table 3. The embedding-based retrieval variants outperform their keyword-based counterparts significantly. Furthermore, the data indicate the impact of the score fusion method and the need for its careful selection, whith a min-max-normalized sum fusion working best.

The bottom part of Table 3 shows the retrieval results when using sample *passages* as queries, which significantly improves the performance of the keyword-based retrieval approach. Again, the combination of sum fusion and min-max normalization appears most effective, while the max-pooling method significantly lags behind. A possible reason for this observation is the tendency of keyword searches to assign higher scores to longer queries. Given the different lengths of the sample passages, the longer examples are predisposed to receive higher scores, potentially leading to their dominance in a max-pooling fusion. In contrast, the embedding-centric search has an intrinsic normalization within the $[-1, 1]$ interval, making max-pooling the superior choice. However, it is noteworthy that – with example-based queries – the embedding-based approach performs much worse compared to the keyword-based search. To summarize, the sample-based retrieval method exhibits commendable performance, achieving a 95.2% recall rate for positive passages within the top 100 ranks. To do so, a keyword-based approach suffices.

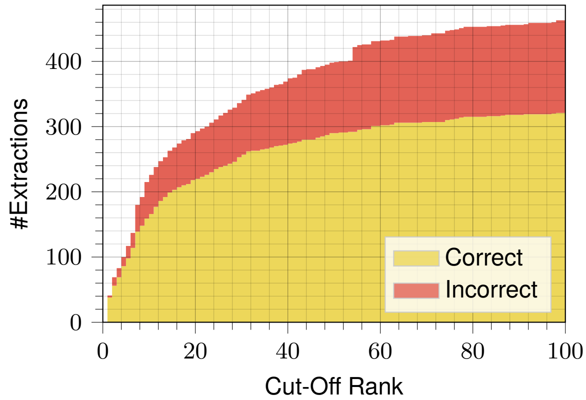


Figure 3: **End-to-end evaluation**: The number of extracted facts yield by our pipeline (*joint* prompting was used for extraction).

5.3. End-to-end Evaluation

Finally, we evaluate the entire pipeline of retrieval and extraction. We focus on the best-performing retrieval setup (keyword search with examples as queries) and evaluate extraction both with separate prompts or joint prompts (cmp. Section 3.3). For both extraction methods, an expert inspected the top-100 extracted facts (according to the associated passages' retrieval score).

Table 4 gives a comparison of both prompting methods. Joint prompting clearly outperforms separate prompts, which may be due to two reasons: First, we found the separate prompts to yield many false positives in which goal types were confused (e.g., together with a net zero goal, a relative goal with `target_rate=100%` would be extracted). Obviously, offering the LLM both goal types in the same prompt improves disambiguation between the types. Second, it should also be mentioned that separate and joint prompts were optimized independently (and ad-hoc), such that the joint prompt might per se be better suited. We share all prompts in Appendix B, and also make the prompts available with the *NetZeroFacts* dataset.

Figure 3 plots the number of facts extracted end-to-end, plotted against the cut-off rank (i.e., the number of passages per document fed to the extraction step). We observe that the correct facts flatten out at Rank 50, which yields 90% of recall compared to Rank 100. This indicates that manually reviewing only relatively few facts per report may suffice, and that the majority of facts to revise is correct.

Table 3: Evaluation measures for the retrieval stage. The best results are highlighted in boldface. Underscores indicate insignificant differences ($p \leq 0.05$) to the best result, according to a paired Student's t-test.

Search by	Method	Norm.	Fusion	Recall@10	Recall@20	Recall@50	Recall@100
Questions	Keyword	-	max	0.211	0.347	0.535	0.750
		-	sum	0.256	0.420	0.615	0.766
		min-max	sum	0.230	0.381	0.598	0.787
	Embedding	-	max	0.323	0.473	0.732	0.834
		-	sum	0.352	0.535	0.727	0.837
		min-max	sum	0.342	0.527	0.735	0.847
Examples	Keyword	-	max	0.280	0.414	0.618	0.817
		-	sum	0.385	0.595	0.853	0.949
		min-max	sum	0.379	**0.612**	**0.886**	**0.952**
	Embedding	-	max	0.354	0.510	0.728	0.851
		-	sum	0.392	0.526	0.737	0.858
		min-max	sum	**0.407**	0.543	0.778	0.869

Table 4: Performance metrics for joint and separate fact extraction.

Prompt	Goal Type	Extracted@100	P@5	P@15	P@50	P@100
Separate Prompt	Net Zero	402	0.445	0.452	0.478	0.440
	Relative	541	0.517	0.460	0.412	0.381
	Total	943	0.483	0.456	0.441	0.406
Joint Prompt	Net Zero	193	0.910	0.922	0.874	0.870
	Relative	258	0.774	0.653	0.631	0.585
	Total	451	0.838	0.757	0.727	0.693

6. Discussion

Workflow Efficiency In practice, analysts currently search with a list of keywords and manually inspect the detected keywords' textual contexts. Compared to this, our pipeline offers the following benefits: (1) our retrieval operates with a larger set of sample queries over which we pool, yielding a better prioritization to begin with, (2) analysts can skip passages for which extraction has yield no results (i.e., the extractor acts as an additional filter), and (3) while analysts must read and digest each detected passage so far, they only have to check pre-filled facts when using our approach. This leads to significant speed-ups in the extraction process: While processing one report with the current workflow takes 91 minutes on average (estimated on the 222 base documents from which *NetZeroFacts-Small* was constructed), we found the inspection of LLM results up to Rank 100 to take ≈ 15 minutes per report (which corresponds to a $6\times$ speedup).

Accuracy and Bottlenecks: According to our results, LLM processing is not sufficient for dark processing without expert cross-checking yet. However, we found many results to be *partially correct*, or semantically correct but formatted inadequately. True error cases for extraction are often tied with complicated passages in which multiple goals coincide (see Appendices C+D for examples). Accuracy could definitely be improved significantly with more thorough postprocessing, and future research could investigate domain-specific fine-tuning.

When it comes to accuracy bottlenecks, we consider the extraction step the limiting factor towards a fully automated extraction. Retrieval appears to be sufficiently accurate (with a recall@100 of $> 95\%$). Accordingly, we have refrained from fine-tuning task-specific LM-based retrieval models. Also, we found a well-tuned keyword-based approach employing example passages as queries to outperform LLM-based embeddings, which is good news from a cost perspective.

Future Research: One direction of future work could be to investigate NetZero classifiers based on fine-tuned LMs (Schimanski et al., 2023) for retrieval, another one to incorporate analysts' feedback when correcting LLM results. Note that both steps of our pipeline can take such feedback into account: In retrieval, passages can be used as example-based queries. In extraction, similar or particularly challenging passages can be included as few-shot examples (see, e.g., (Zhao et al., 2021)). Our *NetZeroFacts* dataset offers a good basis to explore these issues further.

7. Bibliographical References

2024. OpenAI docs prompt engineering.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara colesanti Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. Chatclimate: Grounding conversational ai in climate science. *Communications Earth & Environment*, 4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. *CoRR*, abs/1909.07755.

Jun Gao, Huan Zhao, Yice Zhang, Wei Wang, Changlong Yu, and Ruifeng Xu. 2023. Benchmarking large language models with augmented instructions for fine-grained information extraction.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6250–6255, Hong Kong, China. Association for Computational Linguistics.

Yizhu Jiao, Ming Zhong, Sha Li, Ruining Zhao, Siru Ouyang, Heng Ji, and Jiawei Han. 2023. Instruct and extract: Instruction tuning for on-demand information extraction.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6769–6781. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training. *CoRR*, abs/2201.10005.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Thomas Proisl and Peter Uhrig. 2016. SoMaJo: State-of-the-art tokenization for German web and social media texts. In *Proceedings of the 10th Web as Corpus Workshop (WAC-X) and the EmpiriST Shared Task*, pages 57–62, Berlin. Association for Computational Linguistics.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2020. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Medicine*, 4.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Tobias Schimanski, Julia Bingler, Mathias Kraus, Camilla Hyslop, and Markus Leippold. 2023. ClimateBERT-NetZero: Detecting and assessing net zero and reduction targets. In *Proceedings of the 2023 Conference on Empirical*

Methods in Natural Language Processing, pages 15745–15756, Singapore. Association for Computational Linguistics.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. Climatebert: A pretrained language model for climate-related text. *ArXiv*, abs/2110.12010.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. 2022. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19).

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. *ACL*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

A. Annotation Criteria

Annotations for *NetZeroFacts* were collected by domain experts deciding whether given text passages express a (net zero, or relative) goal according to the following criteria. Particularly, the expressed goals ...

- **must be measurable**: The mere description of climate-relevant activities without setting a reduction goal is labeled as irrelevant (such as *"We are working on infrastructure to make our facilities more energy efficient"*).
- **must be self-imposed**, meaning that constraints imposed through laws or other actors rather than the report's authors themselves do not count as goals (such as *"The Clean Shipping Act calls for the elimination of carbon emissions by 2024."* or *"70% of our customers have set the goal to reach carbon neutrality by 2024"*).
- **must directly address a CO_2 metric**: Statements addressing an indirect effect on emissions are not considered emission goals (such as *"We announce our commitment to phase out of coal by 2030"*).
- **must not report past emission statistics**: Obviously, the reporting of past emission statistics (such as *"We have reduced our absolute scope 1 and 2 emissions by 3% in 2022"*) does not count as a goal.
- **may not match the given schema**: Rarely, we discovered goals expressing an *absolute* reduction, but *not* to net zero (such as *"By 2025, we will reduce GHG emissions by 700 tons"*). While these cases were so rare that we did not measure them in the extraction + end-to-end benchmarks, we labeled them as relevant in the retrieval evaluation.

B. Prompts

For the sake of transparency, we are sharing our prompts with the community. All six prompts used in our extraction evaluation are listed below.

B.1. Net Zero Goal Filter

```
System: You are an information
    extraction tool for net zero climate
     goals. I will provide you with
    statements taken from asset manager
    reports. You will determine whether
    the given statements indicate an
    commitment to achieving net zero or
    carbon neutrality. Relative
    reductions in greenhouse gas
    emissions are excluded. Refraining
    from fossil fuels entirely does
    count as becoming carbon neutral.
    Your answers are 'True' if the
    statement contains an absolute net
    zero goal and 'False' in any other
    situation.
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investor s commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: True
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI:  False
Human: In our climate targets published
    in 2019, we are committed to exiting
     from investments in thermal coal by
     2025. We are also committed to
    excluding oil exploration from our
    investments by 2030.
AI: False
Human: <input>
```

B.2. Relative Goal Filter

```
System: You are an information
    extraction tool for net zero climate
     goals. I will provide you with
    statements taken from asset manager
    reports. You will need to determine
    whether the statements given express
     a percentage of reduction in
    greenhouse gas emissions or not.
    Absolute zero emission targets or
    goals to achieve net zero or become
    carbon neutral do not matter. Your
    answers will be 'True' if the
    statement contains a relative
    reduction target expressed as a
    percentage, and 'False' otherwise.
    Only answer 'True' if a concrete
    relative reduction rate is given.
    Achieving net-zero does not count a
    relative reduction.
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investors commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: False
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI: True
Human: In our climate targets published
    in 2019, we are committed to exiting
     from investments in thermal coal by
     2025. We are also committed to
    excluding oil exploration from our
    investments by 2030.
AI: False
Human: <input>
```

B.3. Net Zero Goal Extraction Prompt

```
System: You are an information
    extraction agent for climate goals.
    I will provide you with statements
    taken from asset manager reports
    that may contain one or more goals
    concerning net zero or neutral
    carbon emissions. For each goal you
    wil extract the target year until
    which the goal should be achieved.
    In addition, a goal can be dedicated
     to a certain sub-division, meaning
    that the reduction applies only to
    this area. Extract all of these
    goals in a list of json objects.
    Ignore all relative goals that state
     a specificreduction percentage.
    Only consider goals regarding
    absolute carbon neutrality or net
    zero emissions. If the target year
    or the sub-division is not given,
    fill the respective field with 'null
    '. Examples are provided below.
Human: We have also emphasised our green
     ambitions by announcing that, from
    2019-2025, we will reduce the carbon
     footprints of our investments by 29
     per cent.
AI: []
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
     (AOA) where a large number of the
    worlds biggest investors commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: [
  {
    "target_year": "2050",
    "division_subset": null
  }
]
Human: The company has published its
    commitment to decarbonisation,
    setting stringent objectives: to
    reduce the intensity of its CO2
    emissions to 50% below those of 2007
     by 2030, and to be carbon- neutral
    by 2050, forecasting virtually zero
    emissions in Europe by 2030. These
    goals have been recognised as being
    based on science in accordance with
    the Science Based Targets initiative
AI: [
  {
    "target_year": "2050",
    "division_subset": null
  },
  {
    "target_year": "2030",
    "division_subset": "Europe"
  }
]
Human: <input>
```

B.4. Relative Goal Extraction Prompt

```
System: You are an information
    extraction agent for climate goals.
    I will provide you with statements
    taken from asset manager reports
    that contain one or more goals
    concerning relative reductions in
    carbon emissions and you will
    extract these goals including
    specific facts: Each goal must state
     a concrete reduction percentage.
    Optionally, each goal has a target
    year until which the goal should be
    achieved, and a reference year to
    which the relative reduction is
    compared. In addition, a goal can be
     dedicated to a certain sub-division,
     meaning that the reduction applies
    only to this area. Extract all of
    these goals and their in a list of
    json objects. Ignore all absolute
    net zero or carbon neutrality goals,
     only considerrelative reductions
    with a reduction percentage. If a
    reference year or the sub-division
    is not given, fill the respective
    field with 'null'.Examples are
    provided below.
```

```
Human: We have also emphasised our green
    ambitions by announcing that, from
    2019-2025, we will reduce the carbon
    footprints of our investments by 29
    per cent.
AI: [
  {
    "relative_target_value": "29 per
        cent",
    "target_year": "2025",
    "reference_year": "2019",
    "division_subset": null
  }
]
Human: Therefore, together with the Net-
    Zero Asset Owner Alliance, we have
    spec- ified concrete CO2 reduction
    targets for the different parts of
    the real estate portfolio for the
    2019 to the end of 2024 period. For
    our Danish commercial prop- erties,
    the target is to reduce our CO2
    emissions by 39 per cent, and for
    our Danish residential properties,
    the target is 35 per cent. For the
    international part of the property
    portfolio, we are following the
    recommenda- tions of the EU-backed
    CRREM (Carbon Risk Real Estate
    Monitor) initiative, which is based
    on the Paris Agree- ments target of
    keeping global temperature increases
    below 1.5 degrees Celsius this
    century. In this context, the
    reduction target is 20 per cent for
    our German residential properties.
AI: [
  {
    "relative_target_value": "20 per
        cent",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "German
        residential properties"
  },
  {
    "relative_target_value": "39 per
        cent",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        commercial prop- erties"
  },
  {
    "relative_target_value": "35 per
        cent",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        residential properties"
  }
]
Human: <input>
```

B.5. Joint Goal Filter

```
System: You are an information
    extraction tool for climate goals
    that classifies whether a given text
    contains a statement about the
    commitment to a goal regarding
    carbon emissions. I will present to
    you passages from asset managers'
    reports. You will determine whether
    the given text contains a commitment
    to either a specific relative
    reduction in carbon emissions or to
    achieving net zero or carbon
    neutrality. Ignore any vague
    statements; a target is only a
    target if it states by when the
    target is to be achieved. For
    relative emission reductions, a
    specific percentage reduction must
    be stated. Ignore goals of third
    parties. Your answers are only 'True'
    if the statement contains such a
    climate target of the asset manager
    and 'False' if it does not.
Human: A standout feature of the
    GreenTech Solutions Factory is its
    pledge to function entirely on
    renewable energy sources, aiming for
    net-zero emissions across its
    operations.
AI: False
Human: We are participating in the UN-
    backed Net-Zero Asset Owner Alliance
    (AOA) where a large number of the
    worlds biggest investors commit
    themselves to being net carbon
    neutral in their investments by no
    later than 2050 and to continually
    make five-year sub-targets for CO2
    footprints.
AI: True
Human: Net zero emissions means
    achieving a balance between
    greenhouse gas (GHG) emissions
    produced and the amount removed from
    the atmosphere, consistent with
    limiting global warming to 1.5C and
    neutralising the impact of any
    residual emissions by permanently
    removing an equivalent amount of
    carbon dioxide (CO2). For BTPS this
    will mean reducing the portfolio's
    emissions through changing
    investments and investing in
    technologies which reduce emissions.
AI: False
Human: We have also emphasised our green
    ambitions by announcing that, from
    2019-2025, we will reduce the carbon
    footprints of our investments by 29
    per cent.
AI: True
Human: In our climate targets published
    in 2019, we are committed to exiting
    from investments in thermal coal by
    2025. We are also committed to
    excluding oil exploration from our
    investments by 2030.
AI: False
Human: <input>
```

B.6. Joint Goal Extraction

```
System: As an information extractor
    specialized in climate goals, your
    task is to analyze reports from
    asset managers. You are going to
    extract specific goals related to
    reducing carbon emissions or
    achieving net-zero carbon emissions,
    whether through their investments
    or policies within their own company.
    You'll respond with JSON objects
    detailing these goals.
 There are two types of targets:
   1. Relative Reduction (goal_type: '
      relativeReduction'): Specifies a
      percentage reduction with a target
      and base year.
   2. Net Zero (goal_type: 'netZero'):
      Specifies a commitment to carbon
      neutrality with a target year.
A 100 per cent relative reduction is
    also classified as netZero and not
    as relativeReduction. In addition, a
    goal can be dedicated to a certain
    sub-division, meaning that the goal
    applies only to this area, such as '
    energy consumption', 'fossil fuels'
    or emissions in a certain scope. If
    no target year or subdivision is
    specified, use 'null'. However, a
    relative reduction goal MUST specify
    a concrete reduction percentage;
    otherwise it is not a relative goal.
    Do not extract goals of third
    parties. Return an empty list if no
    targets are found. Ensure that the
    JSON objects are valid.
Human: The Albert Jackson Processing
    Centre will operate with net-zero
    emissions.
AI: []
Human: We aim to reduce CO2 emissions by
    39% for Danish commercial
    properties and 35% for residential
    by 2024, from a 2019 baseline. For
    German residential properties, the
    target is a 20% reduction by 2024,
    aligning with CRREM's guidelines.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "39%",
```

```
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        commercial properties"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "35%",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "Danish
        residential properties"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "20%",
    "target_year": "2024",
    "reference_year": "2019",
    "division_subset": "German
        residential properties"
  }
]
Human: Our commitment includes reducing
    CO2 emission intensity by 50% from
    2007 levels by 2030, and 100% by
    2050.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "50%",
    "target_year": "2030",
    "reference_year": "2007",
    "division_subset": null
  },
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": null
  }
]
Human: From 2019 to 2025, we will cut
    our investments' carbon footprint by
     29%.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "29%",
    "target_year": "2025",
    "reference_year": "2019",
    "division_subset": "investments"
  }
]
Human: Joining the Net-Zero Asset Owner
    Alliance, we commit to net carbon
    neutrality for investments by 2050.
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": "investments"
  }
]
Human: <input>
```

C. Exemplary Extraction Showcase

Based on the joint prompt extractor, we present some extractions, some of which reflect remarkable capabilities of the underlying LLM.

C.1. Successful Single-Target Extraction

```
Human: This means that with the same
    construction measures, by 2040 a
    CO² intensity of 2.41 kg CO²eq/m 2a
    will be reached, which corresponds
    to a reduction of 88% compared with
    the 2019 reference year. The
    targeted objectives are thus largely
     already being achieved. In order to
     plan the specific measures covering
     the three action areas -energy,
    heat and building shell renovations -
    holistically and co-ordinate them
    with each other, Swiss Prime Site
    integrated the requirements of the
    CO² reduction pathway into the
    property strategy for each
    individual property.
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "88%",
    "target_year": "2040",
    "reference_year": "2019",
    "division_subset": null
  }
]
```

C.2. Multiple Goals from Bullet Point List

```
Human: Sustainability Report 2022 |
    Environment 24 Our climate roadmap
    Our climate roadmap -what we are
    aiming for Founding member of the
    Net Zero Asset Managers initiative
    Addressing our own emissions
    Addressing the emissions of our
    lending activities (from 2020 levels)
     Addressing our supply chain Net
    zero across our activities (scopes 1,
     2 and 3) -in line with fi duciary
    duties •Achieve net-zero scope 1 and
     2 emissions •Reduce our own energy
    consumption by 15% from 2020 levels
    •Offset historical emissions from
    own operations back to 2000 •Net-
    zero GHG emissions by our key
    vendors Commitment to net zero by
    2050 and founding member of the Net-
    Zero Banking Alliance (the NZBA) and
     the Glasgow Financial Alliance for
    Net Zero (GFANZ) 2021 by 2025 by
    2030 by 2035 by 2050 •Reduce
    emissions intensity associated with
    UBS lending to: -residential real
    estate by 42% -commercial real
    estate by 44% -power generation by
    49% -cement by 15% •Reduce absolute
    fi nanced emissions associated with
    UBS lending to: -fossil fuels by 71%
     2020 2022Advisory vote on the
    climate roadmap passed by
    shareholders at the Annual General
    Meeting 24
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": null
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "42%",
    "target_year": "2030",
    "reference_year": "2020",
    "division_subset": "residential real
        estate"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "44%",
    "target_year": "2030",
    "reference_year": "2020",
    "division_subset": "commercial real
       estate"
  },
  <...>
]
```

C.3. Translating "halve" to 50%

```
Human: We aim to halve our carbon
    footprint this decade from 2018
    levels and will report our progress,
     transparently. To achieve that, we
    are working on transforming every
    aspect of our business -from
    switching our trucks to alternative
    fuels and our factories to renewable
     electricity, to improving circular
    packaging solutions that reduce
    plastic and can be more easily
    recycled or reused. These
    transformations are not just good
    for the climate: becoming a business
     that puts regeneration at the core
    of our company reduces our costs and
     makes us more resilient -and this
    is good for Nestlé, too.
AI: [
```

```
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "50%",
    "target_year": "2030",
    "reference_year": "2018",
    "division_subset": null
  }
]
```

C.4. Successful Interpretation of Table Data

```
Human: These metrics and targets are
    shown in the table on the right.
    STRATEGY PILLAR 2026 MILESTONE
    TARGETS 2030 ASPIRATIONS 2030 SBTi
    COMMITMENTS 2050 TARGET METRIC 2022
    BASELINE END 2026 TARGETS ENERGY %
    reduction in Scopes 1&2 CO2e
    emissions 212.3k TONNES 22%
    REDUCTION 70% of energy to come from
     renewables -46.2% in Scopes 1&2
    emissions vs 2019 baseline Net-Zero
    emissions in our value chain
    MATERIALS % volume free from new oil-
     extraction raw materials 26% 60%
    100% -33% in Scope 3 emissions vs
    2019 baseline WATER % of water to be
     recycled 23% 33% INCREASE IN
    RECYCLING RATE WASTE No waste to
    landfill 1.96k TONNES ZERO TO
    LANDFILL All effluent to meet ZDHC
    limits 92% 100% PEOPLE % employees
    in units with Great Place to Work
    certification 86% 88% 90% % of
    females in Senior Leadership
    positions 21% 30% 40% 12 Coats Group
     plc Sustainability Report 2022
    STRATEGY OVERVIEW CLIMATE REPORT
    WATER WASTE PEOPLE MATERIALS ABOUT
    ENERGY MANAGING SUSTAINABILITY
    PERFORMANCE SUMMARY
AI: [
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "22%",
    "target_year": "2026",
    "reference_year": "2022",
    "division_subset": "energy"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "46.2%",
    "target_year": "2030",
    "reference_year": "2019",
    "division_subset": "energy"
  },
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": "energy"
  },
  <...>
]
```

D. Exemplary Error Cases

We usually encounter similar types of errors during extraction. Some examples of typical errors are listed below.

D.1. Confusion with Non-Climate Goals

```
Human: Performance Report // Climate
    protection targets in harmony with
    the Paris Agreement and net zero
    emissions by 2050 // Proportion of
    women in top management 33% by 2025
    and 50% by 2030 // Sustainability
    targets as part of the variable
    compensation of the Board of
    Management
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2050",
    "division_subset": null
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "33%",
    "target_year": "2025",
    "reference_year": null,
    "division_subset": "proportion of
       women in top management"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "50%",
    "target_year": "2030",
    "reference_year": null,
    "division_subset": "proportion of
       women in top management"
  },
  {
    "goal_type": "null",
    "target_year": null,
    "division_subset": "sustainability
       targets as part of the variable
       compensation of the Board of
       Management"
  }
]
```

D.2. Additional Irrelevant Fact

```
Human: Reducing impacts from production
    Net zero Operations Downstream
    Upstream Promoting environmentally
    preferable products Improving
    transit management and supporting
    mixed-use buildings Improving energy
     efficiency of products
    Transitioning to renewable
    electricity Building resilient
    supply chain management and
    procurement Investing in digital
    solutions for real estate management
     Reducing impact from maintenance
    and service Climate action In 2021,
    we were the first global E&E company
     to declare a commitment to achieve
    net zero by 2040, with mid-term 2030
     targets covering scope 1, 2, and 3
    emissions against a 2020 baseline
    aligned to a 1.5-degree pathway.
    These were formally approved in 2022
     by the Science Based Targets
    initiative (SBTi), a multi-
    stakeholder climate initiative. We
    have also committed to shorter-term
    targets: fully transitioning to
    renewable electricity by 2025 when
    joining RE100 in 2021, the global
    environmental initiative led by the
    Climate Group in partnership with
    CDP; and imple- menting the ISO
    50001 management system for all
    major production sites by 2025.
AI: [
  {
    "goal_type": "netZero",
    "target_year": "2040",
    "division_subset": null
  },
  <...>
  {
    "goal_type": null,
    "target_year": "2025",
    "division_subset": "ISO 50001
       management system"
  }
]
```

D.3. Reporting Instead of Goals

```
Human: Canada Post Corporation 2022
    Sustainability Report 7 Introduction
     Governance Environment Social
    Transparency 2022 ESG highlights
    Environment Cl imate action •One of
    the first 150 companies in the world
     (one of only four Canadian
    companies) with a net-zero target
    approved by the SBTi •6% decrease in
     scope 1 and 2 emissions in 2022
```

```
    versus 2019 •12% decrease in scope 1
     and 2 emissions intensity (kt CO 2
    e/billion dollars revenue) versus
    2019 <...>
AI: [
  {
    "goal_type": "netZero",
    "target_year": null,
    "division_subset": null
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "6%",
    "target_year": null,
    "reference_year": "2019",
    "division_subset": "scope 1 and 2
        emissions"
  },
  {
    "goal_type": "relativeReduction",
    "relative_target_value": "12%",
    "target_year": null,
    "reference_year": "2019",
    "division_subset": "scope 1 and 2
        emissions intensity"
  },
]
```

FB-GAN: A Novel Neural Sentiment-Enhanced Model for Stock Price Prediction

Jainendra Kumar Jain, Ruchit Agrawal
Department of Computer Science, Department of Computer Science
University of Birmingham, University of Birmingham
Birmingham, United Kingdom, Dubai, United Arab Emirates
jkj293@alumni.bham.ac.uk, r.r.agrawal@bham.ac.uk

Abstract

Predicting stock prices remains a significant challenge in financial markets. This study explores existing stock price prediction systems, identifies their strengths and weaknesses, and proposes a novel method for stock price prediction that leverages a state-of-the-art neural network framework, combining the BERT language model for sentiment analysis on news articles and the GAN model for stock price prediction. We introduce the FB-GAN model, an ensemble model that leverages stock price history and market sentiment score for more accurate stock price prediction and propose effective strategies to capture the market sentiment. We conduct experiments on stock price prediction for five major equities (Amazon, Apple, Microsoft, Nvidia, and Adobe), and compare the performance obtained by our proposed model against the existing state-of-the-art baseline model. The results demonstrate that our proposed model outperforms existing models across the five major equities. We demonstrate that the strategic incorporation of market sentiment using both headlines as well summaries of news articles significantly enhances the accuracy and robustness of stock price prediction.

Keywords: Stock Price Prediction, Sentiment Analysis, GAN, NLP for Finance, BERT, Opinion Mining

1. Introduction

Accurate stock price prediction is a crucial challenge amidst rapid information transmission and complex market dynamics. Traditional quantitative models, while somewhat effective, often fail to grasp market sentiment nuances, relying heavily on historical data. The rise of social media, financial news sites, and online forums has revolutionised the accessibility to stock market-related information. Consequently, investor sentiment, characterized by emotions, opinions and beliefs has emerged as a dynamic force capable of swiftly altering market trends. Based on recent studies done by Xiaodong Liu and Li (2023) and Marshan et al. (2023), the qualitative aspects of investor sentiment profoundly impact market movements, affecting the desired rate of return of the investors. By harnessing the power of Natural Language Processing (NLP), Deep Learning models can parse and comprehend vast amounts of textual data generated daily, and gauge the collective sentiment of market participants (Sidogi et al., 2021). [1]

Current Machine Learning approaches for stock price prediction primarily rely on autoregressive models such as LSTMs or RNNs (Selvin et al., 2017), (Heaton et al., 2017). The application of modern deep learning approaches for stock price prediction has been limited to the use of the Generative Adversarial Network (GAN) proposed by Goodfellow (2016) or the Bidirectional Encoded Representations from Transformers (BERT) model proposed by Devlin et al. (2019), without extensively exploring the market sentiment for the current stock (Lin et al., 2021).

In this research, we propose a novel neural method for stock price prediction called FB-GAN, which not only relies on historical stock price data, but also leverages the market sentiment for the particular stock in a strategic manner. We explore and assess the performance of our sentiment-enhanced stock price prediction model using multiple strategies for capturing the market sentiment.

The predictive accuracy and performance of the proposed model FB-GAN is rigorously evaluated and compared against contemporary stock price prediction models using appropriate metrics such as RMSE. We demonstrate that our model strategically incorporates market sentiment data along with historical stock prices and outperforms contemporary approaches for stock price prediction.[2]

The major contributions of this paper are summarized below:

- We propose a robust neural framework called FB-GAN based on the BERT and GAN models, which leverages market sentiment in a strategic manner along with stock price history for the prediction of upcoming stock prices.
- We conduct experiments using three different

[1]Manuscript accepted for publication at FinNLP, LREC-COLING 2024 (https://lrec-coling-2024.org/).

[2]The code will be shared publicly upon acceptance.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 85–93

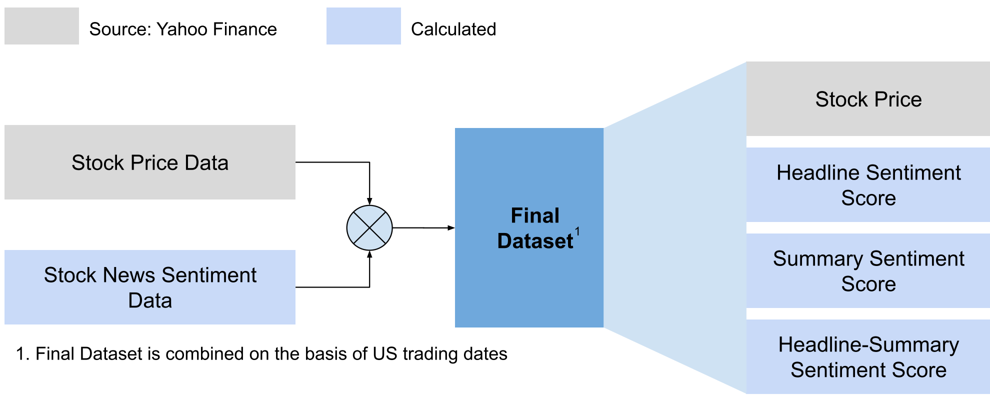


Figure 1: Merging stock price and sentiment score data

strategies to integrate sentiment information from news articles with our stock price prediction model, namely (i) headline, (ii) summary, and (iii) headline and summary combined.

- We demonstrate that the proposed model FB-GAN outperforms contemporary appraoches and the combination of headline and summary of the news articles yields the best results for stock price prediction.

2. Related Work

In this section, we briefly review the state-of-the-art techniques for stock price prediction and highlight their limitations, setting the context of our work.

The stock price prediction task dates back to the 1960s, wherein traditional time series analysis methods were used to capture the serial correlation in stock prices (Fama, 1965). These methods, however, often assume stationarity and are not capable of capturing the complexities of time series data.

The advancements in Machine Learning led to the exploration of neural models such as LSTMs (Hochreiter and Schmidhuber, 1997) and CNNs (LeCun et al., 2015) for stock price prediction. (Mehtab and Sen, 2020), (Heaton et al., 2017), (Selvin et al., 2017). Chung et al. (2014) demonstrated that Gated Recurrent Units (GRU) supplemented LSTM networks, which accelerated the training and mitigated the problem of overfitting. Heaton et al. (2017) suggested that the LSTM neural network could be used as an oscillator and were among the first approaches to demonstrate that deep neural networks can detect patterns in financial data. However, LSTMs may have difficulty distinguishing between meaningful patterns and random noise, especially when the data exhibits high volatility or irregular patterns.

While autoregressive models such as LSTMs and RNNs have been explored extensively for stock price prediction, the application of modern neural network architectures for this task remains relatively unexplored. Lin et al. (2021) explored the usage of Generative Adversarial Networks (GANs) and proposed WGAN-GP, an improved GAN model, to make accurate stock price predictions. However, while the WGAN-GP model yields better performance than the previous model, GAN, it only leverages historical stock prices and is unable to capture additional information such as market sentiments.

Akita et al. (2016) explored the usage of LSTMs to incorporate sentiment analysis for stock price prediction. While this method demonstrated the importance of sentiment analysis for this task, it is based on an outdated neural architecture and is unable to capture the market sentiment in a strategic manner. Devlin et al. (2019) introduced BERT, a Transformer-based language model, greatly impacting a number of NLP tasks. BERT has gained popularity for sentiment analysis, extracting valuable insights from news articles, social media, and financial reports.

With the advent of language models such as BERT, research has been conducted to use sentiment analysis on social media and news data for stock price prediction (Weng et al., 2022; Sidogi et al., 2021). These methods are similar to the one proposed by Akita et al. (2016), and employ LSTM with BERT to predict stock prices based on historical prices, with sentiment analysis done on the news headlines of a set of chosen stocks. While these approaches employ the headlines of the news articles for stock price prediction, they fail to capture the entire sentiment of the news articles and are based on the sub-optimal LSTM framework for the time-series prediction.

We provide substantial arguments that sentiment

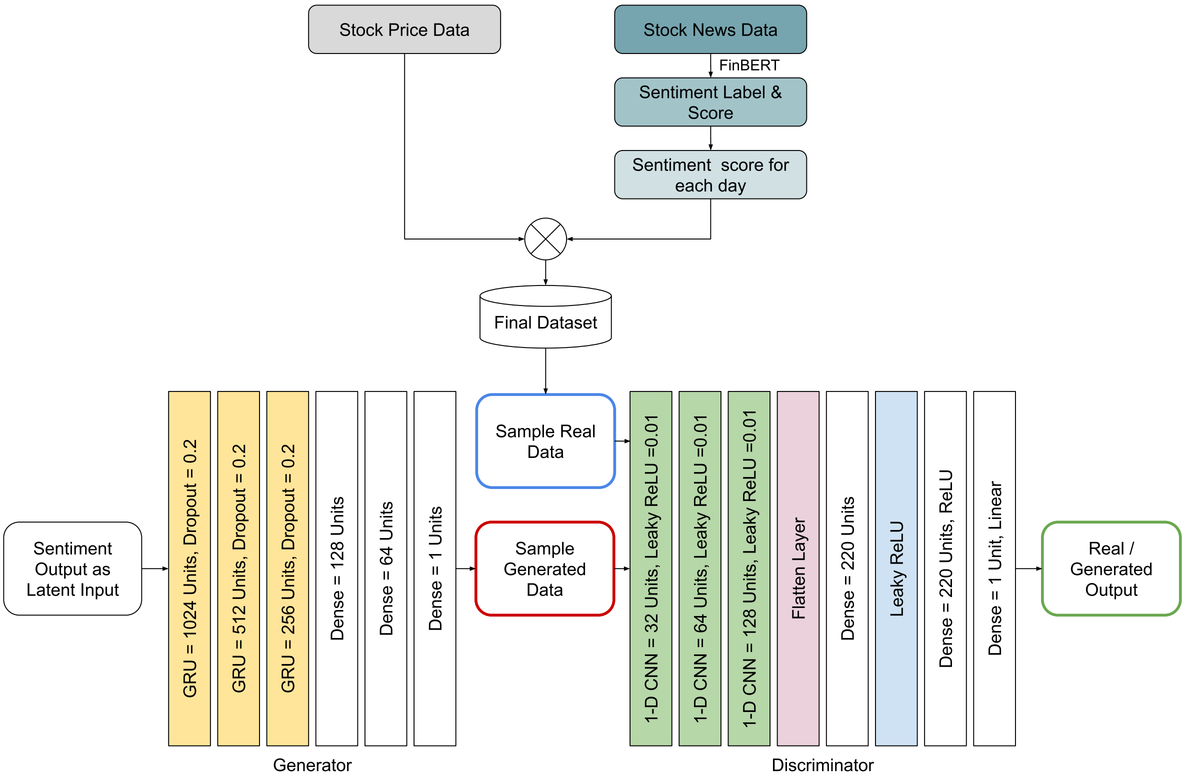


Figure 2: Proposed Model: FB-GAN

analysis done only on headlines can be misleading and result in poor stock price predictions. In addition to highlighting that headlines do not convey the entire sentiment of the news article, we also propose a neural model which improves stock price prediction by leveraging both historical data as well as market sentiment information, which is done by capturing the entire sentiment of the news article in a strategic manner.

3. Methodology

This section introduces the data collection, data preprocessing, feature engineering, experimental setup, study of existing stock price prediction models and our proposed model.

3.1. Data Collection and Preprocessing

For this study, we selected five stocks: Amazon, Apple, Microsoft, Nvidia and Adobe for stock price prediction based on their 5-year stock price history and market sentiments. The data collection was done in two phases for this project. In the first phase, we gathered news articles related to a particular stock, and in the second phase, we collected the historical price history of the stocks. The news articles related to a particular stock were collected using the Alpha Vantage API.[3]

We conduct this study only with high-quality news articles from trustworthy sources. We employ the publicly available news aggregator Alpha Vantage which provides high-quality news articles published by renowned publishers such as The Wall Street Journal, Financial Times, Motley Fool, MarketWatch, etc. We extract information from both the headline and a summary of the news articles, which are essential data points to study the performance of stock price prediction. We extract news articles published during the period 01 Mar 2022 to 31 Jul 2023[4]. The statistics of the data used for our experiments are mentioned in Table 1. The dataset is split randomly into the training and testing sets, such that 80% of the samples are employed for training, and the remaining 20% are used for testing the models.

Historical price data of the stocks was collected using Yahoo Finance's python package yfinance, which gave us data related to a particular stock's close price, open price, high, low and volume for the given time frame. The historical price history collected is from 01 Aug 2018 to 31 Jul 2023.

After performing Exploratory Data Analysis (EDA), we sanitized our dataset to ensure we didn't

[3]Available at https://www.alphavantage.co/

[4]Alpha Vantage does not contain articles published before 01 Mar 2022

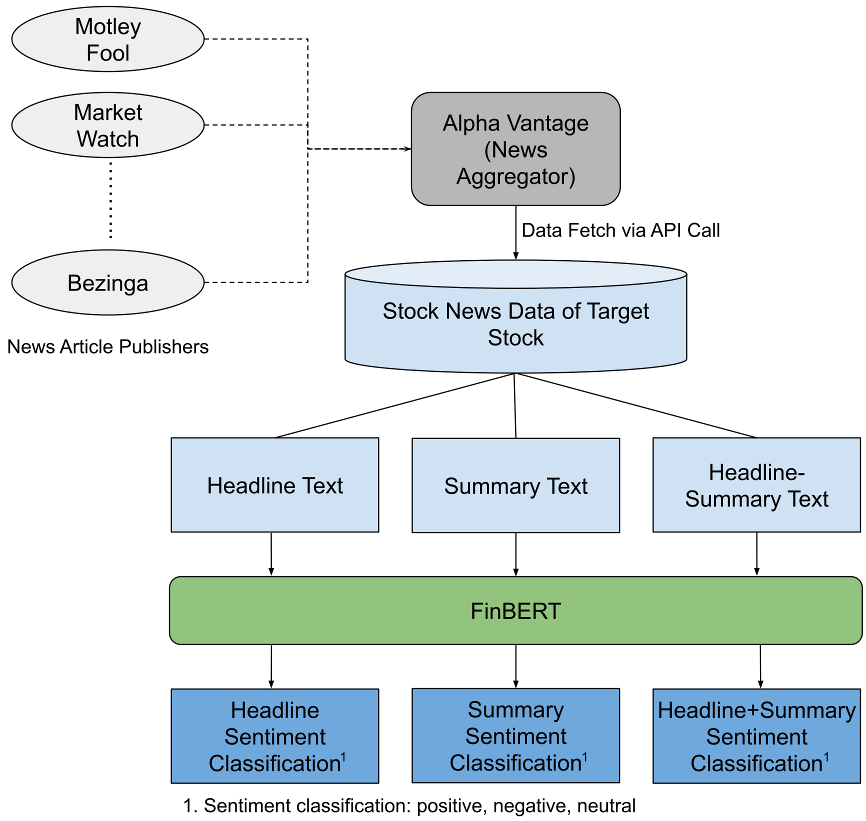


Figure 3: Sentiment Analysis Diagram Flow

Stock Name	Total Articles	Training (80%)	Testing (20%)
Amazon	10.2K	8.2K	2K
Apple	13.8K	11.1K	2.7K
Microsoft	27.6K	22.1K	5.5K
Nvidia	10.8K	8.6K	2.2K
Adobe	1.5K	1.2K	0.3K

Table 1: Count of the news articles captured for each stock from 01 Mar 2022 - 31 Jul 2023 (in thousands)

have any duplicate news articles in our dataset during the data collection process of news articles from the Alpha Vantage API. While performing EDA, it was observed that different news articles vary in terms of their relevance for the stock price prediction task, and some articles could be irrelevant. Hence, in order to ensure that our sentiment analysis is accurate we employ the relevance score provided by Alpha Vantage, which is a measure of how relevant a news article is to a certain stock.

To develop the final dataset used for our experiments, we used the stock price from Yahoo Finance and sentiment scores for each day and combined them based on US Trading dates, as illustrated in Figure 1. The details about the computation of the sentiment scores and strategies to incorporate the same with the stock price history are presented in Section 3.2. While combining the stock price and sentiment data, we assumed market sentiment for a particular day would have an effect on the next day's closing price. To handle the dates with no news articles, we have assumed the sentiment for those dates to be neutral i.e., 0 sentiment score. We pass stock news information of all three types: headline, summary, and headline+summary

3.2. Proposed Model: FB-GAN

Our proposed model, FB-GAN is inspired by WGAN-GP which incorporates market sentiment generated by FinBERT for stock price prediction. FB-GAN has two major components, i.e. the generator and the discriminator. The generator is made up of three GRU Units having 1024, 512 and 256 neurons in the three layers respectively; each layer has a dropout ratio of 0.2, followed by three dense layers. The discriminator is made up of three 1-dimensional Convolutional Neural Networks having 32, 64 and 128 neurons in the three layers respectively, with a flattened layer followed by three dense layers and, finally, the output layer, which used linear activation function. The architecture of the proposed model, FB-GAN, is presented in Figure 2.

As shown in Figure 2, the generator transforms random noise along with sentiment scores as an input into data samples that are indistinguishable from real stock price data. The generator aims to produce data which is realistic enough to fool the discriminator. We feed the discriminator with two sample data i.e. real data and generated data. The discriminator aims to classify the real and fake data correctly. The generator and discriminator work in an adversarial manner, where each one tries to outperform the other. Our proposed model, FB-GAN, is trained on 7 features: Adj. Close, High, Low, Close, Open, and Market Sentiment Score. The market sentiments are fed to the neural network as a latent input (co-variant) along with other inputs.

To categorise a piece of news into a particular category, we performed sentiment analysis on each news article using a language model specialized on financial data known as FinBERT. FinBERT (Araci, 2019) is a pre-trained BERT model fine-tuned for financial sentiment classification. FinBERT analyses a textual input and provides an output between 0 and 1 and the sentiment label: positive, negative and neutral. A higher score indicates a higher confidence in the label. In order to assess and analyse the performance of the type of market sentiment information on the stock price prediction, we conduct ablation studies using three different strategies to compute the sentiment scores:

- Using the headline of the article
- Using the summary of the article
- Using the headline and summary of the article

Each of the above are passed to the FinBERT model to obtain the category label and the confidence score for the given news article, as shown in Figure 3. We scaled the sentiment score, obtained from FinBERT by the relevance score of each article obtained from Alpha Vantage for a true estimation of the overall sentiment of each news article. Since neural networks can only process numerical input, we pre-process the data before feeding it to the network. We feed two types of inputs to the neural network i.e. stock price data and market sentiment data. The stock price data is already in numerical form; however, the output from FinBERT sentiment classification is in the form of textual labels, namely, positive, negative and neutral. In order to transform it to numerical form, we assign positive articles a value of 100, negative articles a value of -100 and neutral articles a value of 0. To calculate the sentiment score for a particular day, we define the Sentiment Score SS_n as follows:

$$SS_n = \frac{\sum_{i=1}^{N}(CS_{pos} \times 100) + \sum_{j=1}^{M}(CS_{neg} \times -100)}{N + M + P} \tag{1}$$

where N, M, and P represent the total number of positive, negative and neutral articles for a particular day, respectively, CS_{pos} (Confidence score - positive) represents the confidence score of the positive article(s), CS_{neg} (Confidence score - negative) represents the confidence score of the negative article(s).
The scores computed using these mechanisms are then fed to our model alongside the stock price history to perform the prediction of the upcoming stock prices. The optimizer used is Adam, with a learning rate of 0.000128, the number of epochs is 160, and a batch size of 128.

4. Experimental Setup

To conduct the experiments, we employ the Python 3 Google Compute Engine. The hardware setup includes a Nvidia Tesla T4 GPU with a 2-core Intel Xenon CPU 2.2 GHz, supported by 13GB RAM and 16 GB GPU Memory. 80% of the samples are used for training, and 20% are used for testing the model. The models are implemented using the deep learning framework Keras, with a Tensorflow backend.

We compare the performance of our model with the following existing approaches:

- Vanila RNN model: The predictions are done based on the Adjusted close price as the input feature (3 days of Adjusted close price to predict the Adjusted close price of the next day). The RNN model comprises 5 layers: 1 input layer, 3 hidden layers, and 1 output layer. The optimizer used is Root Mean Square Propagation (RMSprop), the loss function is the Mean Squared Error (MSE), the number of epochs is set to 100, and the batch size is set to 150.

- LSTM: In the LSTM model, we use a similar input vector as we did in the case of Vanilla RNN, where we use 3 days of Adjusted close price to predict the Adjusted close price of the next day. The LSTM model contains 5 layers: 1 input layer, 3 hidden layers, and 1 output layer. The optimizer is Adam, the loss function is Mean Squared Error (MSE), the number of epochs is set to 100, and the batch size is set to 150. A dropout layer is added after each LSTM layer to prevent overfitting. The dropout ratio is set to 0.2.

- GAN: The generator uses a three-dimensional array of tensors, time steps, and features, similar to the vanilla RNN. The model GAN is trained on 6 features: Adj. Close, Open, High, Low, Close and Volume, using 3-time steps to give the prediction of the next day's Adj. close price. The optimizer used is Adam, with a

	Amazon	Apple	Microsoft	Nvidia	Adobe	Average
Vanilla RNN	5.30	9.34	16.24	25.98	16.66	14.71
LSTM	4.31	6.53	**9.44**	28.19	15.76	12.85
GAN	4.49	12.73	16.74	23.02	17.76	14.95
WGAN-GP	5.03	6.98	18.29	18.30	**14.67**	12.65
FB-GAN (Headline)	4.78	7.61	12.53	15.58	21.10	12.32
FB-GAN (Summary)	4.30	8.13	12.26	19.01	21.67	13.07
FB-GAN (Headline+Summary)	**4.01**	**4.35**	10.08	**14.19**	15.73	**9.67**
Average	4.52	7.49	13.00	20.21	17.21	–

Table 2: Comparison of results of different models based on RMSE.
Best performing model highlighted in bold, second best performing model underlined

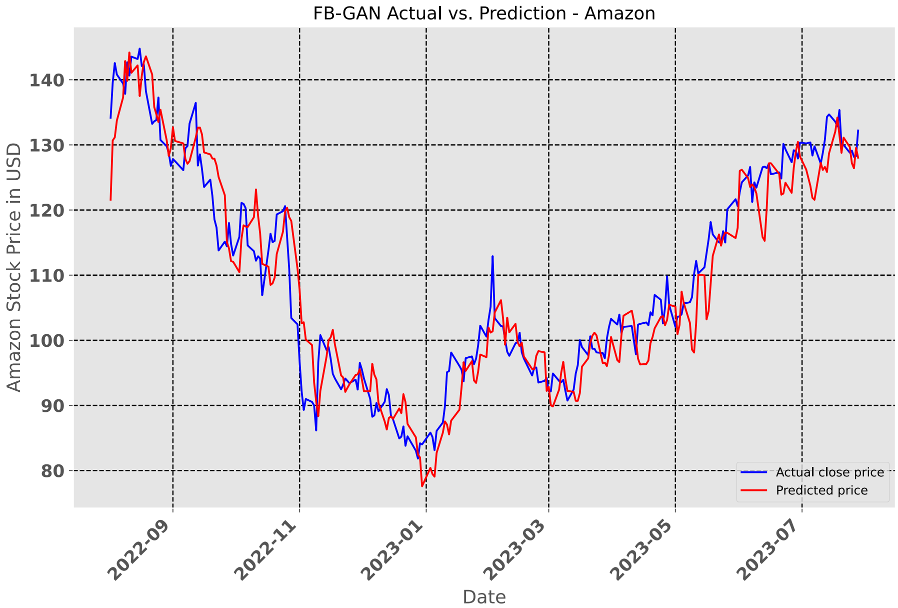


Figure 4: FB-GAN (Headline+Summary) Actual vs. Predicted Stock Price Graph of Amazon

learning rate of 0.00016, the model is trained for 165 epochs with a batch size of 128. Leaky Rectified Linear Unit (ReLU) is used as an activation function among all layers except the output layer, which is a sigmoid activation function. The model is tuned with a learning rate between 0.0003, number of epochs of 300 and a batch size between 64 to 512.

- WGAN-GP: The architecture of WGAN-GP is based on the GAN model; however, the output layer of the discriminator of the WGAN-GP is a linear activation function instead of a sigmoid function, and an additional gradient penalty is added to the discriminator. The optimizer used is Adam, with a learning rate of 0.000115. The model is trained for 100 epochs, with a batch size of 128. The discriminator and generator are the same as the basic GAN; however, the discriminator is trained once, and the generator is trained thrice.

5. Results and Discussions

5.1. Quantitative Analysis

We compare the performance obtained by our proposed model (FB-GAN) with five existing neural network baseline models. The results obtained by the models are presented in Table 2. We used Root Mean Square Error (RMSE) as the evaluation metrics, defined as Equation 2:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \qquad (2)$$

where y_i is the actual (true) value of the i^{th} data point, $\hat{y}_i$ is the predicted value of the i^{th} data point and N is the total number of data points. A lower RMSE value signifies a better model as the predicted values are as close as possible to the target values.

Data Fields	News Article #1	News Article #2	News Article #3
Ticker	AMZN	AMZN	AMZN
Date	25/04/23	07/03/22	10/03/22
Time	15:13:49	16:08:48	13:20:25
Headline	"What's Going On With Amazon Shares - Amazon.com (NASDAQ:AMZN)"	"Why Amazon, Meta Platforms And Microsoft Shares Are Falling Today"	"Why Amazon Shares Are Rising"
Summary	"Amazon.com, Inc. AMZN shares are trading lower by 1.96% to $104.13. The stock is trading lower possibly in anticipation of the company's first-quarter earnings report, confirmed for Thursday's after-hours session."	"Shares of several companies in the broader technology sector, including Amazon.com, Inc. (NASDAQ: AMZN), Meta Platforms Inc (NASDAQ: FB) and Microsoft Corporation (NASDAQ: MSFT), are all trading lower as stocks fall amid the continued escalation of the Russia-Ukraine conflict."	"Amazon.com, Inc. (NASDAQ: AMZN) shares are trading higher by 4.7% at $2,917.75 after the company reported a 20-for-1 stock split and a $10 billion share buyback. Amazon says, subject to shareholder approval of the stock split, each company shareholder of record at the close..."
Headline+Summary	"What's Going On With Amazon Shares - Amazon.com (NASDAQ:AMZN) Amazon.com, Inc. AMZN shares are trading lower by 1.96% to $104.13. The stock is trading lower possibly in anticipation of the company's first-quarter earnings report, confirmed for Thursday's after-hours session."	"Why Amazon, Meta Platforms And Microsoft Shares Are Falling TodayShares of several companies in the broader technology sector, including Amazon.com, Inc. (NASDAQ: AMZN), Meta Platforms Inc (NASDAQ: FB) and Microsoft Corporation (NASDAQ: MSFT), are all trading lower as stocks fall amid the continued escalation of the Russia-Ukraine conflict."	"Why Amazon Shares Are RisingAmazon.com, Inc. (NASDAQ: AMZN) shares are trading higher by 4.7% at $2,917.75 after the company reported a 20-for-1 stock split and a $10 billion share buyback. Amazon says, subject to shareholder approval of the stock split, each company shareholder of record at the close..."
Source	Benzinga	Benzinga	Benzinga
Relevance Score	0.9267	0.5502	0.9836
Headline FL[a]	Neutral	Negative	Positive
Headline FS[b]	1	0.8112	1
Summary FL[a]	Negative	Negative	Positive
Summary FS[b]	0.9999	0.9706	0.9903
Headline+Summary FL[a]	Negative	Negative	Positive
Headline+Summary FS[b]	0.9963	0.9913	0.9999

[a] FL stands for FinBERT Label
[b] FS stands for FinBERT Score

Table 3: Sample News Data after Sentiment Analysis

The results obtained by our proposed model FB-GAN are compared with existing approaches and are presented in Table 2. The effect of stock price prediction with and without sentiment information can be observed by comparing the result obtained by FB-GAN (Headline+Summary) with WGAN-GP, the best-performing stock price prediction model which uses only historical price data. FB-GAN outperforms WGAN-GP by 23.6% in terms of the RMSE value (Table 2, rows 4-6). Additionally, FB-GAN also outperforms other baseline models, namely the RNN, LSTM and GAN models (Table 2, rows 1-3).

In addition to demonstrating the impact of sentiment, we conduct ablation studies using three different strategies to incorporate information from news articles, namely (i) headline, (ii) summary, and (iii) summary and headline combined. FB-GAN yields the best results based on the sentiment obtained from the headline and summary combined for each stock, with an average RMSE of 9.67, followed by the headline sentiment with an average RMSE of 12.32 and lastly using the summary sentiment with an average RMSE of 13.07.

Figure 4 juxtaposes the actual stock price of Amazon against that predicted by the proposed model FB-GAN on the test data, using the Headline+summary strategy. It can be observed that our predicted stock price mimics the actual close price very closely, demonstrating its efficacy for stock price prediction. Although each stock comprises complex time series data, our FB-GAN model performs well in predicting the stock price of each stock.

5.2. Qualitative Analysis

In addition to providing a quantitative analysis of the results, we also present a qualitative analysis of the results obtained by FB-GAN, and demonstrate the importance of sentiment analysis in stock price prediction by comparing the different strategies to capture market sentiment.

Table 3 presents a sample of our stock news dataset after performing sentiment analysis using FinBERT (refer to Figure 3). In our stock news dataset, we have three text parameters: Headline, Summary and Headline+Summary, which we had passed through FinBERT and obtained the FinBERT Label (FL) and FinBERT Score (FS) for each parameter. The FinBERT Label can be any one of three labels: Positive, Negative and Neutral; and the FinBERT Score can be any value between 0 and 1, where a lower score represents low confidence and higher score represents high confidence.

On comparing the Headline FL, Summary FL and Headline+Summary FL of News Article #1 of Table 3, we observe Headline FL is classified as Neutral, Summary FL is classified as Negative and Headline+Summary is classified as Negative. In general, news article headlines could be incomplete and misleading to attract readers' attention and could lead to incorrect classification when sentiment analysis is performed on them. The headline of News Article #1, *"What's Going On With Amazon Shares"* may spark curiosity in reader; and while a human might delve deeper to understand the topic to make an informed opinion, an ML-model might fail to capture the sentiment based on the headline alone. Previous studies have solely relied on the headlines, for incorporating sentiment analysis in stock price prediction (Sidogi et al., 2021; Weng et al., 2022; Li et al., 2023). In this case, the headline suggests neutrality, but the summary paints a negative picture, with *"Amazon shares trading lower by 1.96% to $104.13"*. This highlights the limitation of relying solely on headlines or summaries for sentiment analysis, as they may only present half the picture. By combining the headline and summary, FinBERT can accurately classify the article as negative with 99.63% confidence, demonstrating the importance of complete information for accurate stock price prediction. Similarly, for News Article #2 of Table 3, on comparing the FinBERT Score (FS) obtained after passing Headline, Summary and Headline+Summary through FinBERT, we observe that FS of Headline+Summary is 99.13%, followed by FS of Summary which is 97.06%, followed by FS of Headline which is 81.12%, which proves statistically as well that Headline+Summary provides a higher confidence on the estimated label than its counterparts.

Based on News Article #3 of Table 3, we perform pre-hypothesis testing, where we compare the results from the FinBERT classification with the actual stock price movement. It can be observed that based on the Headline and Summary of the News Article - *"Why Amazon Shares are rising. Amazon [. . .] shares [. . .] trading higher by 4.7% at $2917.75 after the company reported 20-for-1 stock split and a $10 billion share buyback [. . .],"* the News Article is classified as positive with a confidence score of 99.99% and following this news, the stock price shows a bullish (upward) trend for several days.

The qualitative analysis thereby corroborates the finding that our sentiment-enhanced model yields improved performance owing to the correlation between market sentiment and stock price movement. It also confirms that the Headline+Summary combined strategy provides a more accurate estimation of the sentiment than individual strategies, leading to better stock price prediction.

6. Conclusion and Future Work

This paper presents a novel sentiment-enhanced neural model called FB-GAN, and demonstrates that that it outperforms existing approaches for stock price prediction. The experimentation validates our hypothesis that integrating market sentiment in a strategic manner using state-of-the-art language models improves the performance of stock price prediction. We demonstrate that the Headline & Summary combined strategy yields the best results for stock price prediction (an improvement of 21.5% and 26% respectively in the average RMSE scores when considering Headline alone and summary alone respectively).

Future directions to improve our proposed model could be inspired from the Efficient Market Hypothesis (EMH), wherein more correlated factors, such as gold prices, bank rates, etc., are leveraged while training the model for stock price prediction. Another possible direction for future work involves modifying our proposed model to consider the real-time stock price and market sentiment data to predict the stock prices which can be used for Intra-day trading.

7. Bibliographical References

Ryo Akita, Akira Yoshihara, Takashi Matsubara, and Kuniaki Uehara. 2016. Deep learning for stock prediction using numerical and textual information. *2016 IEEE/ACIS 15th International Conference on Computer and Information Science, ICIS 2016 - Proceedings*.

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *CoRR*, abs/1908.10063.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova Google, and AI Language. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, pages 4171–4186.

Eugene F. Fama. 1965. The behavior of stock-market prices. *The Journal of Business*, 38:34.

Ian Goodfellow. 2016. *Deep learning / Ian Goodfellow, Yoshua Bengio and Aaron Courville.*

J. B. Heaton, N. G. Polson, and J. H. Witte. 2017. Deep Learning for Finance: Deep Portfolios. *Applied Stochastic Models in Business and Industry*, 33:3–12.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9:1735–1780.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

J.-B. Li, S.-Y. Lin, F.-Y. Leu, and Y.-C. Chu. 2023. Stock Price Trend Prediction Using LSTM and Sentiment Analysis on News Headlines. Advances on Broad-Band Wireless Computing, Communication and Applications: Proceedings of the 17th International Conference on Broad-Band Wireless Computing, Communication and Applications (BWCCA-2022). Lecture Notes in Networks and Systems (570), pages 282 - 91, Berlin, Germany.

Hung Chun Lin, Chen Chen, Gao Feng Huang, and Amir Jafari. 2021. Stock price prediction using Generative Adversarial Networks. *Journal of Computer Science*, 17:188–196.

A. Marshan, M. Mbedzi, and A. Ioannou. 2023. Exploring the relationship between news articles and stocks market movements: A sentiment analysis perspective. Intelligent Computing: Proceedings of the 2023 Computing Conference. Lecture Notes in Networks and Systems (739) vol.2 pages 1167 - 83, Berlin, Germany.

Sidra Mehtab and Jaydip Sen. 2020. Stock price prediction using CNN and LSTM-based Deep learning models. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 447–453. IEEE.

Sreelekshmy Selvin, R Vinayakumar, EA Gopalakrishnan, Vijay Krishna Menon, and KP Soman. 2017. Stock price prediction using LSTM, RNN and CNN-sliding window model. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1643–1647. IEEE.

Thendo Sidogi, Rendani Mbuvha, and Tshilidzi Marwala. 2021. Stock price prediction using sentiment analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 46–51.

Xiaojian Weng, Xudong Lin, and Shuaibin Zhao. 2022. Stock price prediction based on LSTM and BERT. In *2022 International Conference on Machine Learning and Cybernetics (ICMLC)*, pages 12–17.

Bing Han Xiaodong Liu and Luanfeng Li. 2023. Impact of investor sentiment on portfolio. *Emerging Markets Finance and Trade*, 59(3):880–894.

Unveiling Currency Market Dynamics: Leveraging Federal Reserve Communications for Strategic Investment Insights

Martina Menzio[1,2], **Davide Paris**[1,2], **Elisabetta Fersini**[2]
[1]Intesa Sanpaolo
[2] University of Milano-Bicocca
{martina.menzio,davide.paris}@intesasanpaolo.com
elisabetta.fersini@unimib.it

Abstract

The purpose of this paper is to extract market signals for the major currencies (EUR, USD, GBP, JPY, CNY) analyzing the Federal Reserve System (FED) minutes and speeches, and, consequently, making suggestions about going long/short or remaining neutral to investors thanks to the causal relationships between FED sentiment and currency exchange rates. To this purpose, we aim to verify the hypothesis that the currency market dynamics follow a trend that is subject to the sentiment of FED minutes and speeches related to specific relevant currencies. The proposed paper has highlighted two main findings: (1) the sentiment expressed in the FED minutes has a strong influence on financial market predictability on major currencies trend and (2) the sentiment over time Granger-causes the exchange rate of currencies not only immediately but also at increasing lags according to a monotonically decreasing impact.

Keywords: Federal Reserve Communications, Sentiment, Currency Exchange Rate

1. Introduction

Within the financial sector, the qualitative analysis of central bank communications, encompassing Federal Reserve (FED) minutes and speeches, has emerged as a crucial practice for investors to predict market trends, evaluate economic conditions, and guide strategic decision-making policies. The role of the FED has been recently investigated in (Benchimol et al., 2021) to study the changes across different communication types (Fed fund rate announcements, Federal Open Market Committee minutes, and Fed chairman speeches) during a few economic crises (Global Economic Crisis, Dot-Com Bubble and COVID-19) paying particular attention to financial stability and monetary policies. Additional evidence about the importance of the FED communications and the corresponding sentiment outlook can be grasped in (Wischnewsky et al., 2021), where the authors highlighted that a negative sentiment (estimated on the Humphrey–Hawkins hearings) matters to a greater extent than positive sentiment to the financial stability. A similar conclusion can be derived in (Tadle, 2022) where the FED document's sentiment has been shown as to proxy to predict interest rate tilt. Finally, a very recent investigation is reported in (Fischer et al., 2023) where the authors estimated the extent to which market-implied policy expectations could be improved with further information disclosure from the FED documents, highlighting that the forecasting of future monetary policy could be strongly affected by the sentiment of FED communications.

Although the above-mentioned investigations represent a fundamental step towards the understanding of the FED communications role to explain the general marked behaviours, they are focused on coarse-grained document sentiment (entire document or at most topics), do not pivoting on specific currencies, and on long-term impact on monetary policies, do not assessing the short term implications on the Forex market.

In this paper, we provide two main contributions:

1. FedSent Index: we introduce a metric to evaluate the content of FED meeting minutes and speeches. We create an index that proxies the sentiment expressed in the FED meetings, which has a strong influence on financial market predictability on major currency trends.

2. Forex Market Sentiment Impact: we demonstrate that the sentiments expressed in FED minutes have a significant influence on financial market predictability, especially on major currency trends.

2. Related Work

The literature about Natural Language Processing techniques related to the currency market dynamics has received several efforts in the last five years. The main contributions can be roughly distinguished in two main directions: (1) language models and tools for the broad-ranging financial sector and (2) investigations about specific currencies focusing on different sources of information to perform long-term predictions.

In the first area, the panorama is dominated by several models such as FinBERT (Liu et al., 2021), FLANG (Shah et al., 2022), InvestLM (Yang et al., 2023), FinMA (Xie et al., 2023), BloombergGPT

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 94–102
20 May, 2024.

(Wu et al., 2023) and FinGPT(Liu et al., 2023), where most of them require considerable computational resources to function optimally, making their implementation challenging. On the other hand, the investigations about specific currency are still in their infancy especially focusing on FED communications. While in (Seifollahi and Shajari, 2019), the authors proposed an NLP-based model employing news headlines to predict the upward and downward trends of a Forex currency pair, in (Lee et al., 2021) an interpretable and user-friendly Natural Language Processing (NLP) system has been developed to decode Federal Reserve communications providing tools to deal with sentiment analysis, topic modelling and summarization without deepening the relationships between the available communication and specific market behaviour. Additional investigations relate to the use of Deep Learning techniques for forecasting foreign exchange volatility (Jung and Choi, 2021). Notably, these approaches have been explored without incorporating exogenous variables, offering intriguing perspectives for central banks and financial institutions seeking to enhance their forecasting strategies.

In what follows, we will bridge the gap by extracting sentiment market signals for the major currencies (EUR, USD, GBP, JPY, CNY) analyzing the FED minutes and speeches, and consequently, making suggestions about going long/short or remaining neutral to investors.

3. Federal Reserve Communications

3.1. Data Collection

The datasets used in our paper were sourced from the official websites of the Federal Reserve to obtain distinct time series for speeches, statements, and minutes. Three separate datasets were compiled, each containing pertinent details regarding speeches, statements, or minutes. These datasets include information such as the URL link to access the data, title, date, text content, and associated paragraphs. The datasets contain communications from 1993 to (September) 2023, obtaining 1.671 speeches, 252 minutes and 224 statements (details as reported in Appendix 1 (Table 4). From an initial overview of the collected communications, we can highlight two main aspects:

- publication of minutes and statements are almost constant over the years (∼8 per year per data source), starting from 1993 and 1994 for the minutes and statements respectively.

- publication of speeches is the most variable over time, due to the larger number of publications per year, starting from 1996.

Subsequently, we examined the lexical diversity within the text, encompassing all words present in the documents while excluding stopwords, for minutes, speeches, and statements individually. Our findings align closely with Zipf's Law (Piantadosi, 2014). For instance, when analysing the distribution of terms in the minute dataset, particularly on the left side where frequently occurring words can be observed, notable terms indicative of market trends and sentiment (such as *increase*, *decline*, *risk* and *rise*) emerged. This observation suggests the hypothesis that the currency market dynamics could be related to the sentiment embedded within the FED documents. Analogous observations can be drawn also for speeches and statements.

3.2. Forex Data

In this section we focus our attention to specific Forex data, considering only those minutes, speeches and statements which contain at least one keyword related to the following currency: EUR (euro, €, EUR), USD ($, USD, dollars), GBP (GBP, pounds, sterling), JPY (JPY, yens), CNY (CNY, yuan, renminbi), and general (fx, forex, currency, currencies).

We finally obtained the following datasets to be used in the subsequent analysis:

1. Minutes: all of them mention at least one FX $(100\% - 252/252)$, with a medium number of days between citations of ~ 44 days;

2. Speeches: almost half of them contain at least one FX $(47.34\% - 791/1'671)$, with a medium number of days between citations of ∼12 days;

3. Statements, in the last 30 years, quote in just 6 documents at least one FX keyword $(2.68\% - 6/224)$, with a medium number of days between citations of ∼880 days. Given the reduced number of available observations, the Statements dataset has been disregarded.

Given the Minutes and Speeches datasets, only those sentences containing the above-mentioned currencies have been considered (see the distributions reported in Appendix 1 Figure 4 and 5).

According to the resulting selection, minutes appear longer than speeches, being in line with what we expect. Minutes are published less frequently than speeches therefore containing more sentences mentioning the considered Forex. In the end, however, only a few sentences contain the considered FX keywords: only 4.60% of sentences contain at least one Forex in Minute documents, while for speeches the percentage is 1,62%.

4. Fed Sentiment Index

The core idea is to consider each sentence mentioning an FX and subsequently compute the corresponding overall sentiment index for the referenced FX. To this purpose, we exploited one of the most widely used language models known as *FinBERT* (Liu et al., 2021) to classify each sentence (mentioning a given FX) as positive, negative and neutral and obtain the corresponding probability distribution. To exemplify, we provide FinBERT's score of USD sentences of the Fed minute relative to September 21st 2022 in Table 1.

Let c be a given currency and S_{cd}^{t} the set of sentiments obtained from FinBert for those sentences in given FED document d at timestamp t that mentions c. At time stamp t the **FedSent Index (FSI)** can be estimated as mean or median aggregation of sentiment probabilities. In particular, $FSI_{\mu}(c,t)$ representing a mean aggregation of sentiment probabilities is computed as:

$$FSI_{\mu}(c,t,d) = \begin{cases} \dfrac{\sum_{i=1}^{n} p_i^{=} \times \epsilon}{n} & \text{if } \bigcup_{i=1}^{n} S_{cdi}^{t} = neutral \\ \dfrac{\sum_{i=1}^{k} p_i^{+} - \sum_{i=1}^{m} p_i^{-}}{k+m} & \text{otherwise} \end{cases} \tag{1}$$

where:

- $p_i^{=}$, p_i^{+} and p_i^{-} denotes the probability of a sentence i of being neutral, positive or negative respectively;
- $\bigcup_{i=1}^{n} S_{cdi}^{t}$ denotes the unique sentiment values obtained from FinBERT related to document d mentioning c at timestamp t (with $n = |S_{cdi}^{t}|$)
- k and m represent the number of sentences that are respectively predicted as positive and negative.

An analogous estimate could be computed similarly by adopting a median aggregation of sentiment probabilities. We will denote such median aggregation as $FSI_{\tilde{\mu}}(c,t,d)$.

In practice, for those documents containing only *neutral* sentences mentioning the given FX, we computed the FSI by median the mean scores multiplied by coefficient ϵ to obtain an aggregated score close to 0. For those documents containing at least polarized sentences all neutral probabilities are disregarded, computing the mean and the median of the residual non-neutral probabilities. In this way, for each document d (speech or minute) and for each FX (EUR, USD, GBP, JPY, CNH and general currencies), we get an overall sentiment index.

In order to take into account how specific a minute/speech is with respect to a given FX, we computed a **specificity coefficient**. Such coefficient μ is estimated as the ratio between the number of sentences mentioning an FX and the total number of sentences in the considered minute/speech:

$$\beta_{ctd} = \frac{r_{ctd}}{r_{dt}} \tag{2}$$

where r_{ctd} represents the number of sentences in document d at time stamp t that mention a currency c, while r_{dt} denotes the total number of sentences contained in document d at timestamp t. The specificity coefficient β_{ctd} tends to 1 where all sentences in a document mention an FX at least once.

The above-mentioned FedSent Indexes $FSI_{\mu}(c,t,d)$ and $FSI_{\tilde{\mu}}(c,t,d)$ can be finally smoothed according to the specificity coefficient β_{ctd} by simple multiplication. In Figure 1 and Figure 2 the time series computed as $FSI_{\mu}(c,t,d) \times \beta_{ctd}$ and $FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$ are reported for EUR and USD. The time series of all currencies are reported in Appendix 2.

(a) Smoothed (median) FedSent Index time series, i.e. $FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$, for USD and EUR.

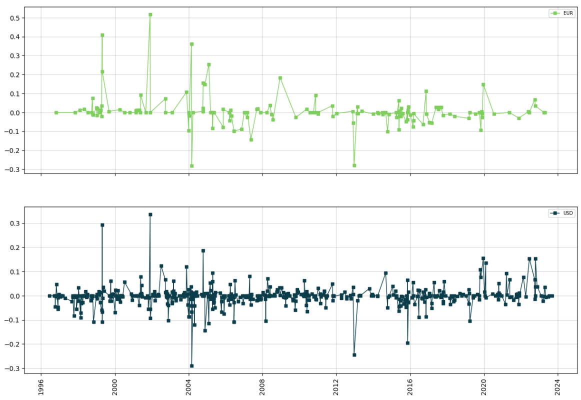
(b) Smoothed (mean) FedSent Index time series, i.e. $FSI_{\mu}(c,t,d) \times \beta_{ctd}$, for USD and EUR.

Figure 1: Smoothed FedSent Index on Minutes.

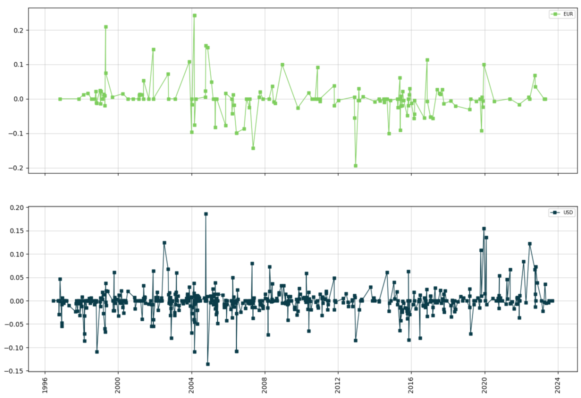
Analyzing the time series of Minutes and Speeches sentiment scores, we observe that:

SENTENCE	LABEL	PROBABILITY
The exchange value of the dollar appreciated notably, reaching multi-decade highs in real terms, as market participants perceived mounting economic challenges abroad.	positive	0.96
The U.S. dollar appreciated further against most major currencies, reaching multi-decade highs against the euro, the British pound, and the Japanese yen.	positive	0.94
The dollar's strength largely reflected increasing investor concerns about the global growth outlook as well as widening interest rate differentials between the United States and Japan.	positive	0.83

Table 1: Example of FinBERT sentiment across USD sentences in Fed minute of September 21^{st} 2022.



(a) Smoothed (median) FedSent Index time series, i.e. $FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$, for USD and EUR.



(b) Smoothed (mean) FedSent Index time series, i.e. $FSI_{\mu}(c,t,d) \times \beta_{ctd}$, for USD and EUR.

Figure 2: Smoothed FedSent Index on Speeches.

- CNH are rarely quoted in speeches (only 3 times in the whole dataset) and never mentioned in minutes;
- GBP and JPY are mentioned only a few times within the documents and there is a long time between signals: their score time series are discontinuous over time;
- the EUR and USD are the FX most mentioned by the FED's minutes and speeches.

Given the frequency and non-discontinuity of EUR and USD sentiment scores within the time series and the expected relevance of the FED communications with USD, we decided to focus on such currency in the subsequent analysis. In particular, from now onward, we focus on USD sentiment time series to check if there exists a relationship between the computed score with the exchange rate USD/EUR over time. To this purpose, we downloaded USD/EUR exchange rate time series from the official site of *Banca di Italia*, obtaining all the estimates available from January 1999 to September 2023. Since the historical USD currency has a daily frequency, it is necessary to have a comparable sentiment score time series at daily basis. For this purpose, the sentiment index between two subsequent communications has been estimated according to the following imputation methods:

- **Ffill**: we repeat the last available sentiment score until another value is found. From a financial point of view, we are considering that the sentiment between a minute/speech and the next one remains constant between two adjoining communications;
- **Exponential Decay**, with different decay rates: in particular we use 0.1, 0.05, 0.01 and 0.001. In this case, the sentiment score degrades over time representing a scenario where the sentiment index at the FED communication day has more relevance to the currency value on the corresponding day and then a decreasing impact;
- **Most Recent Value**: in this case our assumption is that we can estimate the sentiment index between two consecutive FED communications using the last available one associated with previous dates with a similar percentage change in currency;
- **Delta Median**: in this case, we replace the missing values of sentiment score with the median of previous ones associated at previous dates with a similar percentage change in currency.

5. Forex Market Sentiment Impact

We compute the percentage change of the USD currency with respect to the previous day, and then we shift this variation using 1-4 days lags, to verify if there exists a relationship over time with the sentiment indices previously introduced.

5.1. Correlations

As a first analysis, we computed the *Pearson correlation* (Pearson, 1895) coefficients, separately per minutes and speeches considering all the imputation methods but also non-filled sentiment scores. In Table 2, the most relevant results for minutes are shown. First of all, we can observe that the most important correlations are positive. This is in line with what we hypothesized: sentiment increases, the exchange is favourable and therefore the currency price increases, i.e. they are positively correlated.

Focusing on **minutes** correlations, in a nutshell, we find out that:

- all non-filled scores are positively correlated with 3-days shift of percentage variations ($\sim 20\%$ of correlation). In our analysis, we observe positive correlation on the day of the Federal Reserve minute. Subsequently, the following day displays an inclination towards a negative correlation. This discernible pattern resembles a characteristic behavior in financial markets, where, following a noteworthy event, initial days witness a depreciation in market value, followed by an attempt at rebounding on the subsequent day. It is important to note, that in the subsequent two and three days, the correlation consistently remains positive.
- Adopting imputation strategies on the FedSent Index time series, we obtain the highest correlation percentage between the daily percentage change and $FSI_{\mu}(c,t,d) \times \beta_{ctd}$, getting 26.45% of correlation, immediately followed by $FSI_{\mu}(c,t,d)$, which obtains 25.23%.

This is reflected in Figure 3, corresponding to the minute of September 21st 2022 whose sentiment was positive (Table 1). In such a figure, we observe the rising trend of the USD over the next 5 days, which is strongly correlated with the sentiment previously estimated. Therefore, this example corroborates the hypothesis that the FedSent Index has a relevant impact on the USD.

Regarding **speeches**, where no relevant correlation coefficient has been found, we can hypothesize that, while mentioning the USD, the documents are not specific enough to influence the dollar's trend, and their impact may be overshadowed or moderated by other significant drivers, captured more specifically in the minutes.

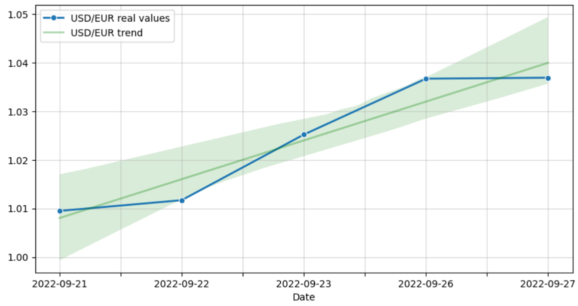


Figure 3: Trend of USD dollar

5.2. Granger Causality

In this section we approach the Granger causality test (Granger, 1969; Shojaie and Fox, 2022) to establish causation by predicting the actual state of a currency using past estimates. Specifically, in our case the sentiment index is the *Grangercause* of the currency exchange value (USD/EUR) if and only if the sentiment index of minutes uniquely improves the predictability of the currency exchange value. This implies that that when forecasting the future states of the currency exchange value (USD/EUR) based on its own past values can be improved when the past sentiment index is also included in the model. In particular, we aim to test the null hypothesis $H0$, i.e. the sentiment score time series does not Granger cause the percentage daily change of USD dollar. More specifically, Granger causality means that past values of the sentiment score have a statistically significant effect on the current value of percentage daily change. By rejecting the null hypothesis, we assume that sentiment score time series Granger causes a percentage daily change of USD dollar if the p-values (using F-Test) are below a given value (0.05).

We test all the sentiment index time series with percentage change of USD/EUR with different lag (from 1 to 4) focusing on those scenarios with the highest positive correlations highlighted in Table 2.

In Table 3 we present the p-values when comparing the sentiment time series identified within minutes with the USD series of LAG at 1-2-3 and 4-days. Focusing on Non-Filled sentiment time series, and in particular, on the mean sentiment index multiplied by specificity coefficient (i.e., $FSI_{\mu}(c,t,d) \times \beta_{ctd}$), we can observe that the p-values for all tests and lags (1, 2, 3, and 4) consistently register a estimation below the 0.05 threshold. Consequently, in such cases, we reject the null hypothesis suggesting that there is a Granger causality between the sentiment score and the lagged percentage change of USD/EUR time series. From a financial point of view, this means that when a minute is published, the corresponding sentiment index per currency has an impact on the subsequent four

Score type	Score	% Daily Change	Shift 1-day	Shift 2-days	Shift 3-days	Shift 4-days
Non-Filled	$FSI_{\tilde{\mu}}(c,t,d)$	10.46%	−1.70%	13.56%	**21.94%**	5.93%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	12.47%	−1.15%	15.41%	**22.18%**	3.27%
	$FSI_{\mu}(c,t,d)$	12.04%	−7.15%	12.67%	15.51%	5.07%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	12.22%	−5.80%	13.67%	**20.20%**	2.48%
Ffill	$FSI_{\tilde{\mu}}(c,t,d)$	0.94%	0.65%	0.92%	1.63%	2.00%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	0.76%	0.31%	0.80%	1.75%	2.05%
	$FSI_{\mu}(c,t,d)$	0.89%	0.41%	0.78%	1.23%	1.65%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	0.77%	0.22%	0.73%	1.70%	1.96%
Exponential Decay (0.1)	$FSI_{\tilde{\mu}}(c,t,d)$	2.65%	2.08%	3.05%	4.74%	4.51%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	2.21%	1.78%	2.91%	4.67%	4.34%
	$FSI_{\mu}(c,t,d)$	3.15%	2.03%	2.81%	3.91%	3.82%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	2.36%	1.47%	2.45%	4.12%	3.90%
Exponential Decay (0.05)	$FSI_{\tilde{\mu}}(c,t,d)$	2.28%	1.93%	2.65%	4.06%	4.07%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	1.98%	1.68%	2.54%	4.02%	3.95%
	$FSI_{\mu}(c,t,d)$	2.57%	1.83%	2.47%	3.42%	3.51%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	2.06%	1.44%	2.21%	3.62%	3.59%
Exponential Decay (0.01)	$FSI_{\tilde{\mu}}(c,t,d)$	1.28%	1.01%	1.40%	2.29%	2.59%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	1.11%	0.73%	1.31%	2.40%	2.63%
	$FSI_{\mu}(c,t,d)$	1.32%	0.82%	1.26%	1.85%	2.20%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	1.15%	0.63%	1.20%	2.29%	2.48%
Exponential Decay (0.001)	$FSI_{\tilde{\mu}}(c,t,d)$	0.97%	0.68%	0.97%	1.70%	2.06%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	0.80%	0.35%	0.85%	1.82%	2.12%
	$FSI_{\mu}(c,t,d)$	0.94%	0.45%	0.83%	1.30%	1.71%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	0.81%	0.27%	0.78%	1.77%	2.01%
Most Recent Value	$FSI_{\tilde{\mu}}(c,t,d)$	18.55%	−0.61%	−1.15%	1.07%	0.69%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	**21.38%**	0.76%	−1.62%	1.01%	0.74%
	$FSI_{\mu}(c,t,d)$	18.29%	−1.39%	−0.43%	1.25%	0.28%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	19.68%	0.41%	−0.78%	1.16%	0.69%
Delta Median	$FSI_{\tilde{\mu}}(c,t,d)$	19.29%	0.52%	−1.12%	1.33%	−1.23%
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	**22.76%**	2.35%	−0.48%	2.69%	−1.33%
	$FSI_{\mu}(c,t,d)$	**25.23%**	0.26%	−0.60%	1.09%	−1.27%
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	**26.45%**	2.43%	0.11%	2.84%	−1.59%

Table 2: Correlations related the USD currency. The coloured cells are those with the highest correlation found with daily percentage change and its shift by 1 to 4 days.

Score type	Score	LAG-1	LAG-2	LAG-3	LAG-4
Non-Filled	$FSI_{\tilde{\mu}}(c,t,d)$	0.2372	0.3962	0.3666	0.3789
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	0.0275*	0.0541	0.0328*	0.0681
	$FSI_{\mu}(c,t,d)$	0.0912	0.2507	0.3603	0.5632
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	0.002*	0.0118*	0.0251*	0.0389*
Most Recent Delta	$FSI_{\tilde{\mu}}(c,t,d)$	0.5209	0.199	0.1688	0.2325
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	0.0964	0.2236	0.3044	0.0562
	$FSI_{\mu}(c,t,d)$	0.85	0.3885	0.4897	0.4698
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	0.1351	0.3236	0.4654	0.0552
Delta Median	$FSI_{\tilde{\mu}}(c,t,d)$	0.9225	0.0688	0.0772	0.065
	$FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$	0.3799	0.0962	0.1391	0.0127*
	$FSI_{\mu}(c,t,d)$	0.4652	0.1839	0.3333	0.2718
	$FSI_{\mu}(c,t,d) \times \beta_{ctd}$	0.3085	0.1902	0.3164	0.032*

Table 3: F-Test p-values of Granger causality to test whether USD sentiment score causes daily percentage change at different lags (1 to 4).

days. As expected, however, the p-values increase at increasing lags, denoting a decreasing impact of the sentiment extracted from the Fed Minutes.

Another important aspect to underline is the role of the specificity score when replacing the missing values according to the Delta Median strategy. In this case, the p-values (0.0127 and 0.032) at lag-4 suggest that the estimated sentiment index series can be used to forecast the USD/EUR exchange rate 4 days later than the availability of the FED minutes.

These considerations would help when making decisions about going short/long on currencies. If the sentiment is negative at a given timestamp t, then it implies that the exchange rate USD/EUR will decrease, suggesting to sell the USD currency at t, buying EUR. On the contrary, if the sentiment is positive at a given timestamp t, then it implies that the exchange rate USD/EUR will increase, suggesting buying the USD currency at t, selling EUR.

6. Conclusion

In conclusion, our analysis finds out a significant correlation between the sentiment expressed in Federal Reserve (FED) meeting minutes and the percentage change of the USD dollar. Through our investigation, we have demonstrated that shifts in sentiment reflected in these crucial documents tend to influence market perceptions and subsequently impact the value of the USD dollar. This correlation highlights the importance of carefully monitoring FED communications and sentiment analysis as integral components of currency market analysis and forecasting. Understanding the nuanced implications embedded within FED minutes can offer valuable insights for investors, policymakers, and financial institutions navigating the complexities of the global economy: it becomes increasingly evident that sentiment analysis remains a pivotal tool for deciphering market movements and informing strategic decision-making in the realm of international finance. Although this study provides valuable insights, further in-depth investigations into additional determining factors influencing currency markets are necessary to enrich the breadth of analysis. Factors such as GDP, inflation rates, and unemployment levels of countries utilizing the currencies may have an impact on currency exchange rates. Furthermore, currency market trends may also be influenced by speeches or communications issued by other central banks (e.g., the European Central Bank), forecasts issued by other industry experts, specific social signals within the industry, financial agencies, and regulatory institutions pertaining to both currencies under consideration.

7. Additional material

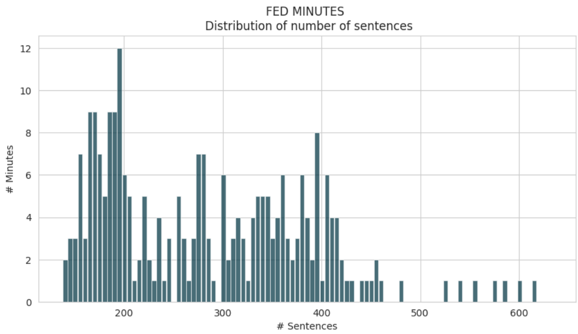
We report here, a set of statistics and distributions related to the available dataset gathered from the FED official site `https://www.federalreserve.gov`. In Table 4 the number of Minutes, Statements and Speeches from 1993 to 2023.

Year	#Minutes	#Statements	#Speeches
2023	5	6	57
2022	8	8	49
2021	8	8	69
2020	8	12	53
2019	8	9	81
2018	8	8	44
2017	8	8	59
2016	8	8	44
2015	8	8	54
2014	8	8	41
2013	8	8	53
2012	8	8	41
2011	8	8	48
2010	8	9	60
2009	8	8	55
2008	7	11	73
2007	10	10	72
2006	8	8	73
2005	8	8	87
2004	8	8	102
2003	8	8	71
2002	8	8	76
2001	12	11	58
2000	9	8	62
1999	8	6	68
1998	8	3	57
1997	8	1	45
1996	8	1	19
1995	8	3	0
1994	9	6	0
1993	8	0	0

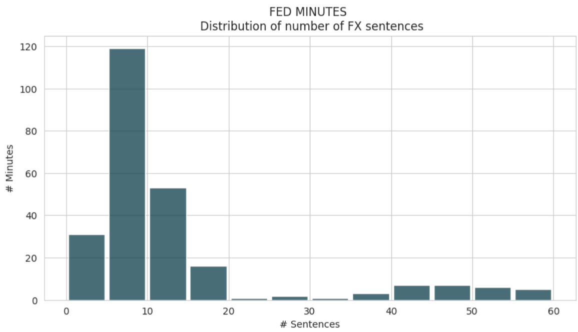
Table 4: Data available per year for each source.

Additionally, we depict in Figure 4 and 5 the overall distributions of the sentences and the corresponding distributions of sentences mentioning an FX, for Minutes and Speeches respectively.

In Figure 6, we report the time series of the smoothed FedSent Index related to Minutes and Speeches, considering both median and mean aggregation functions. As mentioned before, the time series concerned with GBP, JPY and CNY are discontinuous. Considering speeches, we can easily note that the time series of the FedSent Index is significant only for EUR and USD. This is because for GBP, CNY and JPY are rarely mentioned in the analyzed speeches.

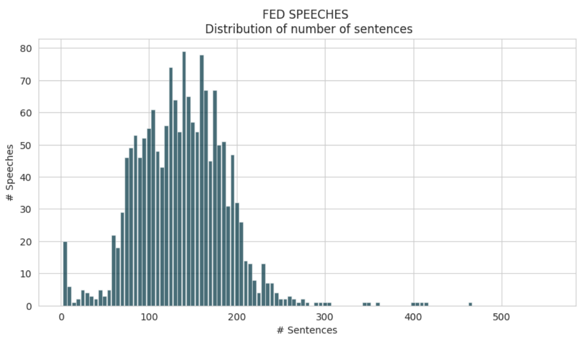


(a) Number of sentences

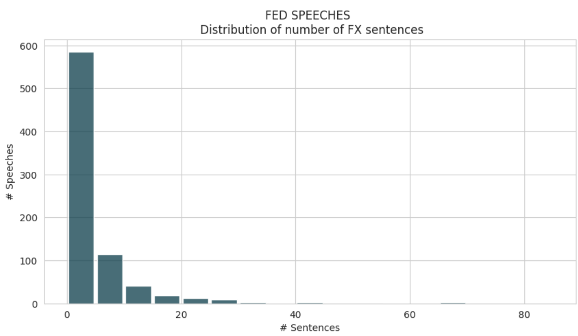


(b) Number of FX sentences

Figure 4: FED minutes distributions.

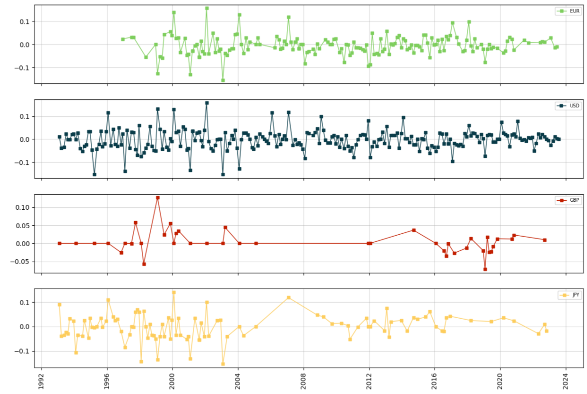


(a) Distribution of sentences

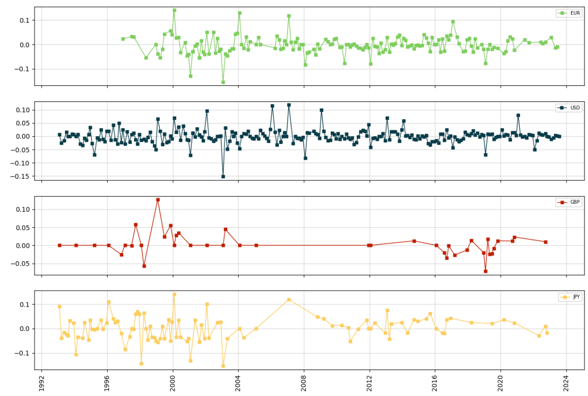


(b) Distribution of sentences mentioning currencies

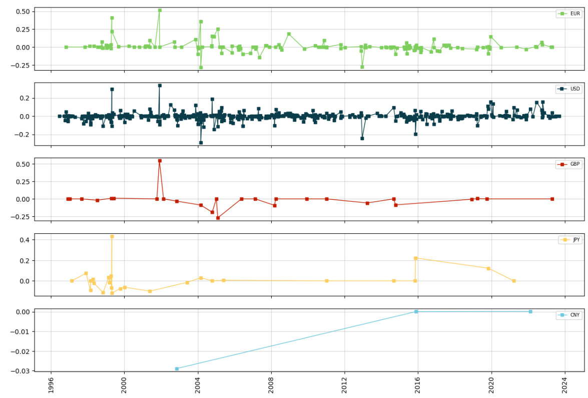
Figure 5: FED speech distributions.



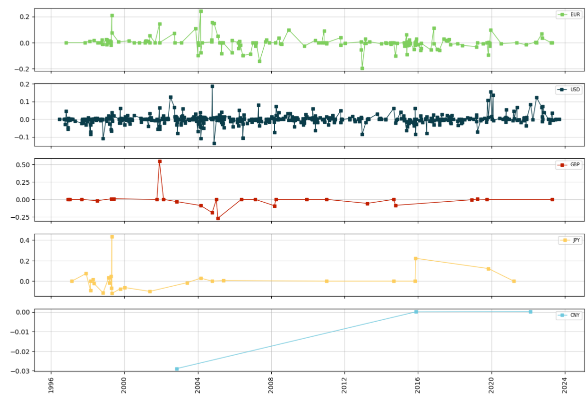
(a) Minutes - Smoothed (median) FedSent Index time series, i.e. $FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$, for major currencies.



(b) Minutes - Smoothed (mean) FedSent Index time series, i.e. $FSI_{\mu}(c,t,d) \times \beta_{ctd}$, for major currencies.



(c) Speeches - Smoothed (median) FedSent Index time series, i.e. $FSI_{\tilde{\mu}}(c,t,d) \times \beta_{ctd}$, for major currencies.



(d) Speeches - Smoothed (mean) FedSent Index time series, i.e. $FSI_{\mu}(c,t,d) \times \beta_{ctd}$, for major currencies.

Figure 6: Smoothed FedSent Index.

8. Bibliographical References

Jonathan Benchimol, Sophia Kazinnik, and Yossi Saadon. 2021. Federal reserve communication and the covid-19 pandemic. *Covid Economics*, 79:218–256.

Eric Fischer, Rebecca McCaughrin, Saketh Prazad, and Mark Vandergon. 2023. Fed transparency and policy expectation errors: A text analysis approach. *FRB of New York Staff Report*, (1081).

Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438.

Gunho Jung and Sun-Yong Choi. 2021. Forecasting foreign exchange volatility using deep learning autoencoder-lstm techniques. *Complexity*, 2021:1–16.

Jean Lee, Hoyoul Luis Youn, Nicholas Stevens, Josiah Poon, and Soyeon Caren Han. 2021. Fednlp: an interpretable nlp system to decode federal reserve communications. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2560–2564.

Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. 2023. FinGPT: Democratizing internet-scale data for financial large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: a pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130.

Saeed Seifollahi and Mehdi Shajari. 2019. Word sense disambiguation application in sentiment analysis of news headlines: an applied approach to forex market prediction. *Journal of Intelligent Information Systems*, 52:57–83.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

Ali Shojaie and Emily B Fox. 2022. Granger causality: A review and recent advances. *Annual Review of Statistics and Its Application*, 9:289–319.

Raul Cruz Tadle. 2022. Fomc minutes sentiments and their impact on financial markets. *Journal of Economics and Business*, 118:106021.

Arina Wischnewsky, David-Jan Jansen, and Matthias Neuenkirch. 2021. Financial stability and the fed: evidence from congressional hearings. *Economic Inquiry*, 59(3):1192–1214.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A comprehensive benchmark, instruction dataset and large language model for finance. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning.

Analysis of Material Facts on Financial Assets: a Generative AI Approach

Gabriel Assis[1], Daniela Vianna[4], Gisele L. Pappa[3], Alexandre Plastino[1], Wagner Meira Júnior[3], Altigran Soares da Silva[2], Aline Paes[1]
[1] Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil
[2] Institute of Computing, Universidade Federal do Amazonas, Manaus, AM, Brazil
[3] Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[4] JusBrasil, Brazil
assisgabriel@id.uff.br, daniela.vianna@jusbrasil.com.br, glpappa@dcc.ufmg.br, plastino@ic.uff.br, alti@icomp.ufam.edu.br, meira@dcc.ufmg.br, alinepaes@ic.uff.br

Abstract

Material facts (MF) are crucial and obligatory disclosures that can significantly influence asset values. Following their release, financial analysts embark on the meticulous and highly specialized task of crafting analyses to shed light on their impact on company assets, a challenge elevated by the daily amount of MFs released. Generative AI, with its demonstrated power of crafting coherent text, emerges as a promising solution to this task. However, while these analyses must incorporate the MF, they must also transcend it, enhancing it with vital background information, valuable and grounded recommendations, prospects, potential risks, and their underlying reasoning. In this paper, we approach this task as an instance of controllable text generation, aiming to ensure adherence to the MF and other pivotal attributes as control elements. We first explore language models' capacity to manage this task by embedding those elements into prompts and engaging popular chatbots. A bilingual proof of concept underscores both the potential and the challenges of applying generative AI techniques to this task.

Keywords: financial asset narratives, material fact, text generation, natural language instructions

1. Introduction

Material Facts (MF) are critical disclosures issued by companies and investment funds (Pallister and Law, 2006). They hold paramount importance in the market due to their potential to affect asset values. Owing to their substantial impact on the financial sector, these announcements are mandatory by law to maintain market transparency (Easterbrook and Fischel, 1984; Mahoney, 1995; Boatright, 2013). Their scope varies from alterations in a company's shareholder composition to significant acquisitions and disposals. Typical examples include executing agreements for transferring share control, changes in partnership, renegotiating debts, managing stock splits or mergers, reporting profits or losses, distributing dividends, adjusting financial forecasts, or initiating bankruptcy proceedings (Finnerty, 2013).

The mandatory announcement of material facts implicates a vast amount of documents. To have an idea, in the last trimester of 2023, the US Securities and Exchange Commission (SEC) had registered more than 23,000 material facts related to worldwide companies[1]. After the announcements, financial analysts and other industry experts might produce an analysis relating the MF to financial assets. Arguably, the analysis is not only an expansion of the MF as they also might implicitly incorporate analysts' background and other relevant events that are not explicitly related to the announcement (Subramanyam, 2014; Revsine et al., 2021).

This way, writing financial asset analyses from an MF presents various challenges. Material facts vary widely – some are brief, while others are lengthy and detailed; although all facts must be disclosed, their importance and impact can differ (Reichert and Perlin, 2022). Arguably, crafting accurate and valuable asset analyses from an MF requires additional information from other sources to point out how the MF might influence an asset, considering the market position, prospects, and potential risks (Chauvin and Hirschey, 1993). Additionally, the analysts' perspectives can bias the final recommendations, potentially compromising the accuracy of the information provided to the public (Hawes, 1975; Boatright, 2013). Furthermore, investment firms typically charge for access to their reports. While they are justified in doing so, this practice can limit the accessibility of financial information for those unable to afford their fees.

Conversely, with the advent and growing popularity of Generative AI powered by large language models (LLMs), automatic text generation (Erdem et al., 2022) has achieved remarkable results. However, generating a text that complies with another piece of information – as in financial asset analyses – still challenges modern LLMs (Peng et al., 2023; Zhang et al., 2024a). While the generated text must convey the material fact, it must, in essence,

[1] `https://www.sec.gov/edgar/search/`

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 103–118
20 May, 2024.

discuss its implications, bringing related information, including extended analyses and correlating important related events. Arguably, while LLMs are pre-trained with a massive amount of text and modern techniques offer the possibility of expanding them with Retrieval Augmented Generation (Lewis et al., 2020b) and reasoning abilities (Wei et al., 2022), they might still not be fully equipped to deal with the listed challenges.

This paper formalizes this task as an instance of controllable text generation (CTG) (Prabhumoye et al., 2020; Zhou et al., 2023; Zhang et al., 2024b). While previous works have explored several forms of CTG with LLMs (Prabhumoye et al., 2020; Zhang et al., 2024b; Zhou et al., 2023), most of the content-based category relies on simple elements, such as terms and topics, or self-contained texts (Liu et al., 2018; Yan, 2022). They have not examined such a diverse and, at the same time, untied element of control as a material fact.

Our initial strategy is implementing this task with crafted structured prompts embedded with a material fact and leveraging the in-context learning emerging abilities of state-of-the-art chabtots (Brown et al., 2020b; Dong et al., 2023) to write financial asset analyses. This preliminary investigation focuses on assessing the abilities of state-of-the-art LLMs to produce analyses solely based on their prior training stages.

Our proof of concept adopts a bilingual approach, generating analyses in both English and Portuguese, thereby probing the capabilities of LLMs in linguistic contexts beyond their primary training. To assess the efficacy of our method, we developed a proof of concept with a small set of reference financial analyses, sourced from reputed investment analyses. Then, we compare them against the outputs of various chatbots, encompassing a spectrum from open to closed systems and from small to large-scale models. We include two models of the GPT family (Ouyang et al., 2022; OpenAI, 2023), three Mistral models (Jiang et al., 2023, 2024), three Llama models (Touvron et al., 2023) and the recently released Gemini-Pro (Google, 2023).

While LLMs often blur the line between human and machine-generated texts, accurately assessing their quality remains challenging, especially in sensitive and specialized fields like finance. This way, this paper relies on traditional and modern text generation metrics to evaluate the output of chatbots in contrast to reference reports. This comparative analysis aims to assess the challenges and potential of the proposed task and at which point LLMs can tackle it.

To sum up, this paper contributes with

1. A novel task proposal for automatically generating asset analysis from material facts.
2. The task implementation leveraging generative AI, guided by well-crafted instruction incorporating a material fact.
3. A proof of concept with LLMs encompassing analyses in Portuguese and English.
4. We explore this problem with nine chatbots and evaluate the results with classical and state-of-the-art text generation metrics.

2. Related Work

2.1. Automating Financial Narratives

Prior research has investigated methods for generating financial reports from different inputs. One of the key areas of focus is generating reports from tabular data using table-to-text (TTG) techniques (Kale and Rastogi, 2020). Cm et al. (2023), for instance, explored pre-trained TTG transformers and how selecting crucial, correlated table attributes can aid in producing accurate money laundering reports. While material facts are typically presented in textual format, they may include numerical values both within tables and the text itself. Therefore, incorporating logical steps to integrate these values into reports accurately will be crucial in future efforts.

More aligned with our method, Yan (2022) developed a technique for creating financial reports from brief news articles. They focused on learning separate latent variables that capture the themes of the input news and the intended reports. The aim is to incorporate the natural uncertainty in reports, acknowledging that human experts contribute diverse perspectives and approaches to their analysis. Material facts may also consist of brief texts, but sometimes they have detailed information. Our proposed task targets producing reports that capture the source document's essential elements while enriching it with additional discussion and inferred insights. We begin our exploration of this issue by leveraging chatbots' innate capability to address these complexities.

2.2. Controllable Text Generation from Content

Frequently, text generation requires that the output agrees with a predefined specific element, such as style, structured data, or content (Erdem et al., 2022). Most recent work that generates controlled texts conditions the input to the required attribute (Prabhumoye et al., 2020). Zhang et al. (2024b) divides the strategies into the three following: (a.) adjusting (some of) pre-trained language models weights to produce texts with specific features (Ziegler et al., 2019; Liu et al., 2020), (b.) training controllable models with injecting controllers (Wang et al., 2021; He, 2021; Chan et al., 2021), and (c.) post-processing PTLM signals that work only when decoding texts to incorporate the desired attributes (Hua and Wang, 2020; Dathathri et al., 2020). In our case, the report must be conditioned on the material fact while conveying related information. Our strategy aligns with the third aforementioned approach, as we include the material fact text in the prompt body and analyze whether chatbots attain it without further control.

3. Task Formulation

This paper introduces a novel task of generating analyses on financial assets (e.g., stocks, funds, private pensions, etc.) from material facts, leveraging text generation techniques. We formulate the task as an instance of controllable text generation, named *material fact controllable text generation* (MF-CTG). This way, MF-CTG is defined as $P(Y|MF, C, \mathbf{A})$, where the controllable elements are the material fact source text $MF = x_1 \ldots x_m$, the company's name releasing it (C), and a set of other controllable attributes($\mathbf{A} = \{A_1, A_2, \ldots, A_k\}$), which could be style, impartial tone, structure, among others. The goal of the task is to generate a financial asset analysis report Y according to a vocabulary $\mathcal{V}$ where $Y = y_1 \ldots y_m$ and $y_k \in \mathcal{V}$.

This paper addresses this task using prompt-based generative pre-trained language models (PTLM). Therefore, $\mathcal{V}$ is the PTLM vocabulary, and MF, C, and $\mathbf{A}$ are included into a prompt PM, together with other elements, for example, context and instruction. This way, the task is $P(Y|PM) = \prod_{i=1}^{n} p(y_i|y_{<i}, PM)$, where $PM = \mathbf{z_1} \ldots \mathbf{A} \ldots \mathbf{z_l} \ldots C\ \mathbf{z_j} \ldots MF\ \mathbf{z_w}$ with $\mathbf{z_i}$ being (possibly empty) sequences of words representing other information added to the prompt.

4. Instance and Evaluation of Material Fact Controllable Text Generation

This section describes our proposed method to address a concrete instance of MF-CTG and an evaluation routine. The method consists of the following procedures: (A.) Prompt Crafting, that assembles a prompt incorporating the related material fact, company's name, and analysis format as the controllable elements, among additional text;

(B.) Analyses Generation, rooted in activating the generative AI models through the prompt; and (C.) Analyses Evaluation, to evaluate the output analyses with automatic metrics. Those metrics assess analyses' lexical and syntactic aspects and rate semantic conformity regarding a reference report. While (A.) and (B.) implement MF-CTG, (C.) is responsible for evaluating its feasibility. Figure 1 depicts an overview of the proposed method.

4.1. Prompt Crafting

How a prompt is constructed significantly impacts the effectiveness of an LLM in performing downstream tasks (Liu et al., 2023a). In this context, prompt engineering techniques potentially optimize a model's performance. Those strategies encompass adding personas, using different delimiter symbols, incorporating reference materials, integrating examples for in-context learning, and outlining steps for task execution, often referred to as chain of thought (White et al., 2023; Brown et al., 2020a; Wei et al., 2022). While embracing those prompt engineering techniques, our strategy also entails directly integrating the controllable elements (material fact, its originating company's name and the format) into the prompt.

After conducting a series of preliminary assessments using these techniques, we observed that incorporating a persona and adding detailed instructions on the document's intended audience and format, along with an explicit formatting template, led to more appropriate responses. Specifically, the models produced content that was not only correctly formatted but also returned an analysis that considered the provided material fact and showed adherence to the vocabulary. Nonetheless, it is still crucial to thoroughly evaluate the content quality and the overall analysis. Further details on this are discussed in Section 4.3. The final prompt is as follows[2]:

Context:
You are a financial analyst with a background in economics who writes for a general investor audience.

Instruction:
Write an analysis about the Company considering the Material Fact and follow the determined Format. The analysis must contain ALL the elements specified in the following Format.

Format:
<Title>
<Body of the Analysis>
<Recomendation>

Company:
{company}

Material Fact:
{source_document}

###Response:

4.2. Analyses Generation

Generating analyses from material facts with LLMs involves several factors. Those include the model's ability to process the given material fact and derive contextually pertinent information to compose a robust and relevant analysis. Therefore, assessing

[2] See Appendix B for the Portuguese prompt.

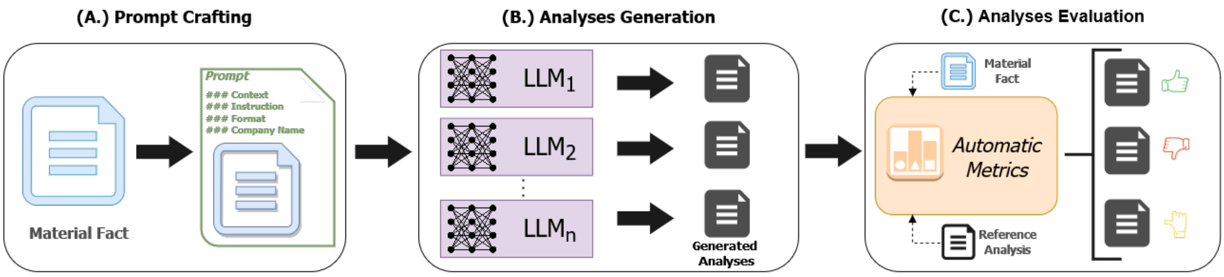


Figure 1: Schematic pipeline of the proposed MF-CTG solution.

the effectiveness of different models in carrying out the proposed MF-CTG task is critical to determine its feasibility.

The evaluation encompassed models based on the popular GPT architecture, specifically (i.) GPT-4 (OpenAI, 2023) and (ii.) GPT-3.5 (Ouyang et al., 2022). Additionally, the (iii.) Gemini-Pro 1.0 model (Google, 2023), known for its remarkable performance in recent benchmarks, and models from the Mistral family — (iv.) Mistral Medium, (v.) Mixtral 8x7B (Jiang et al., 2024), and (vi.) Mistral 7B (Jiang et al., 2023) — were included. We also adopted models from the Llama 2 series (Touvron et al., 2023), the (vii.) Llama 2 7B, (viii.) Llama 2 13B, and (ix.) Llama 2 70B. All models were utilized in their chat or instruction-based versions.

The selected models vary significantly in size, with their number of parameters spanning from a few billion to over a trillion. They also differ in terms of availability, including both closed (models i. to iv.) and open (subsequent models) sources. We aimed to assess how these variations affect the MF-CTG in both English and Portuguese. In this context, another aspect under investigation is the capability to generate analyses in a language for which all models were not predominantly pre-trained.

4.3. Analyses Evaluation

Evaluating texts produced by Generative AI methods remains a significant challenge to NLP (Kryscinski et al., 2019; Liu et al., 2023b). It is even more critical when considering texts from specific areas, such as finances. Nevertheless, characteristics addressed in NLG tasks, like summarization, transduction, and creative writing, establish a robust foundation for automated evaluation. Some critical criteria for evaluation encompass lexico-syntactic structure; fluency, which evaluates the clarity of the writing (Zhong et al., 2022); consistency, which checks how well the generated text aligns with the source document (Cao et al., 2020); coherence, assessing the logical connection of information (Ye et al., 2021); relevance, measuring the pertinence of the generated text to the critical information based on the reference (Zhong et al., 2022); and groundedness, evaluating the text's ability to reflect reasoned knowledge from the source document (Dinan et al., 2019). These criteria are crucial for the MF-CTG task, as they affect the report's readability and can influence its credibility.

A widely adopted approach relies on metrics that measure the similarity of generated texts with references written by humans (Deng et al., 2021). In this context, our work incorporates the material fact and analyses gathered from financial firms' websites, serving as reference reports.

We apply three groups of metrics. The first group regards semantic aspects with the following metrics: BERTScore (Zhang et al., 2020), which computes similarity based on contextualized embeddings; and BARTScore (Yuan et al., 2021), which proposes a unified evaluator based on the likelihood of the encoder-decoder model upon which it is based, and changes in the combination of its inputs. In our evaluation strategy, we propose segmenting both the evaluated text and the reference text. Thus, for an analysis report denoted as a, we define $\{a_t, a_o, a_b, a_c\}$ representing its title, overview, body, and conclusion, respectively. Similarly, for a reference denoted as r, we have $\{r_t, r_o, r_b, r_c\}$. We then calculate the scores for each pair $score(a_t, r_t)$, $score(a_o, r_o)$, $score(a_b, r_b)$, and $score(a_c, r_c)$, and subsequently calculate the average of the previous results. We conjecture that this method allows a more refined evaluation between the components of the analysis report, while also addressing the limitation of the context window often found in using these metrics with lengthy documents.

Conversely, UniEval (Zhong et al., 2022) was employed to evaluate fluency and coherence. This metric utilizes a binary question-answering (QA) pipeline built upon a generative LLM to calculate its value leveraging the probabilities of responses to questions like *"Is this text fluent?"*. Given its evaluation method and the aspects this metric covers are inherent to the text as a single piece, the previous segmentation approach was not applied.

The second group includes morphological and parsing analyses, including tokenization, part-of-speech tagging, and dependency parsing-related metrics conveyed in UDPipe (Straka et al., 2016). It relies on the Universal Dependencies treebank annotations that include analyzers for both English and Portuguese. We compute the number of sentences, tokens, tokens per sentence, and the mean dependency distance (MDD). This last one aims to predict the syntactic difficulty of sentences according to psycholinguistics experiments (Liu, 2008).

Finally, the third group focuses only on the Portuguese analyses as we leverage the large set of metrics provided in NILC-metrix portal (Leal et al., 2023)[3]. Those metrics extract values from several linguistic proxies to assess morphosyntactic, cohesion, coherence, and textual complexity information. Once again, we compare the automatically generated analysis reports with the reference. We compute the metrics for the following groups: Referential Cohesion (seven metrics), Syntactic Complexity (27 metrics), Morphosyntactic Information (42 metrics), and Readability (five metrics).

5. Experimental Setup

This section describes the process of gathering material facts and references, the experimental settings employed in the inference process with LLMs, and further details on implementing the metrics.

5.1. Data

Collection Methodology Our evaluation includes documents in both English and Portuguese. To find openly accessible analyses in English, we benefit from Yahoo Finance, a popular tool in the financial context for indexing news and public reports[4]. We adopted keywords associated with typical topics in material facts as filters, e.g., reports of changes in partnership or organizational restructuring. More examples can be found in Section 1. Subsequently, the material facts issued on the same day or the day before by the entities mentioned in the reports were reviewed through the public system of the SEC[5]. The objective is to match the reports with the forms that contain this specific type of information, namely Forms 6-K for foreign companies, and Forms 8-K for US-based companies.

The reports in Portuguese were directly collected from the websites of financial analysis firms. The same keyword strategy was adopted. Moreover, the system of the regulatory agency equivalent to the SEC in Brazil, the *Comissão de Valores Mobiliários (CVM)*[6], was used for the collection of the material facts, in Portuguese, "*Fatos Relevantes*" forms.

Companies and Material Fact Selection We selected two reference analyses for each language. For Portuguese, the companies examined are BTG Pactual, a Brazilian investment bank specializing in investments and venture capital, and Eneva, a comprehensive Brazilian energy company engaged in power generation, oil and gas exploration and production, and electricity trading. For BTG Pactual, the critical event highlighted in the material fact was the acquisition of three properties, representing a multimillion-BRL transaction. In the case of Eneva, the significant event was a report on the rejection of a previously attempted merger by the company.

The cases in English concern Petrobras, Brazil's largest oil company, a publicly traded corporation operating in the oil, natural gas, and energy sectors. The other company is Twilio, which offers communication tools and services through service APIs. In the case of Petrobras, the pertinent fact was the announcement of the intention to acquire the Jasper Block in the Campos Basin. Meanwhile, Twilio announced a layoff in its global workforce and the integration of Twilio Flex.

This selection requires that the models show a broad range of abilities and knowledge. They must not only consider the events and companies involved but also reason about factors such as geographical nuances, and the sizes and sectors of the companies. The MFs are presented entirely in the Appendix C.

5.2. Generative Models Inference

We conduct inference on generative models through ChatBot Arena[7] (Zheng et al., 2023), an LLM benchmark platform that features comparisons between models in a crowdsourced manner. The platform provides access to models such as *gpt-4-1106, gpt-3.5-turbo-0613, gemini-pro-dev-api, mistral-medium, mixtral-8x7b-instruct-v0.1, mistral-7b-instruct, llama-2-7b-chat, llama-2-13b-chat,* and *llama-2-70b-chat*, all of which were adopted in this work. Our choice is based on the high computational costs of running huge models. It also fits the objective of conducting a preliminary evaluation of the models' ability to generate financial analyses.

As well known, the choice of hyperparameters such as $temperature$ and top_p significantly influences the responses generated by models (Döderlein et al., 2022). To balance between aspects like creative writing and truthfulness, these parameters were set at $\{temperature = 0.3, top\_p = 0.4\}$. The

[3]http://fw.nilc.icmc.usp.br:23380/nilcmetrix
[4]https://finance.yahoo.com/
[5]https://www.sec.gov/edgar/
[6]https://sistemas.cvm.gov.br/
[7]https://arena.lmsys.org/

max_tokens parameter was set to 2000, the highest possible value in Arena, to not limit the models in terms of the length of the generated analyses.

5.3. Metrics Implementation

Concerning the implementation of automatic semantic metrics, we leveraged models that we conjecture to have a more suited semantic representation for metrics that measure similarity, given the financial context, whether through the vector representation of contextualized embeddings or the likelihood of tokens. Consequently, we employed the BERTScore metric on FinBERT (Araci, 2019), a model pre-trained on a financial *corpus*. However, the vocabulary of this model is limited to English. Thus, for analyses in Portuguese, we used the multilingual version of BERT (Devlin et al., 2019)[8], given that the metric's implementation does not support a specific model for Portuguese. For BARTScore, which was originally proposed based on the encoder-decoder model BART (Lewis et al., 2020a), the implementation for English analysis was carried out on its version tuned on CNN and DailyMail news *corpus*. This choice was made due to the higher correlation with human evaluation reported by the original work (Yuan et al., 2021). For Portuguese, similarly to the approach taken for the previous metric, the multilingual version of the original model was employed[9].

Additionally, the UniEval metric is originally proposed on a fine-tuned version of the T5 model (Raffel et al., 2020). Considering the limitation that may be associated with the number of tokens in lengthy financial analyses and to prevent information loss due to truncation, we implemented the original formulation proposed by (Zhong et al., 2022) within the probability results obtained by querying GPT-4 via its API. Further details in the Appendix D. Finally, the implementations of the other metrics and the other hyperparameters settings for all metrics were kept at their default values.

6. Proof of Concept Results

6.1. Quantitative Analysis

Table 1 showcases the results of the first group of metrics, assessing semantic aspects compared to reference reports. UniEval could not discriminate among the LLMs, as the probabilities returned by GPT-4 are always very close to 1. Conversely, BERTScore and BARTScore failed to conclusively identify a superior model, as scores were generally close and varied models excelled in different scenarios. However, two key insights emerged from the results: (i.) the values point out that the reports generated by LLMs closely resemble the reference collection, underscoring the potential of our proposed task, and (ii.) surprisingly, the metric scores for the Portuguese results are generally higher than those for English. This is notable considering the models applied for Portuguese are multilingual and were not explicitly optimized for the financial sector, unlike FinBERT.

Table 2 displays token and sentence count, average tokens per sentence, and the average dependency distance, together with the difference regarding the reference, for both Portuguese and English scenarios. In most instances, the number of tokens and sentences was lower than the reference. While this could be interpreted as greater conciseness, some significant differences indicate the possibility that the models might not have generated additional and relevant information beyond the content of the material fact. Conversely, the values for MDD are significantly close to the reference, which reinforces the notable ability of LLMs in generating texts with grammatical complexity similar to human-written texts (Liu et al., 2023b).

Finally, Table 3 brings the Euclidean distance between the reference and generated reports using the average of four groups of metrics from NILC-metrix, only for the Portuguese cases. Before computing the distance, we normalized the values between 0 and 1. Overall, the models performed closely to the reference, achieving small distances for most metric groups. The poorest performance was observed in referential cohesion for Eneva, where the distance reached half of its maximum potential value. Examining morphosyntax reveals an interesting case with BTG, in which the entire Llama series achieved below-average results, including the worst overall performance. This is evident from the analyses, which include examples such as *"um área"*, a gender mismatch in Portuguese. Consequently, writing in a language for which the models were not primarily pre-trained may pose a significant hurdle. The Mistral and GPT families each claimed half of the top spots. However, while a Mistral model was among the worst performers alongside Llama and Gemini, no GPT model fell into this category. Nonetheless, the open models show promise, delivering solid performances and allowing for further exploration due to their greater accessibility than GPT.

These metrics also reveal that size does not necessarily equate to consistent behavior. For instance, the 7B versions of Mistral and Llama exhibit several distant values in the BTG reports. Moreover, size is not always a determinant of performance, as pointed out by the results within the Llama family:

[8] https://huggingface.co/google-bert/bert-base-multilingual-cased

[9] https://huggingface.co/facebook/bart-large-cnn

	Portuguese						English					
	BERTScore		BARTScore		UniEval		BERTScore		BARTScore		UniEval	
	BTG	Ene.	BTG	Ene.	BTG	Ene.	Petro.	Twillio	Petro.	Twillio	Petro.	Twillio
GPT-4-turbo	**0.715**	0.681	-7.295	-5.973	**1.000**	0.999	0.627	0.599	-3.188	-3.386	0.998	0.999
GPT-3.5-turbo	0.531	0.504	-6.524	-6.793	**1.000**	0.998	0.591	0.593	-3.633	-3.258	**1.000**	**1.000**
Gemini Pro	0.698	0.668	-7.675	-8.356	**1.000**	0.999	0.579	0.596	-3.556	-3.322	**1.000**	**1.000**
Mistral Medium	0.710	**0.682**	-6.474	-6.062	**1.000**	**1.000**	0.625	0.579	-3.164	-3.391	0.999	**1.000**
Mixtral 8x7B	0.710	0.667	-6.082	-6.546	**1.000**	**1.000**	0.627	0.610	**-3.138**	-3.257	**1.000**	**1.000**
Mistral 7B	0.710	0.672	**-5.213**	-6.023	0.998	0.997	**0.635**	0.599	-3.147	-3.252	**1.000**	**1.000**
Llama2 7B	0.697	0.675	-6.005	-6.118	**1.000**	**1.000**	0.575	0.599	-3.690	-3.300	**1.000**	**1.000**
Llama2 13B	0.696	**0.682**	-5.768	-5.803	0.999	0.999	0.589	**0.615**	-3.747	-3.294	**1.000**	**1.000**
Llama2 70B	**0.715**	0.667	-7.326	**-5.566**	0.993	0.990	0.581	0.600	-3.683	**-3.247**	**1.000**	**1.000**

Table 1: Comparative results of semantic evaluation metrics for Portuguese and English cases. It presents the recall values for BERTScore and BARTScore, alongside the UniEval average scores. All metrics except BARTScore are scored in the range of 0 to 1, with higher scores indicating better quality. Best-performing results are highlighted in **bold** for each case.

	Portuguese								English							
	#sent. *(reference △)*		#tokens *(reference △)*		#tokens/sent. *(reference △)*		MDD *(reference △)*		#sent. *(reference △)*		#tokens *(reference △)*		#tokens/sent. *(reference △)*		MDD *(reference △)*	
	BTG	Ene.	BTG	Ene.	BTG	Ene.	BTG	Ene.	Petro.	Twillio	Petro.	Twillio	Petro.	Twillio	Petro.	Twillio
GPT-4 turbo	18 *(-22)*	26 *(-2)*	356 *(-254)*	403 *(-47)*	19.778 *(+4.528)*	15.500 *(-0.571)*	3.827 *(+0.280)*	3.444 *(-0.150)*	26 *(-38)*	36 *(-4)*	437 *(-279)*	465 *(+7)*	16.808 *(+5.620)*	12.917 *(+1.467)*	4.132 *(+0.527)*	3.808 *(+0.161)*
GPT-3.5 turbo	24 *(-16)*	24 *(-4)*	341 *(-269)*	329 *(-121)*	14.208 *(-1.042)*	13.708 *(-2.363)*	3.796 *(+0.249)*	3.621 *(+0.027)*	30 *(-34)*	36 *(-4)*	438 *(-278)*	427 *(-31)*	14.600 *(+3.413)*	11.861 *(+5.017)*	3.827 *(+0.222)*	3.753 *(+0.097)*
Gemini Pro	18 *(-22)*	22 *(-6)*	235 *(-375)*	267 *(-183)*	13.056 *(-2.194)*	12.136 *(-3.935)*	3.531 *(-0.016)*	3.180 *(-0.414)*	22 *(-42)*	22 *(-18)*	343 *(-373)*	314 *(-144)*	15.591 *(+4.403)*	14.273 *(+2.823)*	3.892 *(+0.286)*	3.812 *(+0.166)*
Mistral Medium	20 *(-20)*	22 *(-6)*	320 *(-290)*	347 *(-103)*	16.000 *(+0.750)*	15.773 *(-0.299)*	3.410 *(-0.138)*	3.521 *(-0.073)*	34 *(-30)*	34 *(-6)*	445 *(-271)*	438 *(-20)*	13.088 *(+1.901)*	12.882 *(+1.432)*	3.703 *(+0.098)*	3.786 *(+0.140)*
Mixtral 8x7B	24 *(-16)*	26 *(-2)*	399 *(-211)*	274 *(-176)*	16.625 *(+1.375)*	10.538 *(-5.533)*	3.519 *(-0.028)*	3.383 *(-0.211)*	20 *(-44)*	30 *(-10)*	368 *(-348)*	384 *(-74)*	18.400 *(+7.213)*	12.800 *(+1.350)*	4.123 *(+0.518)*	3.938 *(+0.292)*
Mistral 7B	40 *(0)*	40 *(+12)*	615 *(+5)*	485 *(+35)*	15.375 *(+0.125)*	12.125 *(-3.946)*	3.763 *(+0.216)*	3.138 *(-0.456)*	32 *(-32)*	28 *(-12)*	505 *(-211)*	407 *(-51)*	15.781 *(+4.594)*	14.536 *(+3.086)*	4.076 *(+0.470)*	3.954 *(+0.308)*
Llama2 7B	28 *(-12)*	30 *(+2)*	391 *(-219)*	327 *(-123)*	13.964 *(-1.286)*	10.900 *(-5.171)*	3.422 *(-0.126)*	3.266 *(-0.328)*	20 *(-44)*	30 *(-10)*	379 *(-337)*	494 *(+36)*	18.950 *(+7.763)*	16.467 *(+5.017)*	3.821 *(+0.216)*	3.743 *(+0.097)*
Llama2 13B	26 *(-14)*	26 *(-2)*	377 *(-233)*	339 *(-111)*	14.500 *(-0.750)*	13.038 *(-3.033)*	3.495 *(-0.053)*	3.390 *(-0.204)*	32 *(-32)*	44 *(+4)*	469 *(-247)*	640 *(+182)*	14.656 *(+3.704)*	14.545 *(+3.095)*	3.704 *(+0.099)*	3.649 *(+0.003)*
Llama2 70B	32 *(-8)*	46 *(+18)*	418 *(-192)*	497 *(+47)*	13.063 *(-2.188)*	10.804 *(-5.267)*	3.468 *(-0.080)*	3.403 *(-0.191)*	36 *(-28)*	28 *(-12)*	445 *(-271)*	423 *(-35)*	12.361 *(+1.174)*	15.107 *(+3.657)*	3.663 *(+0.058)*	3.902 *(+0.256)*

Table 2: Comparative results of Morphological and parsing metrics computed with UDPipe.

the larger model does not necessarily outperform its smaller counterparts.

6.2. Manual Inspection of an Analysis

We selected the EN Petrobras analysis to take a fine-grained look at how a particular analysis addressed the material fact. One representative of each family was selected according to their performance on similarity and morphosyntactic results (Tables 1 and 2). Although Gemini did not achieve the best results, we also bring it here for comparison. The Appendix E discusses an AI-generated analysis and the reference. Regarding the format, all the analyses include a title, body, and recommendation followed or not by additional conclusions or (an attempt of) reasoning strategy.

Table 4 exhibits the titles of the material fact, the reference, and the four selected analyses. We notice that all models extensively add words and phrases from the MF title, whereas the reference title summarizes the main point more concisely. All the models include the company's name and the expression "*Production Sharing regime*" directly from the MF title. The title with more words in common with the MF is Gemini, indicating a lower level of creativity in elaborating beyond merely reflecting the title. Llama is almost the same, but at least it included an expression to suggest further discussion (*A Promising Move?*). Conversely, Mistral 7B included the block name (*Jaspe Block*) and GPT-4-Turbo also included the block location (*Campos Basin*), both of them mentioned in the MF body.

In terms of the main body of the analyses, while all the generated analyses address the primary subject of the material fact, they tend to be quite superficial and merely outline the anticipated outcomes of the acquisition. Moreover, they missed listing positive outcomes. For instance, none mentioned the potential for job creation or the advancement of technology when exploring the block, as the reference did.

While the reference report describes the block as "*a geological treasure trove responsible for roughly 80% of Brazil's oil output....*" the machine-generated analysis lack more information about it. They only superficially mention the Jaspe Block as promising and the Campos Basin as prolific. The way we activate the LLMs does not provide them with direct access to such detailed information, even though it might be within their pre-training data. This gap suggests that the decoded analysis could benefit from external sources of information.

	Referential Cohesion		Syntactic Complexity		Morphosyntactic		Readability	
	BTG	Ene.	BTG	Ene.	BTG	Ene.	BTG	Ene.
GPT-4-turbo	0.3063	1.2193	**0.0456**	0.1103	**0.2337**	**0.2474**	0.0244	0.0216
GPT-3.5-turbo	**0.1591**	1.2652	0.0600	0.1044	0.3089	0.2930	0.0112	0.0138
Gemini Pro	0.2970	1.2110	0.0678	<u>0.1561</u>	0.2952	<u>0.4814</u>	<u>0.0465</u>	0.0373
Mistral Medium	0.1784	**1.0560**	0.0521	0.0869	0.3149	0.3437	**0.0053**	**0.0011**
Mixtral 8x7B	0.3716	1.2967	0.0758	0.1087	0.3022	0.3310	0.0263	0.0299
Mistral 7B	<u>0.6112</u>	1.2686	0.0577	**0.0841**	0.3386	0.3918	0.0244	<u>0.0381</u>
Llama2 7B	0.2789	<u>1.3034</u>	<u>0.0956</u>	0.1213	<u>0.6108</u>	0.3538	0.0061	0.0274
Llama2 13B	0.1904	1.2742	0.0683	0.0905	0.4670	0.3002	0.0293	0.0105
Llama2 70B	0.5066	1.1143	0.0624	0.1101	0.4207	0.4329	0.0204	0.0158
Mean ***(std)***	0.3222 (0.1530)	1.2232 (0.0853)	0.0650 (0.0146)	0.1080 (0.0219)	0.3658 (0.1153)	0.3528 (0.0729)	0.0215 (0.0129)	0.0217 (0.0125)

Table 3: Euclidean Distance between NILC-metrix results for BTG and Eneva automatically generated analyses and the reference analysis. The values in **bold** are the less distant to the reference, and the <u>underline</u> values are the more distant.

Source	Title
Material Fact	Petrobras expresses interest in area under the Production Sharing regime
Reference	Petrobras (PBR) Plans to Acquire Jaspe Block in Offshore Brazil
Mistral 7B	Petrobras' Interest in Jaspe Block Under Production Sharing Regime: An Analysis
GPT-4-turbo	Petrobras Eyes Expansion in Campos Basin with Interest in Jaspe Block under Production Sharing Regime
Llama2 70B	Petrobras Expresses Interest in Production Sharing Regime: A Promising Move?
Gemini Pro	Petrobras Expresses Interest in New Production Sharing Regime Block

Table 4: Titles of the Material Fact, reference analysis, and four selected models.

Lastly, some of the generated analyses mention risks and potential negative outcomes, albeit in a broad manner. For example, GPT-4 points out that "*exploration and development of new blocks carry inherent risks, including geological, technological, and market-related uncertainties*". It also raises concerns about the global move towards renewable energy and the need to meet ESG standards. Mistral highlights "*competition from other oil and gas companies in the region and potential regulatory challenges*". Yet, none address the need for the company to take care of and avoid negative environmental impacts, which is a possible outcome that the reference analysis emphasized. Llama and Gemini fail to cover challenges, with Gemini notably adopting a more positive tone, in line with the material fact.

These aspects underscore the challenges of the task we propose in this paper. The analysis should naturally align with the material fact, being compliant, consistent, and coherent. If one targets the reference analyses style, refining the models with examples, explicitly fine-tuning their weights, or enhancing the prompts with demonstrations is recommended.

7. Conclusions

This paper introduced a novel financial task: automatically generating financial asset analyses based on material facts. The task is approached as an instance of controllable text generation, with the material fact (MF) and the company's name serving as primary control elements, alongside other attributes like report structure and tone. We employed generative AI techniques, incorporating these control elements into the prompts. A bilingual proof of concept with four references, nine LLMs, and using semantic, morphological, and syntactic metrics, highlights the proposal's potential and challenges. Among the models we tested, we highly recommend further exploration of Mistral 7B due to its impressive performance in both languages, coupled with the fact that it is openly available and free to use. As expected, GPT-4 also performed remarkably in the Portuguese analyses. Future directions include improving the analyzed information by incorporating relevant facts, gathering more data to make it possible to fine-tune the models, and investigating possible hallucinations. Although we have not discussed that in the paper, we noticed that one of the reports in Portuguese included unreal affirmations. We also plan to design more precise evaluation metrics tailored to the financial sector and achieve more fine-grained control over the generated text by tuning the models with more precise instructions.

Acknwoledgments

This research was financed by the Brazilian research agencies CNPq, CAPES, and FAPERJ.

8. Bibliographical References

Dogu Araci. 2019. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. *CoRR*, abs/1908.10063.

John R Boatright. 2013. *Ethics in finance*. John Wiley & Sons.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual Error Correction for Abstractive Summarization Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Alvin Chan, Yew-Soon Ong, Bill Pung, Aston Zhang, and Jie Fu. 2021. CoCon: A Self-Supervised Approach for Controlled Text Generation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Keith W Chauvin and Mark Hirschey. 1993. Advertising, r&d expenditures and the market value of the firm. *Financial management*, pages 128–140.

Harika Cm, Debasmita Das, Ram Ganesh V, Rajesh Kumar Ranjan, and Siddhartha Asthana. 2023. LoKI: Money Laundering Report Generation via Logical Table-to-Text using Meta Learning. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 104–110, Macao. -.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text generation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, Transduction, and Creation: A Unified Framework for Evaluating Natural Language Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.

Jean-Baptiste Döderlein, Mathieu Acher, Djamel Eddine Khelladi, and Benoît Combemale. 2022. Piloting Copilot and Codex: Hot Temperature, Cold Prompts, or Black Magic? *CoRR*, abs/2210.14699.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey for In-context Learning. *CoRR*, abs/2301.00234.

Frank H Easterbrook and Daniel R Fischel. 1984. Mandatory disclosure and the protection of investors. *Virginia Law Review*, pages 669–715.

Erkut Erdem, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii, Oleksii Turuta, Aykut Erdem, Iacer Calixto, Elena Lloret, Elena Simona Apostol, Ciprian-Octavian Truica, Branislava Sandrih, Sanda Martincic-Ipsic, Gábor Berend, Albert Gatt, and Grazina Korvel. 2022. Neural Natural Language Generation: A Survey on Multilinguality, multimodality, controllability and learning. *J. Artif. Intell. Res.*, 73:1131–1207.

John D Finnerty. 2013. *Project financing: Asset-based financial engineering*, volume 852. John Wiley & Sons.

Team Google. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR*, abs/2312.11805.

Douglas W Hawes. 1975. Truth in Financial Statements: An Introduction. *Vand. L. Rev.*, 28:1.

Xingwei He. 2021. Parallel Refinements for Lexically Constrained Text Generation with BART. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8666, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2020. PAIR: Planning and Iterative Refinement in Pre-trained Transformers for Long Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 781–793, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR*, abs/2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *CoRR*, abs/2401.04088.

Mihir Kale and Abhinav Rastogi. 2020. Text-to-Text Pre-Training for Data-to-Text Tasks. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann, and Sandra Maria Aluísio. 2023. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. *Language Resources and Evaluation*, pages 1–38.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Haitao Liu. 2008. Dependency Distance as a Metric of Language Comprehension Difficulty. *The Journal of Cognitive Science*, 9:159–191.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023a. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9).

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. In *6th International Conference on Learning Representations,* *ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data Boost: Text Data Augmentation Through Reinforcement Learning Guided Conditional Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Paul G Mahoney. 1995. Mandatory disclosure as a solution to agency problems. *The University of Chicago Law Review*, 62(3):1047–1112.

Team OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

John Pallister and Jonathan Law. 2006. A dictionary of business and management.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring Controllable Text Generation Techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Marcos Henrique Reichert and Marcelo Scherer Perlin. 2022. What drives the release of material facts for brazilian stocks? *Brazilian Review of Finance*.

Lawrence Revsine, Daniel W Collins, and W Bruce Johnson. 2021. *Financial reporting & analysis*. McGraw-Hill.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

KR Subramanyam. 2014. *Financial statement analysis*. McGraw-Hill.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Yufei Wang, Ian D. Wood, Stephen Wan, Mark Dras, and Mark Johnson. 2021. Mention Flags (MF): Constraining Transformer-based Text Generators. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event,* *August 1-6, 2021*, pages 103–113. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *CoRR*, abs/2302.11382.

Sixing Yan. 2022. Disentangled Variational Topic Inference for Topic-Accurate Financial Report Generation. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zheng Ye, Liucun Lu, Lishan Huang, Liang Lin, and Xiaodan Liang. 2021. Towards Quantifiable Dialogue Coherence Evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2718–2729, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3):64:1–64:37.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2024b. A Survey of Controllable Text Generation Using Transformer-based Pre-trained Language Models. *ACM Comput. Surv.*, 56(3):64:1–64:37.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a Unified Multi-Dimensional Evaluator for Text Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled Text Generation with Natural Language Instructions. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 42602–42613. PMLR.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul F. Christiano, and Geoffrey Irving. 2019. Fine-Tuning Language Models from Human Preferences. *CoRR*, abs/1909.08593.

A. Ethical Statement

The *corpora* that generative models were pre-trained on may harbor socio-economic partialities, which can lead to these biases being perpetuated in the financial analyses they produce. Additionally, as a task of creative writing, the output can sometimes be marred by the issue of hallucination, leading to the propagation of illusory and occasionally distorted information. We reinforce the recent literature that argues the need for thorough investigation in this aspect. Moreover, financial asset analyses significantly influence investors' decision-making process, from the experienced to the beginners, which can lead to both monetary gains and losses. In this sense, we believe that the AI-generated analyses could play a new and crucial role by providing open and free content on which people can base their investment strategies. However, we note that recommendations from AI should not be accepted uncritically and emphasize the importance of consulting multiple sources of information.

B. Prompt in Portuguese

Contexto:
Você é um analista financeiro com formação em economia que escreve para um público de investidores geral.

Instrução:
Escreva em Português uma análise sobre a Empresa considerando o Fato Relevante e siga o Formato determinado. A análise deve conter TODOS os elementos especificados no seguinte Formato.

Formato:
<Título>
<Sentença que resuma a Análise>
<Corpo da Análise>
<Recomendação>

Empresa:
{company}

Fato Relevante:
{source_document}

###Resposta:

C. The Material Facts

C.1. BTG Pactual

A BTG PACTUAL SERVIÇOS FINANCEIROS S.A. DTVM, inscrita no CNPJ sob o nº 59.281.253/0001-23 ("Administradora"), e a BTG PACTUAL GESTORA DE RECURSOS LTDA., inscrita no CNPJ sob o nº 09.631.542/0001-37 ("Gestora"), na qualidade de administradora e gestora, respectivamente, do BTG PACTUAL LOGÍSTICA FUNDO DE INVESTIMENTO IMOBILIÁRIO inscrito no CNPJ sob o nº 11.839.593/0001-09 ("Fundo"), serve-se do presente para informar aos cotistas e ao mercado em geral quanto o seguem que: I. O Fundo celebrou, nesta data, Instrumentos Particulares ("Instrumento"), sujeitos a determinadas regras de confidencialidade, tendo por objeto a promessa de venda e compra de três imóveis localizados em São Paulo, sendo 70% em um raio de 30km, com área total de, aproximadamente, 233.000m² totalmente locadas ("Imóveis Performados") e um projeto aprovado de, aproximadamente, 74.000 m² (em conjunto com os Imóveis Performados, "Imóveis"), pelo montante total de R$ 760.000.000,00 (setecentos e sessenta milhões de reais) ("Preço da Venda"), a serem desembolsados da seguinte forma: (i) Primeira Parcela: R$ 440.000.000,00 (quatrocentos e quarente milhões de reais) na data de fechamento; e (ii) Segunda parcela: R$ 320.000.000,00 (trezentos e vinte milhões de reais) após 18 meses da data de fechamento, corrigidos pelo Índice Nacional de Preços ao Consumidor Amplo ("IPCA"). II. O fechamento da operação está condicionado a verificação de condições precedentes usuais para este tipo de operação que, quando verificadas, serão comunicadas ao mercado. III. O pagamento parcelado atrelado ao recebimento total das receitas a partir da data de fechamento proporcionará ao Fundo um Yield estimado de 15% até o pagamento da parcela final. A receita estimada desta operação é de R$ 0,19 por cota. IV. O cap rate envolvido na operação, ou seja, o valor de receita vigente sobre o Preço dos Imóveis Performados é de 9,2%. V. Por fim, a Gestora ressalta que a nova aquisição é resultado do trabalho ativo que vem realizando com o intuito de gerar valor para o Fundo e seus cotistas.

C.2. Eneva

Proposta Não-Vinculante para Fusão de Iguais com Vibra Energia S.A. – Resposta do Conselho de Administração da Vibra

Rio de Janeiro, 28 de novembro de 2023 – ENEVA S.A. ("Eneva" ou "Companhia") (B3: ENEV3), em atendimento ao disposto no artigo 157, § 4.º da Lei n.º 6.404, de 15 de dezembro de 1976, e na Resolução CVM n.º 44, de 23 de agosto de 2021, e em continuidade ao fato relevante divulgado em 26 de novembro de 2023 a respeito do envio de proposta de combinação de negócios ao Conselho de Administração da Vibra Energia S.A. ("Proposta" e "Vibra") (B3: VBBR3), vem comunicar a seus acionistas e ao mercado em geral que, por meio de correspondência recebida na data de hoje (cuja cópia consta anexa), a Vibra informou à Eneva que seu Conselho de Administração rejeitou a Proposta. A administração da Companhia avaliará tal resposta oportunamente e a Eneva se compromete a manter seus acionistas e o mercado em geral informados a respeito de novos desdobramentos relevantes a respeito deste tema na forma da lei e da regulamentação da Comissão de Valores Mobiliários – CVM.

C.3. Petrobras

Petrobras expresses interest in area under the Production Sharing regime

Rio de Janeiro, January 24, 2024 – Petróleo Brasileiro S.A. – Petrobras informs that it expressed today to the National Energy Policy Council (CNPE) its interest in the right of first refusal in a block to be tendered in the Permanent Offer System, under the Production Sharing Regime, under the terms of Law 12,351/2010 and Federal Decree 9,041/2017. Petrobras approved the expression of interest in the right of first refusal in the Jaspe block, located in the Campos Basin, considering the parameters disclosed in CNPE Resolution No. 11, of December 20, 2023, published on December 27, 2023. The expression of interest is in line with the E&P strategy set out in SP 24-28+, focusing on profitable assets and replenishing oil and gas reserves. Material facts on the subject will be disclosed to the market in due course.

C.4. Twilio

Costs Associated with Exit or Disposal Activities.

On December 4, 2023, Twilio Inc. (the "Company", "we" or "our") committed to a further workforce restructuring plan (the "December Plan") intended to streamline operations and accelerate the Company's path to delivering profitable growth. The December Plan includes the elimination of approximately 5As a result of the December Plan, the Company estimates that it will incur approximately $25-$35 million in charges in connection with the workforce reduction, consisting of expenditures for employee transition, notice period and severance payments, employee benefits, and related facilitation costs, substantially all of which are expected to result in future cash outlays. The Company expects that the majority of the restructuring charges related to the December Plan will be incurred in the fourth quarter of 2023 and that the execution of the December Plan, including cash payments, will be substantially complete by the end of the first quarter of 2024. Potential position eliminations in each country are subject to local law and consultation requirements, which may extend this process beyond the first quarter of 2024 in certain countries. The charges that the Company expects to incur are subject to a number of assumptions, including local law requirements in various jurisdictions, and actual expenses may differ materially from the estimates disclosed above. As part of the December Plan, Twilio Flex, the Company's cloud contact center, will be reported as part of the Company's Twilio Communications reportable segment in future periods. Prior periods presented for purposes of comparison will be recast accordingly.

D. UniEval Implementation Details

Given the constraints of the context window in the model originally associated with the UniEval (Zhong et al., 2022) metric, we implemented an approach on GPT-4-turbo that focuses on the originally proposed dimensions: *coherence, consistency, fluency, relevance,* and *groundedness*. The questions were adapted in the following manner, in which $analyses$ pertain to the text generated and $document$ refers to the associated MF form:

Coherence Is this an analysis with ideas that are coherent with each other?

Consistency Is this an analysis consistent with the facts in the document?

Fluency Is this a fluent analysis?

Relevance Is this an analysis that covers the most relevant topics of the document?

Groundedness Does this analysis develop insights derived from the knowledge in the document?

Therefore, GPT-4 was required to respond to the mentioned questions with "Yes" or "No". Subsequently, we used the $top\_logprobs$ API parameter to compute the formulations as originally proposed by the metric. Additionally, it is pertinent to underline that all the dimensions in use have their results ranging from 0 to 1.

E. Examining AI-generated vs. human-written analyses

This section highlights the similarities and differences observed between an AI-generated analysis and the human-written reference. The analyses are presented in Appendices E.1 and E.2, respectively; the specific MF form addressed by these texts is outlined in Appendix C.3, pertaining to the Petrobras case.

Foremost, the analysis by Mistral 7B is notably shorter, as previously indicated in the Table 2. Both texts rely on the events described in the MF. Specifically, the AI-generated text concentrates on detailing the facts presented in the form. Furthermore, the texts diverge in their coverage of certain topics. The AI-generated text focuses on the acquisition regime present in the MF and its potential implications (*"The Production Sharing Regime allows for the sharing of risks and rewards between the government and the private sector, which can encourage investment in exploration and production"*). On the other hand, the human-written text explores additional themes beyond just oil production potential increase, such as job creation, economic stimulus, and technological advancement.

Nonetheless, both texts address topics which are not explicitly mentioned in the MF. Notably, both point out that the company is under Brazilian state control, a potential important detail for contextualizing the information for readers. Similarly, each text acknowledges the company's long-standing expertise in the region, with Mistral 7B noting, *"The basin has a long history of oil and gas production,* *and Petrobras has been a key player in the region for many years..."* while the human analysis adds, *"Petrobras, as the operator of the surrounding fields in the Campos Basin, has the necessary expertise and infrastructure for the swift development of the Jaspe block"*. Both analyses discuss risks, though in different aspects. The AI-generated text highlights competition from other oil and gas companies and potential regulatory challenges. In contrast, the human-written text mentions actual issues like environmental concerns and fluctuating oil prices.

Nevertheless, while both cover important common points, the approach to certain topics varies. For instance, the AI-generated analysis superficially mentions, *"The Jaspe block is located in the Campos Basin, which is among the most prolific oil and gas basins in Brazil"*. It may been inferred from the MF that Petrobras's strategy includes *"...focusing on profitable assets and replenishing oil and gas reserves..."*. Conversely, the human-written text offers a more quantitative perspective and is supported by additional data, with the statement, *"...Campos Basin, a geological treasure trove responsible for about 80% of Brazil's oil production. This offshore block is estimated to contain up to 3 billion barrels of oil equivalent..."*.

In addition, another significant aspect is found in the recommendations sections of both texts. While the Mistral 7B concludes its analysis by revisiting the mentioned risks and pointing out the event as potentially positive move, the human-written reference brings in additional external context by evaluating other company alternatives for potential investment, ultimately recommending a "hold" stance.

Although here we refrain from determining which of the two texts addresses the event more effectively due to our lack of specialized knowledge in the finance sector, it is noteworthy that both texts present topics relevant to the MF, albeit from different styles. We acknowledge that these differences may be pivotal and significant in shaping a reader's direct preference, and therefore justifies further investigation.

E.1. Analysis generated by Mistral 7B

Petrobras' Interest in Jaspe Block Under Production Sharing Regime: An Analysis

Petrobras, the state-owned oil and gas company in Brazil, has recently expressed interest in the right of first refusal in the Jaspe block, located in the Campos Basin, under the Production Sharing Regime. This analysis will examine the material facts surrounding Petrobras' interest in the Jaspe block and provide recommendations for investors.
Petrobras has expressed interest in the right of first refusal in the Jaspe block, located in the Campos Basin, under the Production Sharing Regime. The expression of interest is in line with Petrobras' E&P strategy set out in SP 24-28+, focusing on profitable assets and replenishing oil and gas reserves.
Petrobras' interest in the Jaspe block is significant as it represents a potential opportunity for the company to increase its oil and gas reserves and maintain its position as a major player in the Brazilian energy sector. The Production Sharing Regime allows for the sharing of risks and rewards between the government and the private sector, which can encourage investment in exploration and production.
The Jaspe block is located in the Campos Basin, which is one of the most prolific oil and gas basins in Brazil. The basin has a long history of oil and gas production, and Petrobras has been a major player in the region for many years. The block is estimated to have significant oil and gas reserves, which could provide a significant boost to Petrobras' E&P activities.
Recommendation: Given Petrobras' interest in the Jaspe block and the potential benefits of the Production Sharing Regime, investors should closely monitor the company's activities in the region. Petrobras' E&P strategy is focused on profitable assets and replenishing oil and gas reserves, which could lead to increased production and revenue in the long term.
Investors should also be aware of any potential risks associated with Petrobras' interest in the Jaspe block, such as competition from other oil and gas companies in the region and potential regulatory challenges. However, given Petrobras' long history of success in the Brazilian energy sector and its focus on profitable assets, investors should view the company's interest in the Jaspe block as a positive development.
Conclusion: Petrobras' interest in the Jaspe block under the Production Sharing Regime represents a potential opportunity for the company to increase its oil and gas reserves and maintain its position as a major player in the Brazilian energy sector. Investors should closely monitor the company's activities in the region and be aware of any potential risks associated with Petrobras' interest in the Jaspe block. Overall, Petrobras' focus on profitable assets and replenishing oil and gas reserves could lead to increased production and revenue in the long term.

E.2. Human-written Analysis

Petrobras (PBR) Plans to Acquire Jaspe Block in Offshore Brazil[10]

Petrobras PBR, the leading Brazilian state-run oil company, announced its intention to acquire the

[10] https://bit.ly/zacks-pbr-jaspeblock

Jaspe oil block, located in the Campos Basin, by exercising its preferential rights. The offshore block is expected to be tendered by the country's regulator. The Jaspe block holds considerable significance for Brazil's oil production, and Petrobras anticipates substantial gains from the same. The block is estimated to contain significant oil reserves, which will help Petrobras increase its production capacity. Let's delve deeper into this strategic acquisition and explore its implications.

Jaspe Block: A Jewel in the Campos Crown

The Jaspe block sits nestled within the heart of the Campos Basin, a geological treasure trove responsible for roughly 80% of Brazil's oil output. This offshore block holds estimated reserves of up to 3 billion barrels of oil equivalent, making it a highly attractive prospect for any oil company. Its shallow waters and proximity to existing infrastructure further enhance its appeal, promising lower development costs and faster production timelines.

Petrobras' Strategic Plan: Unlocking Potential

Petrobras, as the operator of the surrounding fields in the Campos Basin, boasts the necessary expertise and infrastructure to swiftly develop the Jaspe block. This acquisition aligns perfectly with the company's strategic focus on maximizing output from mature basins while pursuing exploration in frontier areas. By leveraging its existing assets and knowledge, Petrobras can unlock the potential of the Jaspe block.

Increased Oil Production: Jaspe's reserves hold the potential to significantly boost Petrobras' production, bolstering its position as a key player in the global oil market. This translates to enhanced energy security and economic benefits for Brazil.

Job Creation and Economic Stimulus: Developing the Jaspe block will inevitably create new job opportunities, both directly in the oil and gas sector and indirectly through supporting industries. This economic stimulus will ripple through local communities, fostering growth and development.

Technological Advancement: Extracting oil from the Jaspe block will necessitate cutting-edge technologies and innovative solutions. This fosters a robust research and development environment, propelling Brazil's oil industry forward.

A Win for Brazil's Oil Ambitions. The Jaspe acquisition isn't just a win for Petrobras, it serves as a vital stepping stone for Brazil's broader oil production goals. The successful development of this block sends a positive signal to international investors, potentially attracting further investment and expertise to the Brazilian oil sector. This fosters healthy competition and accelerates exploration and development activities, propelling Brazil toward ambitious oil production targets.

Conclusion: Challenges and Considerations: While the Jaspe block holds immense potential, it's not without challenges. Environmental concerns surrounding offshore drilling activities require careful mitigation strategies. Additionally, fluctuating oil prices and global economic shifts can impact the project's viability. Petrobras must navigate these challenges prudently to ensure long-term success of the Jaspe acquisition.

A Strategic Move With Far-Reaching Implications

Petrobras' plan to acquire the Jaspe block marks a strategic move with the potential to significantly benefit both the company and Brazil's oil industry as a whole. Increased production, economic stimulus and technological advancements are just some of the rewards on the horizon. While challenges remain, the successful development of Jaspe could unlock a new chapter in Brazil's oil production story, solidifying its position as a major player in the global energy landscape.

Recommendation: Currently, PBR carries a Zacks Rank #3 (Hold). Investors interested in the energy sector might look at some better-ranked stocks like Sunoco LP SUN and Oceaneering International, Inc. OII, both sporting a Zacks Rank #1 (Strong Buy), and Enbridge Inc. ENB, carrying a Zacks Rank #2 (Buy) at present. You can see the complete list of today's Zacks #1 Rank stocks here.

Exploring Large Language Models in Financial Argument Relation Identification

Yasser Otiefy, Alaa Alhamzeh
University Of Passau
yasser.otiefy@uni-passau.de, alaa.alhamzeh@uni-passau.de

Abstract

In the dynamic landscape of financial analytics, the argumentation within Earnings Conference Calls (ECCs) provides valuable insights for investors and market participants. This paper delves into the automatic relation identification between argument components in this type of data, a poorly studied task in the literature. To tackle this challenge, we empirically examined and analysed a wide range of open-source models, as well as the Generative Pre-trained Transformer GPT-4. On the one hand, our experiments in open-source models spanned general-purpose models, debate-fine-tuned models, and financial-fine-tuned models. On the other hand, we assessed the performance of GPT-4 zero-shot learning on a financial argumentation dataset (FinArg). Our findings show that a smaller open-source model, fine-tuned on relevant data, can perform as a much larger general-purpose one, showing the value of enriching the local embeddings with the semantic context of data. However, GPT-4 demonstrated superior performance with F1-score of 0.81, even with no given samples or shots. In this paper, we detail our data, models and experimental setup. We also provide further performance analysis from different aspects.

Keywords: natural language processing (NLP), argument mining, large language models (LLMs), zero-shot learning, GPT-4, financial domain

1. Introduction

Argumentation plays an indispensable role for financial professionals and market participants. Many investors wait for the quarterly announcements of publicly traded companies to make their investment decisions. The company presents its reports about the last quarter, and makes expectations for the next one, then has to answer professional analysts' questions during a public event of an Earnings Conference Calls (ECCs) (Price et al., 2012). Alhamzeh et al. studied intensively how to mine the arguments of company executives stated during those calls (Alhamzeh et al., 2022b). They revisited the topic and studied how to rank the quality of those arguments in (Alhamzeh, 2023a). They proposed five quality metrics and various types of premises and claims based on interdisciplinary literature. Their study found a considerable link between the argument quality and the relation type (support or attack) between the used premise and the final claim. In other words, an argument that consists of many supporting premises is more likely to be convincing than an argument with many attacking premises. Although discussing the opponent's view is valuable in some situations, the speaker has to state many supporting premises to win.

While this sounds just logical and straightforward, the argument relation detection or classification did not get fair exploration, in the literature, in comparison to other argumentation tasks (e.g., (Reimers et al., 2019; Wachsmuth et al., 2017)). This could be due to its complexity as a Natural Language Inference (NLI) task. However, as we have mentioned, we believe that the potential of solving this task with high accuracy would empower different directions. To clear any possible confusion, on the one hand, the argument relation identification task considers the detection of the relation between given two sentences, so classify them as "related" or "unrelated". In other words, detection if a relation exists between a given premise and claim (the main argument components). While, on the other hand, the argument relation classification task, considers the classification of related premises and claims into a support or attack relation. In our work, we tackle the first identification task, as it is the core stone to structure the argument in the first place.

Furthermore, we focus on the financial use case of argumentation. (Chen et al., 2021) demonstrated, in their book, the urgent need for the automatic mining of arguments in financial narratives and reports. Argument mining considers, mainly, the automatic detection of argument components (premise/claim), argument relations (support/attack), and argument quality assessment.

However, given the fact that financial language has its jargon and particular terms, the language model performance can vary a lot from other domains, even for a simple task like sentiment analysis (Chen et al., 2020). Therefore, the Financial NLP (FinNLP) domain has emerged as an interdisciplinary field, which thus fostered different shared tasks and workshops (e.g., (El-Haj et al., 2018; Shah et al., 2023; Chen et al., 2023a)).

Hence, we have to consider the financial language peculiarities, but also the argumentation discourse nature. Argumentation is proven to be domain-dependent. The structure of arguments can vary a lot between scientific argumentation (Ac-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 119–129
20 May, 2024.

cuosto and Saggion, 2020), legal argumentation (Urchs et al., 2020), or simply web argumentation (Habernal and Gurevych, 2017).

Therefore, with the recent advances in NLP, the need to examine their performance in financial argumentation becomes more urgent. For example, (Al Zubaer et al., 2023) found that a model like Roberta, fine-tuned on the task data, outperform the Generative Pre-trained Transformer (GPT) both versions GPT-3.5[1] and GPT-4 (Achiam et al., 2023) in the legal argument mining area. This raises a critical consideration for each domain. In this paper, we want to assess the performance of large language models in the financial argumentation domain.

In particular, we compare the zero-shot performance of GPT-4, with a wide range of open source Large Language Models (LLMs). We cluster the latter in three categories: general-purpose models (e.g., BERT (Devlin et al., 2019), Vicuna (Zheng et al., 2023)), debate-fine-tuned models (e.g., ArgumentMining-EN-ARI-Debate[2]), financial-fine-tuned models (e.g., FinBert).

The debate-fine-tuned models are fine-tuned on argumentation debate data, while the financial-fine-tuned models are fine-tuned on financial data. Thus, and as our task considers financial argumentation, we aim to inspect the impact of this background data in enriching the model's local embedding.

All in all, the literature lacks a fair exploration of the financial argument relation identification task[3]. We aim, in this study, to bridge this gap. In particular, the contributions of this paper are:

- Empirical study of zero-shot learning and a wide range of outstanding LLMs on Financial Argumentation dataset (FinArg).
- Comparison between the performance of general-purpose, debated-fine-tuned, and financial-fine-tuned LLMs given the nature of this interdisciplinary task.
- To the best of our knowledge, this is the first intensive study to examine recent LLMs on the argument relation identification task.

In Section 2, we navigate the state-of-the-art dedicated to LLMs in argument mining tasks. We overview our data, and methodology in Section 3. Afterward, we exhibit the evaluation results in Section 4. We further discuss and analyze our findings in Section 5. Finally, we conclude our work and open future perspectives in Section 6.

2. Related Work

The exploration of argument mining and text classification has burgeoned with the advent of LLMs. Those models are heavily trained on massive data to learn general language representations. This learned knowledge can be then transformed to downstream domains (or tasks) through the procedure of fine-tuning. This concept made a remarkable revolution in Natural Language Processing (NLP) and helped to solve many challenges, like the need for huge training datasets. However, the behaviour of fine-tuned models on out-of-domain data cannot be completely expected. For example, (McCoy et al., 2019) found that 100 instances of Bert reported performance inconsistency for out-of-domain tests. Similarly, Bert-like models report performance drop in out-of-domain experiments in (Yogatama et al., 2019).

(Ruiz-Dolz et al., 2021) explored BERT (Devlin et al., 2018), XLNet (Yang et al., 2019), RoBERTa (Liu et al., 2019a), DistilBERT (Sanh et al., 2019a), and ALBERT (Lan et al., 2019) in identifying argument relations, across various domains. They emphasized the challenge of argument mining due to data scarcity and introduced a comprehensive analysis using the US2016 debate corpus[4] and the Moral Maze corpus[5] and others. The study revealed that different models, especially RoBERTa variants, excel in predicting argument relation on all tested datasets ranging from 0.51 to 0.70 of F1-score, the variation depends on the dataset these models fine-tuned on. This work also emphasizes the potential of other transformer architectures in processing complex argumentative structures.

Since the announcement of GPT-3 in 2020, many studies demonstrated its capability to reach state-of-the-art performance on different NLP tasks without extensive training or fine-tuning. For instance, (Brown et al., 2020) presented a detailed exploration of GPT-3 few-shot learning to generate human-like text, answer questions, translate languages, and other tasks.

The prompt is the main hyperparameter to handle in this scenario. (Liu et al., 2021) provided an exhaustive review of prompt-based learning techniques within NLP. They systematically categorized and evaluated various prompting strategies that leverage the capabilities of pre-trained language models.

In terms of argument mining via LLMs, there have been a couple of research papers that study the

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

[2] https://huggingface.co/raruidol/ArgumentMining-EN-ARI-Debate

[3] We found only a sub-task in FinArg -1 considering the argument relation classification, which we will address in Section 2

[4] https://corpora.aifdb.org/US2016

[5] https://corpora.aifdb.org/mm2012

power of open-source models fined-tuned to generate semantically rich local embeddings, in comparison to the general OpenAI embeddings. For example, in the legal domain, (Al Zubaer et al., 2023) analyzed the performance of GPT-3.5 and GPT-4 models in classifying argument components (premise/claim) within the European Court of Human Rights dataset. The study found that baseline models (like Large BERT and Roberta) outperform GPT-3.5 and GPT-4, with no significant improvement of GPT-4, over GPT-3.5. Similarly, (Chen et al., 2023b) explored multiple computational argumentation tasks (e.g., claim detection, stance detection) using LLMs in zero-shot and few-shot settings, without any fine-tuning. They found that introducing more samples (longer context) could result in unnecessary information that might negatively affect the performance of smaller models.

From another perspective, (Hinton and Wagemans, 2023) studied how persuasive is AI-generated argumentation. By analyzing the quality of the GPT-3 generator, they concluded that it generated a variety of argument types, but can include fallacies, lacking a real sense of human realization and a cogent argument structure. This raises considerations about the comprehending and reasoning these models can do in argumentation discourses.

In the frame of FinArg-1 shared task (Chen et al., 2023a), argument relation identification task was proposed on a similar dataset derived from (Alhamzeh et al., 2022a), the best team scored 61.50% and 84.86% of macro and weighted F1-score, respectively. Their approach was based on the T5 model (Raffel et al., 2020), fine-tuned using the financial Phrasebank dataset (Malo et al., 2014).

In addition, (Loukas et al., 2023) investigated the use of GPT-3.5 and GPT-4 for few-shot text classification in finance using the Banking77 dataset (Casanueva et al., 2020), demonstrating that conversational LLMs can quickly deliver accurate results and, in some cases, outperform fine-tuned masked language models with fewer examples. However, the cost of subscription-based LLMs may be prohibitive for individuals or smaller organizations. (Li et al., 2023) investigates the efficacy of generically trained LLMs, including ChatGPT and GPT-4, across various financial text analytics tasks, demonstrating their superiority over domain-specific models in many cases but also noting limitations, particularly in tasks requiring deep semantic and structural analysis, this work provides a comprehensive evaluation across eight datasets from five categories of tasks, marking an initial exploration into the capabilities and limitations of LLMs in financial applications.

Hence, and as no consistent superior performance was demonstrated in the recent works on different domains and tasks, we explore in this paper a wide range of LLMs, inspecting their performance on the financial argumentation dataset. Our study is among the first ones to explore the argument relation detection task in a financial narrative.

3. Method

We provide in this section a detailed overview of the data, models, and our experimental setup.

3.1. Data

We conducted our experiments on the Financial Argumentation dataset *FinArg*, which was collected and annotated by (Alhamzeh et al., 2022b; Alhamzeh, 2023b). This data is publicly available[6], and covers the quarterly earnings conference calls of major corporations (Amazon, Apple, Microsoft, and Facebook[7]) spanning from 2015 to 2019.

The annotation of this data encompasses the following labels: *premise*, *claim*, *non-arg* on the sentence level, as well as *support/attack* label on the relation between related premises and claims. Therefore, and to be able to solve the relation identification problem, we had to deduce the unrelated relation examples from the data. Subsequently, we construct our data as follows:

- **Positive Sampling:** We concatenate each claim with every single corresponding premise using [SEP] token (i.e., claim [SEP] premise), and we label it with class '1', signifying a related pair. This outcome in about 5K samples generated from 2200 arguments.
- **Negative Sampling:** We pair the unrelated claim-premise pairs and label each with class '0'. By this, we got about 1M possible pairs.
- **Data Balancing:** To keep the data balanced, we randomly selected 5K negative samples.

Hence, our problem is a binary classification task, on a balanced dataset. We have approx. 10K data samples formatted as the following:

- **Input –> {Claim} [SEP] {Premise}**
- **Output –> "1" or "0"**

3.2. Models

In this section, we elaborate on our models and experimental setup. We have examined two families of state-of-the-art large language models. On the first hand, fine-tuned models from Huggingface[8],

[6] FinArg Dataset
[7] Currently Known as Meta
[8] https://huggingface.co

and on the other hand, GPT language model from OpenAI[9]. This setting allows us to inspect the impact of the fine-tuning phase on the output in comparison to generative models where the prompt plays a considerable role.

3.2.1. Fine-tuned Large Language Models

To investigate the potential of open-source LLMs in argument relation identification, we examine in our study three categories of models, based on their training data, and intended application. This classification enables a focused analysis of each model's performance, especially in tasks that align with their customized training. We provide in the following an overview of those categories, and the examined models corresponding to each.

1. **General-purpose models:** This category encompasses original models that have been trained on general domain-agnostic data. These models are designed to perform a variety of natural language understanding tasks across different domains due to their diverse training backgrounds. Our used models from this category include:

    - *Bert-base-uncased* (Devlin et al., 2019)
    - *Roberta-base* (Liu et al., 2019b)
    - *Distilbert-base-uncased* (Sanh et al., 2019b)
    - *Bloom (560m,1b,7b)* (Workshop et al., 2022)
    - *BloomZ* (Muennighoff et al., 2022)
    - *LLaMa-2-7B-Guanaco-QLoRA-GPTQ*[10] a fine-tuned version of Llama 2 (Touvron et al., 2023)
    - *Vicuna*: is a chat assistant trained by fine-tuning LLaMA on user-shared conversations collected from ShareGPT. We test two versions *(Vicuna13bv1.5 and Vicuna-13b_rm_oasst_hh*[11]) (Zheng et al., 2023)
    - *GPT4-X-Alpaca*[12] a finetuned on GPT4's responses, for 3 epochs of a base model Alpaca (Taori et al., 2023)

2. **Debate-fine-tuned models:** Models in this category have been specifically fine-tuned on datasets featuring argumentative structures derived from debate content, which can be related to finance. They are optimized to discern and process argumentative nuances, making them well-suited for applications of argument mining. We include in this category:

    - *ArgumentMining-EN-ARI-Debate*, *ArgumentMining-EN-AC-Essay-Fin*, *ArgumentMining-EN-AC-Financial*, *ArgumentMining-EN-CN-ARI-Essay-Fin*[13]: All adopted from (Ruiz-Dolz et al., 2021), as fine-tuned versions of (Conneau et al., 2019) on different datasets such as US2016-test, MM2012, Bank, Money and others. For more details about those models, please refer to (Ruiz-Dolz et al., 2021).
    - *Roberta-argument*[14] trained on 25k heterogeneous manually annotated sentences by (Stab et al., 2018) and *Roberta-base-150T-argumentative-sentence-detector*[15]: A fine-tuned version of RoBerta (Liu et al., 2019b) using FS150T-Corpus dataset by (Schiller et al., 2022).

3. **Financial-fine-tuned models:** Our third category consists of models that have been fine-tuned with financial datasets, aiming to address classification challenges pertinent to the financial sector. These models leverage financial discourse and numeric data to provide insights specific to financial contexts. Namely:

    - *Finbert* (Araci, 2019) involves enhancing the BERT language model specifically for the finance sector. This is achieved by training it on a substantial corpus of financial documents, subsequently refining its capabilities for classifying financial sentiment. For this fine-tuning process, the Financial PhraseBank, created by (Malo et al., 2014), is employed.
    - *Finbert-tone-finetuned-finance-topic-classification* (Hazourli, 2022): Fine-tuned version on sentiment analysis task on Financial PhraseBank by (Malo et al., 2014).
    - *Deberta-v3-base-finetuned-finance-text-classification*[16]: Fine-tuned version of Deberta (He et al., 2021) tuned on financial-classification dataset[17].

[9] https://openai.com
[10] https://huggingface.co/TheBloke/llama-2-7B-Guanaco-QLoRA-GPTQ
[11] https://huggingface.co/reciprocate/vicuna-13b_rm_oasst-hh
[12] https://huggingface.co/chavinlo/gpt4-x-alpaca
[13] https://huggingface.co/raruidol
[14] https://huggingface.co/chkla/roberta-argument
[15] https://huggingface.co/pheinisch/roberta-base-150T-argumentative-sentence-detector
[16] https://huggingface.co/nickmuchi/deberta-v3-base-finetuned-finance-text-classification
[17] https://huggingface.co/datasets/nickmuchi/financial-classification

- *Roberta-Earning-Call-Transcript-Classification*[18]: Fine-tuned model from the base model RoBerta (Liu et al., 2019b) tuned on extracted a decade's worth of earnings call transcripts for 10 corporations, including Apple, Google, Microsoft, Nvidia, Amazon, Intel, Cisco, and others.

In all these categories, we conduct 5-fold cross-validation, with hyperparameter optimization as follows:

- Learning rate (2e^{-5}, 3e^{-5}, 5e^{-5})
- Maximum length of the tokenizer (64, 128, 256)
- Number of epochs (ranging from 2 to 5)

Please note that all fine-tuned models are trained on 2 x NVIDIA A100 80GB GPUs using Pytorch Lightening and HuggingFace frameworks with global seed 42.

3.2.2. GPT-4 Zero-Shot Learning

In our experiments, we explore the capability of the *GPT-4* model (Achiam et al., 2023) to detect the relation between a given claim and premise, using zero-shot learning (Xian et al., 2018).

Zero-shot learning refers to the model's ability to understand and perform tasks without the need for a specific training dataset tailored to that task. Recently, it has shown a very competent performance in various NLP tasks (Wei et al., 2021; Brown et al., 2020).

Prompt Design As prompting has not been yet explored in the task of financial argument relation detection, and due to budget constraints, we chose to follow a basic handcrafted prompt. This is also justified by the fact that the prompt has a significant impact in few-shot learning where choosing the number of shots, and choosing the example(s) play a crucial role, also this is impacted by budget constraints whereas we apply a zero-shot experiment.

Therefore, we decided to follow a straightforward approach that gathers the context and the instruction to the model (Brown et al., 2020). Obviously, we consider carefully OpenAI recommendations and prompt guide[19] as well as the prompt engineering guide[20].

[18] https://huggingface.co/NLPScholars/Roberta-Earning-Call-Transcript-Classification

[19] https://platform.openai.com/docs/guides

[20] https://www.promptingguide.ai/techniques/zeroshot

Since we aim to classify the relation between a given claim and premise as either *Related* or *Unrelated*, we formulate our prompt to clarify those two explicitly and then ask for the output class, as shown in the function `generate_messages` in the following:

```
def generate_messages(claim, premise):
    messages = [
        {"role": "system", "content":
            "You are a helpful
            assistant. Given the
            following claim and
            premise, please classify
            the relation between them
            as either Related or
            Unrelated. Please only
            generate one of the two
            labels."},
        {"role": "user", "content":
            f"Claim: {claim}"},
        {"role": "user", "content":
            f"Premise: {premise}"},
    ]
    return messages
```

This function encapsulates the interaction pattern with the model, where the model is first instructed about its role and the task's objective. Following this, the claim and premise are presented for classification.

Post-Processing of GPT-4 Output Following the interaction with the *GPT-4* model (Achiam et al., 2023), a crucial step is required to accurately extract the classification labels. The model responses are encapsulated within structured formats either as content within the interaction messages or through explicit function call objects which require systematic extraction processes to discern the relation classification between claims and premises. In other words, we had to check the extracted class label, to ensure it aligns with the expected output format and classification options ('Related' or 'Unrelated'). In some cases, the model responds by undefined class, then we have to extract it from the function call[21] output, if it does not exist in both response and function call response we label the sentence with "Unrelated" since this is the safe solution.

4. Results

In our comprehensive evaluation of argument relation identification, we explored a wide spectrum of fine-tuned Large Language Models (LLMs) alongside the innovative zero-shot learning capabilities of *GPT-4*, unveiling a fascinating landscape of performance across models tailored for General-purpose, Debate-fine-tuned, and Financial-fine-tuned tasks.

[21] https://platform.openai.com/docs/guides/function-calling

To have comparable results, we train the fine-tuned models in a cross-validation approach, where each part of the data is a test set at some fold. We then consider all data (all possible test sets) as the test data for *GPT-4*. Therefore, we report in Table 1 the average performance of the fine-tuned models along with the standard deviation, while we report in Table 2 the outcomes of *GPT-4* considering all the data.

Our results show that *GPT-4* was the most efficient performer by a macro F1-score of 0.81, confirming its ability to grasp the nuances of argumentative relations without explicit task-specific training.

However, among the fine-tuned models, *Vicuna-13b_rm_oasst_hh*, and *ArgumentMining-EN-ARI-Debate* showed a good performance with a mean macro F1 Score of 0.751. Despite the huge difference in the number of parameters between those two models, the latter behaved closely to Vicuna, only by having it already fine-tuned on debate data. This reflects the custom data impact on handling domain-specific argumentation. Yet, both models of *ArgumentMining-EN-CN-ARI-Essay-Fin* and *ArgumentMining-EN-AC-Financial* exhibited poor recognition of the argument relation.

In the series of Bloom models, the version of *Bloom 7b* parameters achieved a mean F1-score of 0.65, whereas a random guess behaviour was observed by *Bloom 560 m*, *Bloom 1b*, and *Bloomz 7b*. Similarly, FinBert, llama-2, Bert, and Alpaca showed weak efficiency. At the bottom of the list, lags *Roberta-Earning-Call-Transcript-Classification*, with an F1-score of 0.371, indicating a potential misalignment with the dataset's characteristics or the need for further tuning.

Our zero-shot learning experiment, which was conducted with *GPT-4*, is detailed in Table 2. It reveals *GPT-4* robust classification ability, with a precision of 0.85 for "Related", and 0.77 for "Unrelated" classes, reflecting a balanced understanding of both relationship types. This performance is further encapsulated in the precision-recall balance, with *GPT-4* favouring recall for "Unrelated" (0.87) over "Related" (0.75), suggesting a slight inclination towards conservatively identifying unrelated pairs to mitigate the risk of false positives in argumentative contexts.

The aggregate analysis does not only highlight the superior adaptability and understanding of *GPT-4* in zero-shot learning scenarios but also points to significant variations in the effectiveness of fine-tuned models across different categories. These distinctions underline the importance of model selection tailored to the specific characteristics of the task at hand, where the data domain and the classification task's nature critically influence model performance. The breadth of models evaluated demonstrates a spectrum of capabilities, from the comprehension exhibited by *GPT-4* to the more domain-specific insights offered by models like *Vicuna 13b*, and *ArgumentMining-EN-ARI-Debate*.

5. Discussions

In this section, we will discuss the analysis of hyperparameters, also we will spotlight the models that significantly outperformed the other models and attempt to justify these gaps. Since our data is balanced, we will focus on discussing the mean macro F1-score as it captures the harmonic mean of precision and recall.

The variability in performance as indicated by the standard deviation from the 5-fold cross-validation process as shown in Table 1 reveals insights into model stability. In general, models showed low standard deviations, suggesting consistent performance across different data folds and thus, greater reliability in practical applications.

The impact of model size on the F1-score in Figure 1 was evident from the visual data. While larger models generally achieved higher F1-score, indicating better generalization, the increase of model size did not always correlate with proportional improvements of results. This suggests a point of diminishing returns, where additional model complexity yields minor improvements at a significant computational cost. However, some models with a small number of parameters achieved relatively good performance. Potential reasons are the domain of the data those models used for tuning and also the task that those models tuned on, when possibly similar to our task, argument relation identification.

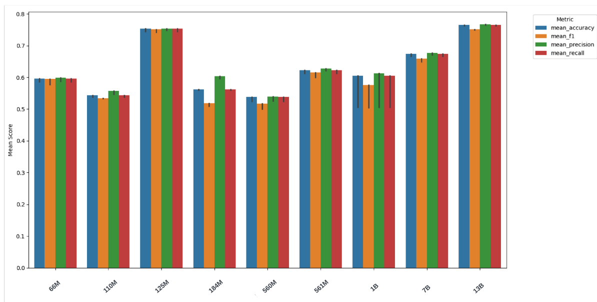


Figure 1: A grouped bar chart displaying the comparison of four metrics mean (accuracy, F1 score, precision, and recall) across models of various sizes.

Figure 2 indicates the performance of the three categories of open-source models we have experimented with. It reflects that Debate-fine-tuned and General-purpose models have a comparable mean macro F1-score, outperforming the Financial-fine-tuned models. This may suggest that general reasoning knowledge learned in debate-fine-

Model	Accuracy	F1-score	Precision	Recall	Model Type
Vicuna-13b_rm_oasst-hh	0.764 ± 0.05	**0.751 ± 0.05**	0.767 ± 0.05	0.764 ± 0.05	
Vicuna-13b-v1.5	0.762 ± 0.05	0.750 ± 0.05	0.762 ± 0.05	0.762 ± 0.05	
Bloom-7b1	0.675 ± 0.04	0.659 ± 0.06	0.677 ± 0.04	0.674 ± 0.04	
Bloom-1b1	0.567 ± 0.04	0.549 ± 0.05	0.572 ± 0.04	0.567 ± 0.04	
Bloomz-7b1	0.567 ± 0.02	0.534 ± 0.03	0.573 ± 0.02	0.567 ± 0.02	
Bloom-560m	0.531 ± 0.02	0.507 ± 0.03	0.530 ± 0.02	0.531 ± 0.02	General-Purpose Models
Bert-base-uncased	0.532 ± 0.01	0.503 ± 0.03	0.541 ± 0.02	0.532 ± 0.01	
GPT4-x-Alpaca	0.558 ± 0.04	0.536 ± 0.04	0.561 ± 0.04	0.558 ± 0.04	
LLaMa-2-7B-Guanaco-QLoRA-GPTQ	0.517 ± 0.01	0.468 ± 0.06	0.504 ± 0.09	0.517 ± 0.01	
Roberta-base	0.547 ± 0.03	0.479 ± 0.09	0.563 ± 0.13	0.547 ± 0.03	
ArgumentMining-EN-ARI-Debate	0.753 ± 0.02	**0.751 ± 0.02**	0.753 ± 0.01	0.753 ± 0.02	
ArgumentMining-EN-AC-Essay-Fin	0.622 ± 0.04	0.615 ± 0.04	0.627 ± 0.02	0.622 ± 0.02	
Roberta-base-150T-argumentative-sentence-detector	0.578 ± 0.01	0.569 ± 0.01	0.584 ± 0.02	0.578 ± 0.02	Debate-fine-tuned Models
ArgumentMining-EN-CN-ARI-Essay-Fin	0.532 ± 0.01	0.492 ± 0.07	0.540 ± 0.06	0.532 ± 0.01	
ArgumentMining-EN-AC-Financial	0.530 ± 0.02	0.480 ± 0.08	0.536 ± 0.09	0.530 ± 0.02	
FinancialBERT-Sentiment-Analysis	0.518 ± 0.02	**0.514 ± 0.02**	0.518 ± 0.02	0.518 ± 0.02	
Roberta-Earning-Call-Transcript-Classification	0.503 ± 0.01	0.371 ± 0.07	0.359 ± 0.14	0.503 ± 0.01	Financial-fine-tuned Models
Finbert	0.516 ± 0.02	0.507 ± 0.03	0.517 ± 0.02	0.516 ± 0.02	
Deberta-v3-base-finetuned-finance-text-classification	0.554 ± 0.01	0.505 ± 0.03	0.589 ± 0.02	0.554 ± 0.01	

Table 1: Classification performance metrics of LLMs on argument relation identification using 5-fold cross-validation. All models reported here are fine-tuned for 5 epochs, except Bloomz-7b1, for 2 epochs. The learning rate for all models is $5e^{-5}$

Class	Precision	Recall	F1-score	Support
Related	0.85	0.75	0.79	4899
Unrelated	0.77	0.87	0.82	4899
Accuracy			0.81	9798
Macro Avg	0.81	0.81	0.81	9798
Weighted Avg	0.81	0.81	0.81	9798

Table 2: Classification performance metrics of *GPT-4* zero-shot learning

tuned models is more valuable than the financial background knowledge learned in the Financial-fine-tuned models. Yet, the performance between Debate-fine-tuned models and General-Purpose Models is comparable, which could rely on the size of the latter. Therefore, we suggest examining smaller LLMs for a low tuning cost before looking for huger models, especially in a small dataset setting.

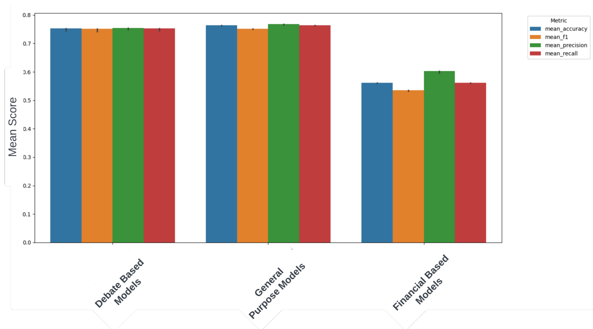


Figure 2: Performance among the three categories of fine-tuned models (Debate-fine-tuned, General-purpose, Financial-fine-tuned)

Figure 3, and the Pearson correlation heat map presented in Figure 4 provide an understanding of the relationship between hyperparameters and F1-score. Certain hyperparameters such as epochs and learning rate showed positive correlations with the F1-score. Potentially, since we give the model the chance to distil the pattern of our data, which means the more epochs we give to the model to train, the better the model learns.

Hyperparameters like maximum input length (max length), did not exhibit a very strong relationship with mean F1-score since most of the data points, as shown in Figure 5, are less than the smallest value of the max length hyperparameter ranging from (64 to 256) and the frequency of the examples that has 64 tokens or less is dominant. However, the correlation still exists which means the longer the sentence is fed to the model without truncation, the better performance the model achieves. However, a complex interplay between these hyperparameters requires careful tuning to optimize performance.

We also have noticed that the standard deviation, in general, is small which means the consistent performance of such models with low standard deviation, however, some models have a slightly larger standard deviation such as Roberta-base and *ArgumentMining-EN-AC-Financial*, One of the reasons could be the type of data these models fine-tuned on which made those models overfitted and stuck in a local minimum because of such past fine-tuning.

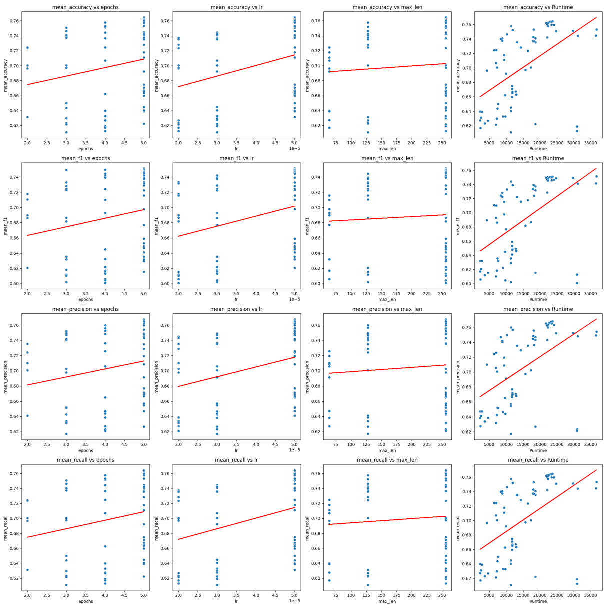


Figure 3: Correlation between hyperparameters (epochs, learning rate, input max length, runtime) and the performance metrics of fine-tuned models (accuracy, F1-score, precision, recall)

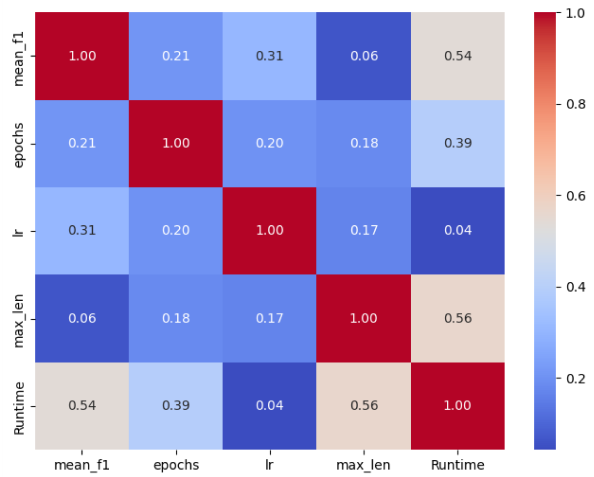


Figure 4: The heat map shows that learning rate and runtime, maximum input length and epochs correlation with mean F1-score.

6. Conclusion

The automatic mining of arguments (components and relations) has become an essential tool for multiple applications like assisted writing, fact-checking, search engines, law, and decision-making aid systems. In this paper, we investigated argument mining in financial texts, In particular, the task of relation detection between given two sentences (potential argument components) within the context of earnings conference calls.

Our experimental study encompasses a wide range of LLMs, including *GPT-4*, debate-fine-tuned models, and financial-fine-tuned models. The performance of open-source models ranged from 0.37 to 0.75 in terms of F1-score, while *GPT-4* zero-shot learning achieved 0.81. This superior performance of *GPT-4* highlights its potential to adapt to complex language understanding tasks, without any further training. Moreover, we believe that this outcome can be significantly improved with few-shot learning, or exploring other prompting techniques in future work.

In closing, our study contributes to the literature of argument mining in the financial domain by providing a comprehensive evaluation of various LLMs and illustrating the potential of zero-shot learning in understanding the nuances of financial discourse.

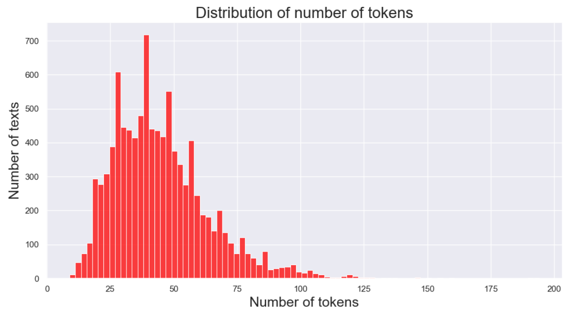


Figure 5: Distribution of sentence length

7. Bibliographical References

Pablo Accuosto and Horacio Saggion. 2020. Mining arguments in scientific abstracts with discourse-level embeddings. *Data Knowledge Engineering*, 129:101840.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abdullah Al Zubaer, Michael Granitzer, and Jelena Mitrović. 2023. Performance analysis of large language models in the domain of legal argument mining. *Frontiers in Artificial Intelligence*, 6.

Alaa Alhamzeh. 2023a. Financial argument quality assessment in earnings conference calls. In *International Conference on Database and Expert Systems Applications*, pages 65–81. Springer.

Alaa Alhamzeh. 2023b. *Language Reasoning by means of Argument Mining and Argument Quality*. Ph.D. thesis, Universität Passau.

Alaa Alhamzeh, Előd Egyed-Zsigmond, Dorra El Mekki, Abderrazzak El Khayari, Jelena Mitrović, Lionel Brunie, and Harald Kosch. 2022a. *Empirical Study of the Model Generalization for Argument Mining in Cross-Domain and Cross-Topic Settings*, pages 103–126. Springer Berlin Heidelberg, Berlin, Heidelberg.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022b. It's time to reason: Annotating argumentation structures in financial earnings calls: The finarg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Issues and perspectives from 10,000 annotated financial social media data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6106–6110.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From Opinion Mining to Financial Argument Mining*. Springer Nature.

Chung-Chi Chen, Chin-Yi Lin, Chr-Jr Chiu, Hen-Hsen Huang, Alaa Alhamzeh, Yu-Lieh Huang, Hiroya Takamura, and Hsin-Hsi Chen. 2023a. Overview of the ntcir-17 finarg-1 task: Fine-grained argument understanding in financial analysis. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo, Japan*.

Guizhen Chen, Liying Cheng, Luu Anh Tuan, and Lidong Bing. 2023b. Exploring the potential of large language models in computational argumentation. *arXiv preprint arXiv:2311.09022*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mahmoud El-Haj, Paul Rayson, and Andrew Moore. 2018. The first financial narrative processing workshop (fnp 2018). In *Proceedings of the LREC 2018 Workshop*.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Ahmed Hazourli. 2022. Financialbert-a pretrained language model for financial text mining. *Research Gate*, 2.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Martin Hinton and Jean HM Wagemans. 2023. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator. *Argument & Computation*, (Preprint):1–16.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are ChatGPT and GPT-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422, Singapore. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55:1 – 35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the bank with ChatGPT: Few-shot text classification for finance. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 74–80, Macao. -.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

R Thomas McCoy, Junghyun Min, and Tal Linzen. 2019. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

S McKay Price, James S Doran, David R Peterson, and Barbara A Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36(4):992–1011.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821*.

R. Ruiz-Dolz, J. Alemany, S. Barbera, and A. Garcia-Fornes. 2021. Transformer-based models for automatic identification of argument relations: A cross-domain evaluation. *IEEE Intelligent Systems*, 36(06):62–70.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019a. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019b. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2022. On the effect of sample and topic sizes for argument mining datasets. *arXiv preprint arXiv:2205.11472*.

Sameena Shah, Xiaodan Zhu, Wenhu Chen, Manling Li, Armineh Nourbakhsh, Xiaomo Liu, Zhiqiang Ma, Charese Smiley, Yulong Pei, and Akshat Gupta. 2023. Knowledge discovery from unstructured data in financial services (kdf) workshop. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3464–3467.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Stefanie Urchs, Jelena Mitrović, and Michael Granitzer. 2020. Towards classifying parts of german legal writing styles in german legal judgments. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 451–454. IEEE.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. 2018. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, et al. 2019. Learning and evaluating general linguistic intelligence. *arXiv preprint arXiv:1901.11373*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *CoRR*, abs/2306.05685.

Keyword-based Annotation of Visually-Rich Document Content for Trend and Risk Analysis using Large Language Models

Giuseppe Gallipoli[1], **Simone Papicchio**[1], **Lorenzo Vaiani**[1], **Luca Cagliero**[1], **Arianna Miola**[2,3], **Daniele Borghi**[2]

[1]Politecnico di Torino, Turin, Italy
[2]Intesa Sanpaolo Innovation Center, Turin, Italy
[3]Università degli Studi di Milano-Bicocca, Milan, Italy
{name.surname}@polito.it, {name.surname}@intesasanpaolo.com

Abstract

In the banking and finance sectors, members of the business units focused on Trend and Risk Analysis daily process internal and external visually-rich documents including text, images, and tables. Given a facet (i.e., topic) of interest, they are particularly interested in retrieving the top trending keywords related to it and then use them to annotate the most relevant document elements (e.g., text paragraphs, images or tables). In this paper, we explore the use of both open-source and proprietary Large Language Models to automatically generate lists of facet-relevant keywords, automatically produce free-text descriptions of both keywords and multimedia document content, and then annotate documents by leveraging textual similarity approaches. The preliminary results, achieved on English and Italian documents, show that OpenAI GPT-4 achieves superior performance in keyword description generation and multimedia content annotation, while the open-source Meta AI Llama2 model turns out to be highly competitive in generating additional keywords.

Keywords: Visually-Rich Document Understanding, Trend and Risk analysis, Large Language Models

1. Introduction

Understanding and exploring the content of visually-rich documents such as PDF files and scanned documents is of primary importance for trend and risk analysts of the banking and finance sectors. Since these documents have variable layout and content, with a mixture of text, images, and tables, their deep understanding requires both advanced multimodal learning capabilities.

The goal of this work is to enhance the research and analysis capabilities of a primary Italian financial institution, focusing on emerging trends within both the national and international contexts. Improving these functions is crucial for the strategic positioning of the bank and for providing value-added services to its customers. The partial automation of the research process allows for the inclusion of a greater number of data sources that were previously untapped due to operational limits. Given the relentless flow of information in today's environment, this represents a strategic step towards expanded informational access and a stronger ability to proactively adapt to market evolution.

In this work, we provide bank analysts with a financial document annotator relying on multimodal Large Language Models (LLM). Given a topic of interest (hereafter denoted by *facet*), the LLM produces a list of facet-related keywords as well as the corresponding textual descriptions and high-dimensional vector representations. In parallel, the multimodal document content is split into textual paragraphs, images, and tabular elements and conveniently processed to generate embeddings of the equivalent text versions. Finally, the annotation process is tackled as a keyword retrieval task on the document elements driven by textual semantic similarity. An extensive empirical analysis, supported by a bilingual testing document collection and a team of experts who validated the keyword descriptions, provide an in-depth performance comparison between the open-source Meta AI Llama2 and the proprietary OpenAI GPT-4 models.

2. Problem statement

Given a set of multi-page financial documents $\mathcal{D}$ and a set of facets $\mathcal{F}$ describing the topics of interest, our purpose is threefold:

1. **Keyword generation and description**. Generate for each facet $f_i \in \mathcal{F}$ a set of keywords $k_j \in \mathcal{K}^i$ related to f_i. Next, annotate each keyword k_j with a free-text description *descr*(k_j) summarizing its general meaning.
2. **Captioning of non-textual document elements**. Produce textual descriptions of multimedia document elements $e_l \in \mathcal{E}^m$, where an arbitrary element e_l in a document $d_m \in \mathcal{D}$ can be either an image, a table, or a textual paragraph.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 130–136
20 May, 2024.

3. **Keyword-based content annotation**. For each element e_l, retrieve the keywords k_j that are most relevant to e_l.

Our goal is to compare the performance of LLMs, in zero-shot or few-shot learning, to address all the above-mentioned tasks. Hereafter, we will consider Llama2 (Touvron et al., 2023) (or its Italian version Camoscio (Santilli and Rodolà, 2023)) as representative open-source model and GPT-4 (OpenAI, 2023) as representative proprietary model.

3. Proposed approach

In the following, we describe the main steps of our method. A sketch of the proposed pipeline is displayed in Figure 1.

3.1. Generation of keywords and keyword descriptions

Given a user-provided facet name f_i, we use the LLM to automatically generate a set of related keywords k_j as well as the corresponding free-text descriptions *descr*(k_j).

We explore the following settings:

- *Zero-Shot learning – Cold Start Setting*: We prompt the LLM with the facet name only, assuming that neither facet-relevant keywords nor examples of textual descriptions are given.

- *Few-Shot learning – Cold Start Setting*: We prompt the LLM with the facet name and h examples chosen randomly from keywords and their corresponding descriptions, previously provided by the domain expert. Here, we assume that some examples of keyword descriptions are already available, but we do not know any facet-related keyword yet, since the selected examples are not necessarily related to the input facet.

- *Few-shot learning – Additional Keyword Recommendation*: We prompt the LLM with the facet name and h examples of facet-related keywords and their corresponding descriptions. Here, we assume that the examples are not chosen randomly but shortlisted by human expert (e.g., by validating a previous output).

In few-shot learning settings, we ensure that the examples of keywords and descriptions provided as input to the LLM do not overlap with the keyword currently being prompted.

The output of this step is then used in the keyword-based content annotation stage.

3.2. Document pre-processing

To process the input PDF documents, we extract the following three main elements: (i) Textual paragraphs (e.g., titles, sections, subsections), (ii) Visual items (e.g., images, sketch of architectures/processes/pipelines, iconography, graphical examples), and (iii) Tables.

Textual paragraphs and tables are extracted from PDF documents using the proprietary Document Intelligence service provided by the Azure AI platform (Azure, 2024). For visual and textual content extraction, we face the following challenges:

- *Slide extraction*: Some input documents consist of slide presentations, which appear to be unsuitable for text and image extraction using standard content extraction tools. To address this issue, we opportunistically generate textual explanations of the slide content using the Multimodal Large Language Model GPT-4 Vision (OpenAI, 2023). Specifically, we train an ad hoc CNN to automatically detect the presence of presentation slides on a PDF document page. If the current page is classified as a *slide*, then the input is processed directly by the Multimodal LLM.

- *Paragraph length*: Some extracted textual elements contain few words (likely due to a misalignment of PDF content). To avoid this issue, we prevent the generation of textual elements consisting of less than 4 words.

- *Redundant table content*: The textual content within table cells sometimes appears incorrectly twice, in separate tabular and textual elements. During table extraction, we early detect possible situations of overlap between the bounding box of the table and the position of the text. The purpose is to disregard duplicated text whenever it is not deemed relevant.

- *Irrelevant images*: The image detector module also recognizes irrelevant visual items such as banners or graphical separators. We define the boundary regions of each document page (e.g., the bottom of the page) and ignore all the images placed in those border regions, as they are unlikely to convey informative content. To prune irrelevant content, we apply the following filters to all visual elements: (1) *Minimum image size*: we drop visual elements containing less than 150 pixels; (2) *Minimum height-width ratio*: we drop visual elements whose absolute ratio is above 500% (i.e., greater than 5:1 or 1:5); (3) *Percentage of pixels of the same color*: we drop visual elements whose percentage of pixels with the same color is above 80%.

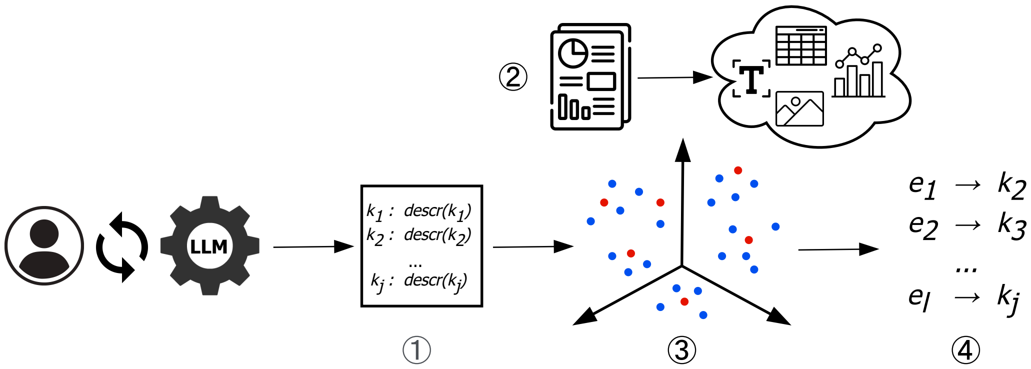


Figure 1: The figure illustrates the main steps of the proposed method: (1) keyword and description generation; (2) document preprocessing; (3) document element and keyword description encoding; and (4) keyword-based content annotation. In step (3), blue and red dots represent the embedding representations of document elements and keyword descriptions, respectively.

3.3. Keyword-based content annotation

For each document element e_l within each multi-page financial document $d_m \in \mathcal{D}$, we retrieve the keywords k_j that are the most relevant to e_l. Specifically, we return a ranked list $k^1_{e_l,d_m}, ..., k^K_{e_l,d_m}$ of the top-K keywords assigned to e_l. Notice that the assigned keywords can arbitrarily refer to any facet and, possibly, the retrieved list can be empty.

Focusing on text-only content, we address the retrieval of keywords relevant to each element in an unsupervised fashion using various textual similarity approaches, including both syntax-oriented and semantic-oriented methods. For each element of the document e_l, we assign the K keywords whose textual descriptions are most similar to e_l according to the following measures:

- Syntactic similarities: (1) **ROUGE-1/2/L F1-Score** (Lin, 2004) measures syntactic overlap in terms of common unigrams, bigrams or longest matching subsequence; (2) The **Levenshtein**, **Jaro**, and **Jaro-Winkler edit distances** measure the number of character-level operations needed to transform one piece of text into another.
- Semantic similarity: **SentenceBERT** (Reimers and Gurevych, 2019) and **proprietary embeddings**, used to compare document elements and keyword descriptions via cosine similarity.

Additionally, we experiment with **prompting GPT-4** with both the document element to be labeled and all possible keywords, asking the model to assign the K most pertinent ones.

Notice that, for the sake of simplicity, in Figure 1 document elements and keyword descriptions are displayed as embedding representations in a latent space. However, we also experiment with the syntactic similarity and prompting approaches discussed above.

4. Experimental evaluation

We run our experiments on a machine equipped with a single NVIDIA® RTX A6000 48GB GPU. We leverage standard Python libraries to calculate syntactic similarity measures, while for semantic similarity we rely on SentenceBERT `paraphrase-MiniLM-L6-v2` model and `text-embedding-ada-002` as proprietary OpenAI model. We employ Llama2-Chat 7B with 16-bit quantization. GPT-4 (`gpt-4-0613`), GPT-4 Vision (`gpt-4-1106-vision-preview`) and `text-embedding-ada-002` have all been accessed through OpenAI API.

Dataset. Business Units provided the following two in-domain datasets: (1) **ICT Risk Analysis**, consisting of 11 documents and annotated with 2 facets and 25 keywords. It contains 991 textual elements, 13 images, and 15 tables. (2) **Trend Analysis**, consisting of 4 documents, annotated with 1 facet and 12 keywords, and including 69 images. Most images are presentation slides, which are handled by the LLM to get the textual reformulation. We also have additional facets and keywords, along with their corresponding descriptions (92 overall), which analysts have not used for element annotation.

Evaluation Metrics. To evaluate the efficacy of element annotation, we employ the following metrics for information retrieval (Manning et al., 2008):

- **Precision at K** (P@K): percentage of returned keywords that occur in the expected keyword list.
- **Recall at K** (R@K): percentage of expected keywords that occur in the returned keyword list.
- **Mean Reciprocal Rank** (MRR): mean of the multiplicative inverse of the rank of the first correctly assigned keyword.

where K is the number of keywords retrieved that are considered. The rank order is based on the similarity score used to retrieve the keywords.

To assess keyword and description generation, we compare the produced and expected outcomes using the following established metrics for evaluating sequence-to-sequence models, i.e., ROUGE-1/2/L (R1/2/L) F1 score (Lin, 2004) for syntactic similarity and BERTScore (BS) F1-score (Zhang et al., 2020) for semantic similarity.

Prompt description. We present in the following some examples of prompts provided to the LLMs to perform keyword and description generation tasks. Prompts were selected according to preliminary experiments and their format may vary depending on the LLM under consideration.

Keyword generation: *The [K] most relevant keywords for the [FACET] domain are:*
where we replace [K] and [FACET] with the desired number of keywords and the facet name of interest, respectively.

Description generation: *Explain in a few lines the word between the quotation marks: "'[KEYWORD]"'*
where we replace [KEYWORD] with the keyword for which to generate a description.

When conducting experiments in the Italian language, we use the corresponding Italian translations as prompts.

4.1. Results on content annotation

Similarity measure	ICT Risk Analysis	Trend Analysis
R1	0.458	0.300
R2	0.367	0.279
RL	0.472	0.258
Levenshtein	0.347	0.247
Jaro	0.483	0.249
Jaro-Winkler	0.483	0.249
SentenceBERT	0.658	0.430
embedding-ada-002	**0.779**	**0.610**
GPT-4	0.729	0.500

Table 1: Mean Reciprocal Ranks.

Textual semantic similarity based on contextual embeddings and LLM prompting achieve very promising results (MMR above 0.7) and outperform both syntactic measures and edit distances (see Table 1). System's precision decreases while increasing the number K of retrieved keywords whereas its recall shows an opposite trend (see Figure 2). Similarity based on OpenAI embedding performs best, e.g., for $K = 3$, P@K $>$ 40% and R@K $>$ 50% on both ICT Risk and Trend.

4.2. Results on keyword and description generation

Tables 2 and 4 summarize system performance on keyword description and keyword generation tasks, respectively. Due to space constraints, we report here only the outcomes on a single dataset, i.e., ICT Risk, for both languages.

- *Proprietary vs. open-source LLM*: Proprietary GPT-4 performance is superior to that of open-source (Llama2/Camoscio) on keyword description generation for both tested languages (e.g., +33% ROUGE-1 on Italian documents). Conversely, open-source LLMs are highly competitive on keyword generation, likely because training examples of smaller models are more focused on specific domains, such as finance. This trend is confirmed by the results on Italian documents (not shown here due to space constraints).
- *Italian vs. English*: Both LLMs perform better on English than Italian text. The gap in performance is more evident for the open-source LLMs, e.g., ROUGE-1 for description generation 0.31 Italian vs. 0.39 English.
- *In-context learning*: Prompting LLMs with few training examples (from 3 to 5) turns out to be beneficial for both keyword generation and description generation. Few-shot learning has shown to be more beneficial for open-source LLMs because of their lower pre-trained model complexity.

4.3. Human evaluation

Each generated description, for both Italian and English languages, was annotated by five domain experts using a 5-point Likert scale based on five criteria (Iskender et al., 2021): (1) **Usefulness** (effectiveness in conveying key information); (2) **Coherence** (logical and semantic coherence); (3) **Non-Redundancy** (conciseness); (4) **Grammaticality** (linguistic correctness); (5) **Overall Quality** (holistic evaluation of the generated description).

Results (see Table 3) are satisfactory and coherent with quantitative outcomes (see Section 4.2). The perceived quality of Italian-written descriptions is lower than that of English ones, likely due to the more limited capabilities of LLMs on languages other than English.

4.4. Qualitative examples

To better illustrate the proposed approach, we provide examples of outputs of the different steps of our method.

Considering the ICT Risk Analysis dataset, one of the keywords associated with the *cyber risk* facet is *third-party risk*.

Reference description: *It refers to the potential risks or threats to an organization arising from relationships with third parties, such as suppliers, business*

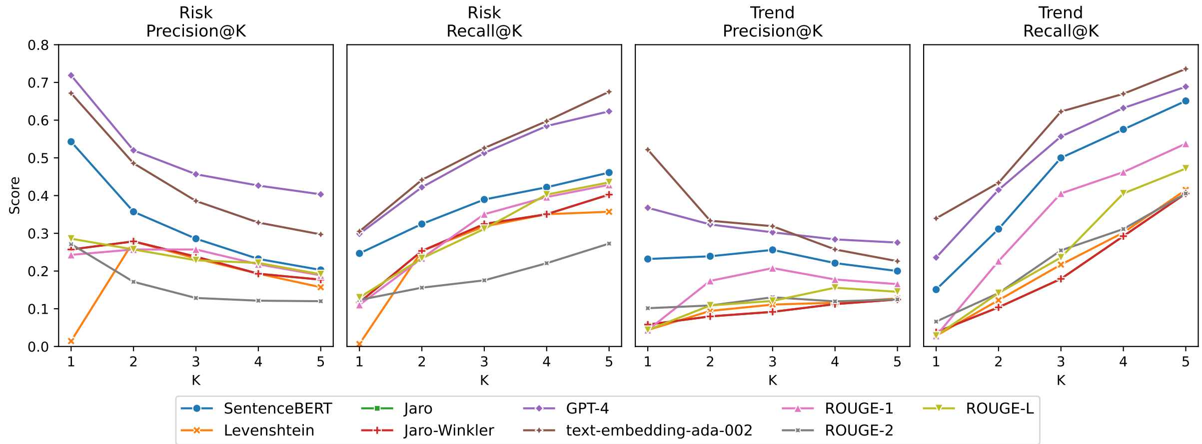


Figure 2: Precision@K and Recall@K values of different similarity measures on the ICT Risk Analysis (left) and Trend Analysis (right) datasets. English language.

	K = unspecified		$K = 3$		$K = 5$		$K = 10$		$K = 20$	
	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2
RL	0.051	**0.066**	**0.070**	0.062	0.058	**0.062**	0.057	**0.065**	0.058	**0.060**
BS	**0.860**	0.859	0.862	**0.864**	0.861	**0.863**	**0.865**	0.860	**0.861**	0.857
P@K	0.771	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.867	**0.944**	0.833	**0.894**
R@K	**0.447**	0.296	**0.133**	**0.133**	**0.221**	**0.221**	0.375	**0.420**	0.721	**0.783**

Table 4: Evaluation of keyword generation for varying K. ICT Risk Analysis dataset. English language.

	Italian		English	
	Camoscio	GPT-4	Llama2	GPT-4
R1	0.310	**0.413**	0.394	**0.437**
R2	0.082	**0.169**	0.131	**0.150**
RL	0.208	**0.279**	0.254	**0.284**
BS	0.719	**0.773**	0.760	**0.902**

Table 2: Evaluation of keyword description generation performance. ICT Risk Analysis dataset.

	Italian	English
Usefulness	**4.38**±1.60	4.33±1.78
Coherence	4.52±0.93	**4.62**±1.15
Non-Redundancy	4.38±1.12	**4.52**±1.11
Grammaticality	4.60±0.73	**4.81**±1.20
Overall Quality	4.33±1.37	**4.34**±1.66

Table 3: Human evaluation of keyword descriptions. ICT Risk Analysis dataset.

partners, or external contractors. These risks [...]

Generated description: *It is the risk that arises from the use of third-party vendors, suppliers, or partners that provide goods or services to an organization. Third-party risk can include a wide range of* [...]

Document element: *The image presents* [...] *in the context of retail banking leaders.* [...] *security providers aim to protect company, payment, card, and consumer data* [...] *the importance of various data privacy and security measures and where they stand in terms of industry focus and market trends.*

Target keywords: *third-party risk*, *regulation*
Assigned keywords: *third-party risk*, *regulation*, *compliance*

5. Conclusions

We presented an automatic pipeline for annotating visually-rich financial documents for Trend and Risk analysis in banking and finance sectors. The main takeaways can be summarized as: (1) *Semantic similarity*: proprietary embeddings outperform open-source solutions for both Italian and English text; (2) *Keyword generation*: open-source LLMs perform as good as or even better than GPT-4 in zero-shot and few-shot learning settings on the tested documents, likely due to a higher in-domain specialization; (3) *Description generation*: GPT-4 performs best, while open-source LLMs perform reasonably well. Human feedback is in line with quantitative results based on established performance metrics.

As future work, we will explore the integration of the proposed method in a Retrieval Augmented Generation system and address the task of zero-shot document classification using the additional keywords that have not been used for annotation yet. Moreover, we plan to assess the capabilities of other Multimodal LLMs (e.g., LLaVA (Liu et al., 2023)) to generate textual descriptions of multimedia document elements.

Limitations

Text-only processing. In this work, we focus on textual content, whether the content is originally text or is converted from a visual format. Consequently, we have not embedded visual and tabular content directly. We deemed such variant as a potentially valuable extension of the present work as enables the adoption of state-of-the-art multimodal learning techniques.

Limited robustness to document layout variety. Despite our efforts, the substantial variety in document structures, both within and across different domains, may introduce inconsistencies in document pre-processing and content extraction. This could potentially lead to suboptimal results in the subsequent content annotation phase. We intend to refine the document pre-processing and content extraction phase in alignment with the availability of new state-of-the-art document layout understanding models.

Limited scope of Multimodal LLM reasoning. When textual content cannot be successfully extracted, in particular from presentation slides, we rely on the GPT-4 Vision model to generate textual explanations. Although the LLM may disregard potential useful content and/or introduce inaccuracies in the generated textual explanation, also based on a manual inspection of a sample of outputs, we are confident that this approach is sufficiently satisfactory. However, we plan to conduct a more in-depth analysis to assess the LLM capabilities in providing textual explanations of visually-rich domain-specific content.

Ethical Considerations

The use of Large Language Models in critical sectors like banking and finance offers significant advantages, including improved efficiency, automation, and enhanced data analysis capabilities. These models can optimize processes, improve customer interactions, and contribute to informed decision-making. However, it is crucial to acknowledge that deploying LLMs in these domains also presents potential risks and undesired outcomes.

The complexity of banking and financial systems, coupled with the intricate nature of language understanding, may result in unintended consequences, such as biased generations or misinterpretation of information. However, in our specific use case, we view the LLM as an assistant to domain experts who remain fully responsible for the process, supervising the system outputs and possibly refining them through a human-in-the-loop approach. We believe that vigilant oversight, continuous refinement, and ethical considerations are essential to fully exploit the potential of LLMs while minimizing any adverse impacts on critical sectors such as the ones of our work.

We also acknowledge that the use of proprietary models may hinder transparency. However, the active involvement of domain experts who supervise the process is expected to alleviate this issue. Additionally, we sought to address this concern by experimenting also with open-source models.

Data and Code Availability Statement

Documents cannot be disclosed due to confidentiality and copyright issues. Code could be available upon request to the authors.

Conflicts of Interest

We have no Conflicts of Interest to declare.

Credits to financial institutions

Intesa Sanpaolo is a leading banking group in the Eurozone, and the most important one in Italy. Intesa Sanpaolo Innovation Center is part of ISP group, and its mission is exploring business models of the future to discover new assets and skills that support the long-term competitiveness of ISP group and its customers. ISP has established the Innovation Center Labs to respond to the complex needs of the bank and the market, determined by the evolution of market trends and exponential growth technologies.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments.

They would also like to thank Giorgio Bella, Anna Polise, Stefania Vigna, Laura Li Puma, Chiara Napione, Giovanni Troiano, Patrizio Paolo Dionisi, Maura Bertaglia, Carla Monferrato, and Simone Scarsi for their useful comments. The authors would like to thank Luigi Ruggerone for supporting the researcher team.

This work was partially carried out within the FAIR (Future Artificial Intelligence Research) and received funding from the European Union Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000013).

The work by Giuseppe Gallipoli was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from Next-GenerationEU (Italian PNRR – M4 C2, Invest 1.3 – D.D. 1551.11-10-2022, PE00000004). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

6. Bibliographical References

Microsoft Azure. 2024. Azure AI Document Intelligence.

Neslihan Iskender, Tim Polzehl, and Sebastian Möller. 2021. Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 86–96, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

OpenAI. 2023. GPT-4 technical report. *ArXiv*, abs/2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: an Italian instruction-tuned LLaMA.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

ESG-FTSE: A corpus of news articles with ESG relevance labels and use cases

Mariya Pavlova. Miaosen Wang, Bernard Casey
Imperial College London, Google DeepMind, SOCial ECONomic RESearch London and Frankfurt
Exhibition Rd, South Kensington, London SW7 2BX, United Kingdom
1600 Amphitheatre Parkway, Mountain View, CA, United States
16 Sandringham Court, Arregon Road, London TW1 3ND, United Kingdom
m.pavlova22@imperial.ac.uk, miaosen@deepmind.com, b.casey@soceconres.eu

Abstract

We present ESG-FTSE, the first corpus comprised of news articles with Environmental, Social and Governance (ESG) relevance annotations. In recent years, investors and regulators have pushed ESG investing to the mainstream due to the urgency of climate change. This has led to the rise of ESG scores to evaluate an investment's credentials as socially responsible. While demand for ESG scores is high, their quality varies wildly. Quantitative techniques can be applied to improve ESG scores, thus, responsible investing. To contribute to resource building for ESG and financial text mining, we pioneer the ESG-FTSE corpus. We further present the first of its kind ESG annotation schema. It has three levels: a binary classification (relevant versus irrelevant news articles), ESG classification (ESG-related news articles), and target company. Both supervised and unsupervised learning experiments for ESG relevance detection were conducted to demonstrate that the corpus can be used in different settings to derive accurate ESG predictions.

Keywords: corpus annotation, ESG labels, annotation schema, news article, natural language processing

1. Introduction

ESG is a framework that aims to capture all the non-financial information arising from a company's day-to-day activities. Financial markets have been going through a seismic shift with the rise of ESG investing. The pressing need to address climate change has led to the ascent of sustainable investing. This has also boosted the proliferation of ESG scores. Their different requirements and quality have added cost, confusion, risk, and complexity to investors. According to recent research, poor data quality is one of the biggest obstacles in ESG investing (Murray, 2021). This, in turn, has prompted concerns over "greenwashing" – i.e., that some investments are not as sustainable as they claim to be. Subsequently, this has negatively impacted the fight against global warming. In 2023, the European Commission sought responses from relevant regulators and has pressed for better disclosure (Jones, 2023). ESG scores are the most widely used metric. Yet, they have been scrutinised by regulators and investors because of their questionable quality. We argue that the limitations of the research methods used to generate ESG scores are one of the main barriers to responsible investing. We discuss this in section 2.1.

Artificial Intelligence (AI) techniques can drastically improve the accuracy of ESG scores and, thus, ESG investing by automatically detecting socioeconomic events and news items that influence ESG scores. Despite that, research in this domain is limited. To encourage it, we take the novel and difficult approach of creating a corpus with ESG relevance labels. To the best of our knowledge, there is no such publicly available corpus. Annotating news articles with ESG context and relevance to a company poses several challenges. ESG relevance detection is highly dependent on one's ESG domain knowledge and own perspective on what constitutes ESG relevance to a company. The lack of a universal ESG score framework further complicates this. In addition, categorising ESG content is highly contextual. Given these challenges, we argue that providing a comprehensive annotation schema is crucial to ensuring consistency and good performance of natural language processing (NLP) tasks. To promote standardisation, our schema was inspired by the EU taxonomy for sustainable initiatives and the United Nations Sustainable Development Goals (SDGs).[1] [2]

Alternative data, such as news articles, are well-documented as a powerful tool for evaluating stock market performance and investment opportunities. By building a corpus that consists entirely of publicly available news articles about FTSE 100 Index constituents, we demonstrate that this approach can also be used for assessing ESG credentials.[3] Specialists with industry and academic experience

[1] https://finance.ec.europa.eu/sustainable-finance/tools-and-standards/eu-taxonomy-sustainable-activities_en

[2] https://sdgs.un.org/goals

[3] https://www.londonstockexchange.com/indices/ftse-100

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 137–149
20 May, 2024.

in ESG, sustainable investing, economics, and finance manually annotated the ESG-FTSE corpus. Further, the annotators undertook training and followed precise guidelines to minimise bias and ensure the accuracy and consistency of the process. Baseline supervised and unsupervised learning experiments were conducted with the corpus. Results demonstrate high ESG salience and that NLP techniques can successfully be applied to the corpus for sustainability research. In addition, our experiments reveal that ESG has small data characteristics, i.e., scarce but relevant data. Despite that, experiments prove that it is feasible to make accurate ESG predictions even from small-volume data. The main contributions of this paper are the following:

- The first corpus with ESG relevance labels: the ESG-FTSE corpus. It consists of 3,913 news articles in English covering the period from late 2018 to summer 2021. To ensure the corpus is suitable for analysing a company's credentials, the news pieces are about the top ten FTSE 100 Index constituents by market capitalisation. By building the corpus entirely from publicly available news articles, we take the view that such alternative data influence not just financial performance but also a company's credentials, thus making it essential to ESG analysis.

- The first of its kind ESG annotation methodology. It consists of a three-level schema: a binary classification (relevant versus irrelevant news articles), ESG classification (environmental, social and governance-related news articles), and target company.

- We revealed a small data characteristic associated with ESG data, i.e. scarce but relevant data. This is an important characteristic of the ESG domain. Acute events, such as Covid-19 and extreme weather, are becoming more frequent. Thus, it is important to be able to utilise small-volume data to accurately predict such events. Our experiments prove that ESG-FTSE can be useful in making accurate ESG predictions.

2. Related Work

2.1. Notion of ESG Investing

ESG investing is an umbrella term for investments that seek positive returns and long-term impact on society, the environment, and the business. Environmental criteria may consider an organisation's pollution, waste, energy use, natural resource conservation, carbon footprint, and treatment of animals. The case of the miner BHP damaging aboriginal sites, which prompted an inquiry in the Australian parliament, is an example. Social criteria examine a company's management and its relationships with employees, customers, suppliers, and the communities where it operates. The case of wages and conditions of workers in the Leicester garment factory, which led to retailers reconsidering their purchasing policies, is an example. Governance looks at a company's leadership, executive pay, audits, internal controls, lawsuits, and shareholder rights. The case of increases paid to AstraZeneca investors, which sparked a rejection by shareholders, is an example. Previously, ESG investing represented a niche area of financial markets. With regulators and investors realising the financial materiality of ESG risks, these financial products have experienced soaring demand. According to a report by Morgan Stanley, sustainable funds' assets under management (AUM) totalled nearly $2.8 trillion in 2022. They are continuing to grow as a proportion of overall AUM: 7% compared to 4% five years ago 2023. Their popularity has led to investors seeking more information on sustainability risks. This has given rise to various initiatives to define ESG disclosure standards, investment and measurement principles, and metrics (Murray, 2021). ESG scores have become the most widely used metric to measure a company's credentials. While they are high in demand, the same cannot be said for their quality. The multitude of choices can explain the lack of consistency surrounding ESG scores regarding disclosure standards and measurement methodologies. The plethora of different ESG scores has left investors frustrated and confused with their competing measurement methodologies. In fact, the latter vary so wildly that organisations have been able to cherry-pick the most appealing providers (Murray, 2021; Li and Polychronopoulos, 2020). It also makes it difficult to compare one ESG score methodology with another. As seen in (Berg et al., 2022), correlations between ESG scores are, on average, 0.54 and range from 0.38 to 0.71. Sustainalytics and Vigeo Eiris, both major ESG score providers, have the highest level of agreement with each other, with a correlation of 0.71. The correlations of the environmental dimension are slightly lower than the overall correlations, with an average of 0.53. This leads to capital markets to not adequately pricing the ultimate costs surrounding sustainable businesses. In addition, the lack of a uniform measurement approach can cause reputational damage, financial loss and regulatory fines. Overall, the major ESG score providers generally follow similar processes for calculating their scores. They use traditional research methodologies.[4] These include manually gathering publicly available information, sending surveys to companies, and receiving issuer feedback on the scores given to them. Thus, producing ESG scores appears to be manual, time-consuming, and prone to human bias and omissions. We take as our premise the view that there are two main barriers to producing accurate ESG scores: the limitations of the research methods used and the lack of robust data in the process. We seek to address this challenge by providing a free reproducible corpus with ESG relevance scores, and an ESG annotation methodology.

2.2. Automated Text Classification of Financial Texts and its relevance to ESG

Traditionally, quantitative financial data have been essential to understanding an investment's sustainability potential. In recent years, alternative data, such as news articles and social media, have become more important for assessing investment opportunities and financial market performance because they capture corporate information outside the realm of traditional financial data. Studies have shown that such information, especially news articles, affects the value and performance of organisations. This, in turn, has boosted research in financial news analysis (Hagenau et al., 2012; Kalyanaraman et al., 2014; Luss and d'Aspremont, 2008; Shah et al., 2018; Zhao et al., 2020). Even though research in AI-based ESG scores and trends is still in its infancy, it has been gaining interest. Several studies have used company disclosures and social media text to extract ESG information (Mehra et al., 2022; Raman et al., 2020; Nematzadeh et al., 2019; Shahi et al., 2014; Hisano et al., 2020). Other studies generated ESG scores or utilised ESG criteria to inform investment decisions and assess their impact on a company performance (Y.Aiba et al., 2019; Sokolov et al., 2021; Ribano and Bonne, 2010; Napier, 2019; Khan, 2019; Ghoul et al., 2011; Krueger et al., 2020; Guo et al., 2020; Brown, 2015).

While analysing company narrative and social media text carry relevant ESG information, we believe such approaches have limitations regarding data robustness and objectivity. To illustrate, detecting ESG relevance from corporate disclosure and earning call reports hampers ESG score objectivity and accuracy by excluding other important information sources, such as news articles. In addition, not all companies produce ESG or Corporate Social Responsibility (CSR) reports or include such sections in their annual reports. Companies tend not to disclose negative sentiments about themselves either voluntarily. Another limitation of the relevant studies is the lack of generalisation of their models to out-of-corpora models. This paper addresses these limitations by creating an ESG news article corpus that applies to other corpora domains.

[4] Based on the lead author's experience in the financial sector and working with ESG score providers

2.3. Small data

Small data refers to an approach that requires less data but still offers useful insights. According to research by Gartner, 70% of organisations will shift their focus from big to small data by 2025 (Gartner, 2021). Small data seek to solve challenges stemming from scarce and disparate data, and historical data abruptly becoming obsolete and thus breaking AI models. To illustrate, breaking news can cause sudden changes in sentiment surrounding an organisation. In accordance with this, we argue that ESG-FTSE can be useful for obtaining relevant ESG insights. In line with recent work (Gururangan et al., 2020), we further demonstrate in our experiments that pretraining a model with a small corpus provides significant benefits: less computational resources and high accuracy in detecting ESG insights.

3. ESG-FTSE Corpus Development

This section describes the building, annotation process, and evaluation of the proposed corpus. Corpus development did not incur any costs and took ten days. Annotation took a week.

3.1. Approach Overview

The corpus construction process was divided into three phases: data collection, annotation, and evaluation. Each stage is described in detail in the following sections. In the first phase, we defined a set of criteria for data collection: category and data source. We selected News API as a retrieval method. We extracted news pieces about the top ten FTSE 100 Index companies to ensure suitability for financial market analysis. We extracted a total of 5,000 raw news articles. This was the initial, unlabelled version of the ESG-FTSE corpus. After this, we performed data cleaning on the initial corpus. In the second phase, we defined ESG relevancy criteria. We also introduced a three-level annotation schema. Due to the complexity of the annotation task, we defined a set of criteria for selecting annotators. Two annotators who met the requirements were selected. They were provided with training, clear guidelines and examples to minimise bias. We evaluated the ESG-FTSE corpus in the last phase and provided corpus statistics.

3.2. Data Collection

3.2.1. Category definition and category selection

The Financial Times Stock Exchange 100 Index (FTSE 100 Index) is a share index tracking the 100 biggest companies by market capitalisation that are listed on the London Stock Exchange (LSEG), which is said to be the most used UK stock market indicator by investors.[5] For data extraction and annotation schema purposes, we define as category the name of each of the ten biggest FTSE 100 index constituents by market capitalisation.[6] [7] The company name "AstraZeneca" is an example of a category. The goal was to ensure the corpus is relevant for analysing the credentials of stock market companies.

3.2.2. ESG Topics

This paper considers a news article with ESG relevance as a topic. An ESG topic can include any news piece related to environmental, social, or governance matters. Table 1 lists some factors under each ESG pillar. We follow the SDG guidelines and the technical screening criteria under the EU taxonomy for sustainable activities. As per Annex I of the above-mentioned EU regulation, the taxonomy is a classification system that determines sustainability criteria for "economic activities aligned with a net zero goal and the broader environmental goals other than climate". For a detailed description of the screening criteria and scope of the regulation, please refer to the EU taxonomy.[8]

3.2.3. Data source selection

News API is a REST API that returns JSON results for current and historic news articles.[9] We utilise it for news article retrieval for each category. Being an established information retrieval method, it proved more suitable for the purposes of this research because it overcame the limitations associated with other news APIs and data collection techniques, such as RSS feed and web scraping. Namely, it allows the collection of historic articles effortlessly. It also has a wide range of endpoints, including full content. Furthermore, it facilitates reproducibility because the developer subscription is free. It ensures data robustness and non-bias as it returns results from over 80,000 news publications. In addition, it solves data sparsity associated with ESG article collection.

3.2.4. News Article Extraction

The data extraction methodology is shown in Table 2. It is broken into three steps. The first step is collecting news articles for each category. The second step is corpus pre-processing. The third step is technical validation. In the first step, the same endpoints, language and time frame were used for each category. The aim was to achieve consistency. News articles were extracted in a csv format. The downloads were performed in ten batches over a ten-day period due to download limits associated with the free NewsAPI subscription. This produced ten csv files - one for each category. Two new columns, "Company Name" and "Number", were added to each file and filled with the corresponding category. For example, the column name of the AstraZeneca file was auto-completed with "AstraZeneca" in each row containing a news article. News articles in English were extracted between October 2018 and July 2021. To ensure data robustness, data were extracted by relevancy. A limit of 500 articles was set for each category. The following endpoints for each article were extracted via News API: title, author, source, description, content, publish date, and URL. In the data cleaning step, the following formatting changes were made to the data file to enhance understandability for future data use. Changes were made using Python. Duplicate news articles were removed. After examining the new corpus, more duplicates were noticed. Different news publications reusing the exact text caused some news pieces to be treated as unique by Python. Thus, another duplicate removal exercise was conducted. There is a word limit for csv files. To avoid losing article content and for consistency, a limit of 4,800 words per news article was applied. Last, the "content" endpoint was renamed to "Text". New columns were added: "Relevance Label" and "Primary Label". In the last step, news articles were checked for personally identifiable information, particularly e-mail addresses. This was done by searching for symbols and domains commonly used in e-mail addresses, i.e., "@" or ".com,". Author names were removed. For the purposes of this study, only the "Text", "Number", and "Label" columns were kept in the final corpus. All other columns were removed. After iterating over each category, 5,000 raw news articles were obtained. After removing duplicate articles, the final corpus consisted of 3,913 articles.

3.3. Annotation Process

Since this paper focuses on producing a corpus with ESG relevance labels, the paper deems rele-

[5] https://www.londonstockexchange.com/

[6] As of 30 June 2021

[7] The top 11 constituents were taken due to Royal Dutch Shell being listed twice via different legal entities

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021R2139

[9] https://newsapi.org/

	Factors
Environ-mental	Greenhouse gas emissions, ground and air pollution, energy usage, carbon footprint, waste and water management, land use, biodiversity loss
Social	Labour practices, fair pay, equal employment opportunities, labour laws, workplace health and safety, responsible supply chain, community engagement, product quality, safety and access
Governance	Shareholder rights, board diversity, executive compensation, corporate governance, compliance, risk management, conflict of interest, corruption, accounting integrity

Table 1: ESG pillars: key factors. The table is not exhaustive.

Data	Numerical or Text Value
Category	AstraZeneca, Unilever, Diageo, HSBC, GlaxoSmithKline, Rio Tinto, BP, British American Tobacco, Royal Dutch Shell, BHP
News Articles	500 per category
Time Period	30/10/2018 – 31/07/2021
Data Source	News API
Language	English
Raw End Points	title, author, source, description, content, publish date, URL
Final End Points	content - renamed to Text
Word Limit	4,800 words per news article
Added Columns	Index, Company Name, Relevance Label, Primary Label

Table 2: Data Collection methodology

vant news articles that include both a category and at least one ESG topic. Table 3 shows examples of three ESG topics.

3.3.1. Annotator Selection

In the first step of the annotation process, we introduced a set of requirements for selecting annotators. This was necessary due to the multi-faceted nature of the annotation task. We decided to select experts recognised by industry and academia for their contribution to the ESG field. In addition, the specialists had to possess industry and academic experience in the following domains: ESG, sustainable investing, economics and finance.

Category	Text	ESG topic
BP	BHP's oil exit would be better sooner than later	Environmental
BHP	Strike at BHP's Chile copper mines continue	Social
Rio Tinto	Rio Tinto appoints three women as non-executive directors	Governance

Table 3: Examples of categories and ESG topics. The text column consists of news headlines.

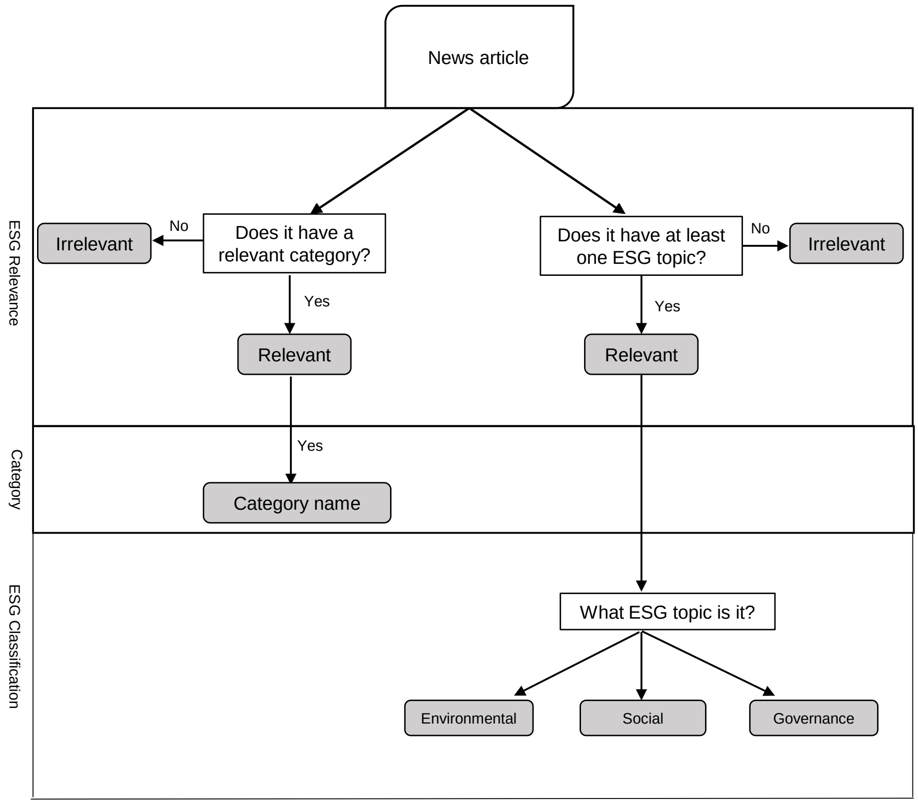
3.3.2. Annotation Schema

Schemas from other domains and the SDGs inspired the annotation methodology.(Zampieri et al., 2019; Lee et al., 2022) Our methodology comprises a three-level schema: a binary classification (relevant versus irrelevant news articles), ESG classification (environmental, social, and governance-related news articles), and target company. Figure 1 outlines the annotation schema. Annotation was done manually. This method was adopted because it is considered the most precise method for document annotations. Two annotators who met the selection criteria were recruited. They conducted the labelling independently and according to two different levels of classification. The first layer is a binary classification: relevant versus irrelevant news articles. These are denoted by "1" and "0" respectively. Relevant news pieces must contain one category and at least one ESG topic. The second layer comprises an ESG classification: environmental, social, and governance-related news articles. These are represented in the "Primary Label" column as "E", "S", and "G", respectively. The ESG topic criteria are described in more detail in the ESG Topic section. It is to be noted that some articles may contain multiple ESG topics. We only classify the dominant topic, i.e., the primary topic. The third layer, the company name, was added during corpus preprocessing.

3.4. Annotation Evaluation

We computed inter-annotator agreement using Cohen's kappa (McHugh, 2012). We produced a Kappa score for both levels of the annotation schema. The Kappa scores show that high inter-annotator agreement was reached for both binary and ESG classification: 0.97 and 0.94, respectively. The small number of codes for each clas-

Figure 1: Annotation schema



sification task should be noted, as well as the high level of expertise of both annotators. We believe that it explains the substantial results.

4. ESG-FTSE corpus statistics

This section presents the ESG-FTSE corpus statistics. The corpus consists of 3,913 document-level annotations. First, a binary classification was performed. Each of the 3,913 news articles received either a ”Relevant” or ”Irrelevant” label: 1,178 and 2,735 news pieces, respectively. Next, the news articles with a ”Relevant” label were also annotated according to an ESG-level classification. 418 news articles were classified as ”Environmental”, 218 news articles as ”Social”, and 542 news articles received a ”Governance” label.

5. Experiments and Discussion

To validate the suitability of the proposed expert annotated corpus for ESG relevance detection, we implemented baseline experiments using both supervised and unsupervised learning methods for text classification. They showed that ESG-FTSE can successfully be used in different types of experiments to derive ESG insights from text. A second objective of the experiments was to describe the corpus and evaluate its quality by applying different tasks, representations, and machine-learning methods. A detailed description of the experiment and a discussion of the results are presented in the following sections.

5.1. Supervised Learning: Text Classification

We performed four supervised learning experiments: three for ESG relevance detection and one for ESG classification.

5.1.1. ESG Relevance Detection

To detect ESG relevancy, we undertook three binary classification experiments. Due to the imbalanced nature of the corpus, a stratified K-Fold cross-validation was implemented in all experiments. It was essential to ensure the data were split randomly while maintaining the same class distribution in each subset. We determined that 5 splits with a class ratio of approximately 0.30 were most suitable. A different representation for each experiment was adopted to decide whether or not it would improve model performance. In general, SVM classifiers produce highly accurate results for binary classification problems (Schölkopf and Smola, 2018). Thus, an SVM classifier with

Table 4: ESG detection: binary classification. Results are rounded to two decimal places

Experiment	Feature set	F1 Score	Accuracy	Precision	Recall
Experiment 1	TF-IDF	79.09	88.62	86.33	72.96
Experiment 2	TF-IDF, n-grams	72.96	85.09	82.43	65.45
Experiment 3	TF-IDF, uni-grams	78.40	87.85	83.33	74.03

a linear kernel and default parameters was chosen as a machine-learning method for all experiments. In addition, we used the pandas, sklearn, nltk and matplotlib packages in our experiments. Accuracy, precision, recall and F1 score with default parameters were adopted as evaluation metrics. Experiment 1 sliced the data in 67% train and 33% test. We used TF-IDF as representation (Sammut and Webb, 1970). Experiment 2 adopted a default training/validation split: 75% and 25% accordingly. In addition, we used *n-grams* and TF-IDF for feature extraction at different levels: word, n-grams (from bi- to four-grams), and character. Experiment 3 was an extension of Experiment 1, with *uni-grams* added to the pre-processing step. Table 4 presents the results of all experiments.

5.1.2. ESG classification

In this experiment, we implemented a 5-fold stratified BERT model to classify Environmental, Social, and Governance labels. The smaller, pre-trained *bert-base-uncased* model was used. PyTorch, tqdm, BertTokenizer, pandas, sklearn and NumPy packages were adopted in the task. Data were split into 15% validation and 85% training sets. We used RandomSampler for training and SequentialSampler for validation. The training was conducted in 5 epochs. Due to class imbalance, weighted evaluation metrics were utilised: F1 score, accuracy, precision and recall. Results are displayed in Table 5

Label	F1	Accuracy	Precision	Recall
E	0.92	0.86	1.00	0.86
S	0.65	0.48	1.00	0.49
G	0.60	0.43	1.00	0.43
I	0.97	0.95	1.00	0.95

Table 5: ESG classification results. Labels: E = Environmental, S = Social, G = Governance, I = Irrelevant. All scores are weighted average.

5.2. Unsupervised learning: Topic Modelling

Topic modelling is an unsupervised probabilistic algorithm that considers the problem of modelling discrete data, such as text corpora. The goal is to discover the main topics that occur in a set of documents by reducing their dimensionality. In our experiments, topic modelling was performed by building a latent Dirichlet allocation (LDA) model (Blei et al., 2003). Gensim, Nltk and Spacy packages were utilised. Hyperparameter tuning was performed in a series of hyperparameter sensitivity tests to improve the default LDA model's accuracy. The following hyperpatameters were tuned: *filter_extremes, random_state, update_every, chunksize, passes, alpha, eta and per_word topics.* The most optimal hyperparameter values were chosen based on the highest coherence score achieved on the LDA model. It achieved a coherence score of 0.60.The sensitivity tests are shown in Table 6. The most optimal hyperparameter values were chosen based on the highest coherence score achieved on the LDA model (Table 6, Test 3). Number of topics (*k*) is one of the most important LDA model inputs. Extracting the right number of topics largely depends on the dataset characteristics. Seven LDA models were built and compared by their coherence values to determine the most optimal *k* number of topics. A limit of forty LDA models was set. The optimal number of topics was chosen based on the highest coherence value achieved on the final LDA model. *K=20* was selected because it achieved the highest coherence score of 0.60. More information is available in Appendix A. The evaluation was conducted via an intrinsic metric (coherence score C_v). The four most dominant topics are visualised via t-SNE in Figure 2. Entity salience was visualised via an interactive LDA model Intertopic Distance Map created via the pyLDAvis package. A snippet of it is available in Appendix B. In the interactive LDA model Intertopic Distance map, each bubble on the plot represents a topic. The larger the bubble, the more prevalent the topic. Additionally, a bar chart representing the top 30 most salient keywords that form a selected topic is available in the interactive LDA model Intertopic Dis-

Table 6: LDA model sensitivity tests

Hyperparameters					
Test	**Alpha**	**Beta**	**filter_extremes**	**Perplexity**	**Coherence**
1	0.01	0.9	no_below=10 no_above=0.20	-7.597	0.538
2	0.01	0.5	no_below=10 no_above=0.15	- 7.670	0.565
3	0.03	0.5	no_below=10 no_above=0.15	-7.652	0.597
4	0.05	0.5	no_below=10 no_above=0.15	-7.673	0.552
5	0.04	0.5	no_below=10 no_above=0.15	- 7.665	0.572
6	0.04	0.4	no_below=10 no_above=0.15	- 7.668	0.551

tance Map. It reveals high ESG salience in each topic.

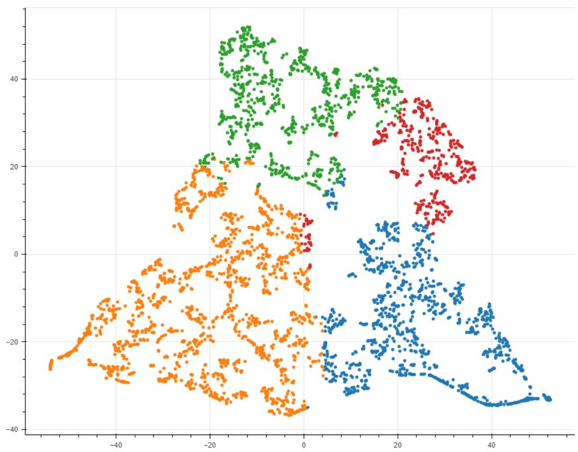


Figure 2: t-SNE clustering of the top 4 topics

5.3. Discussion

We evaluated two different classification tasks: ESG relevance detection and ESG classification. Results demonstrated that the ESG-FTSE corpus is highly quality and can be used successfully for ESG knowledge extraction in supervised and unsupervised models. In addition, the experiments proved that valuable ESG insights can be obtained even from low-volume data. Last, the topic modelling experiment provided a thorough context and description of the corpus. We implemented three baseline experiments for ESG relevance detection. All of them achieved high performance. The best model for this task, Experiment 1, obtained a 79% F1 score. For ESG detection, we performed a stratified 5-fold BERT experiment. As shown Table 5, high performance was also achieved for this task. To illustrate, adequate F1 scores were obtained for each label. We take at our premise the view that this suggests a balanced model. The unsupervised experiment also produced high results. Probabilistic topic models like LDA always produce topic outputs. However, making them valuable and meaningful for this research demanded capturing the correct information, i.e., the minority ESG class. Despite the class imbalance of the corpus, the topic modelling experiment successfully extracted relevant ESG information. The t-SNE plot of the four most dominant topics indicates that the most similar documents are grouped in well-defined clusters (Figure 2). The most salient words also demonstrate the robustness of the model for a given topic. The results validated our initial view that a well-defined annotation schema yields a good model performance in complex and subjective domains like ESG.

6. Conclusion

The most significant contribution of this study is presenting a free, reproducible corpus to facilitate data sharing in a standardised framework. Being the first corpus with ESG relevance, ESG-FTSE provides the first-of-its-kind annotation schema. In addition, it offers a novel solution to the bias associated with ESG scores. The present study revealed class imbalance due to data sparsity. Instead of trimming the corpus or boosting the minority class, we trained on all possible instances to maximise coverage. The evidence confirmed that small data can be insightful in obtaining relevant ESG insights.

7. Ethical Considerations and Limitations

A developer News API license was obtained to download the news articles. Data were downloaded and used following News API Terms. [10] According to the provider, all data are publicly available. No personal data, such as user analytics or cookies, were used in this study. News API is compliant with UK and EU data laws and directives. To illustrate, its privacy policy states that News API "has been prepared to fulfil the obligations under Art. 10 of EC Directive n. 95/46/EC, and under the provisions of Directive 2002/58/EC, as revised by Directive 2009/136/EC, on the subject of Cookies." [11]. Thus, all data used for this study is considered ethical and lawful.

8. Bibliographical References

Florian Berg, Julian F Koelbel, and Roberto Rigobon. 2022. Aggregate confusion: The divergence of esg ratings.

D. Blei, A. Ng, and M. Jordan. 2003. Latent dirichlet allocation - journal of machine learning research.

M. Brown. 2015. Managing nature–business as usual: Resource extraction companies and their representations of natural landscapes. *Sustainability*, 7(12):15900–15922.

Gartner. 2021. Gartner says 70% of organizations will shift their focus from big.

S. Ghoul, O. Guedhami, C. Kwok, and D. Mishra. 2011. Does corporate social responsibility affect the cost of capital? *Journal of Banking & Finance*, 35(9):2388–2406.

T. Guo, N. Jamet, V. Betrix, L.A Piquet, and E. Hauptmann. 2020. Esg2risk: A deep learning framework from esg news to stock volatility prediction.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks.

M. Hagenau, M. Liebmann, M. Hedwig, and D. Neumann. 2012. Automated news reading: Stock price prediction based on financial news using context-specific features. *Decision Support Systems*, 55:1040–1049.

[10] https://newsapi.org/terms
[11] https://newsapi.org/privacy

R. Hisano, D. Sornette, and T. Mizuno. 2020. Prediction of esg compliance using a heterogeneous information network. *Journal of Big Data*, 7(1).

H. Jones. 2023. Eu watchdogs see greenwashing across the bloc's financial sector.

V. Kalyanaraman, S. Kazi, R. Tondulkar, and S. Oswal. 2014. Sentiment analysis on news articles for stocks. In *2014 8th Asia Modelling Symposium*, pages 10–15.

M. Khan. 2019. Corporate governance, esg, and stock returns around the world. *Financial Analysts Journal*, 75(4):103–123.

P. Krueger, Z. Sautner, and L. Starks. 2020. The importance of climate risks for institutional investors. *The Review of Financial Studies*, 33(3):1067–1111.

M. Lee, L. Soon, E. Siew, and L. Sugianto. 2022. Crudeoilnews: An annotated crude oil news corpus for event extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022*.

Feifei Li and Ari Polychronopoulos. 2020. What a difference an esg ratings provider makes!

R. Luss and A. d'Aspremont. 2008. Predicting abnormal returns from news using text classification. *Quantitative Finance*, 15.

M. McHugh. 2012. Interrater reliability: The kappa statistic.

S. Mehra, R. Louka, and Y. Zhang. 2022. ESGBERT: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications*. Academy and Industry Research Collaboration Center (AIRCC).

S. Murray. 2021. Measuring what matters: The scramble to set standards for sustainable business.

E. Napier. 2019. Technology enabled social responsibility projects and an empirical test of csr's impact on firm performance.

A. Nematzadeh, G. Bang, X. Liu, and Z. Ma. 2019. Empirical study on detecting controversy in social media. *CoRR*, abs/1909.01093.

N. Raman, G. Bang, and A. Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models.

J. Ribano and G. Bonne. 2010. A new quality factor: Finding alpha with asset4 esg data - thomson reuters.

C. Sammut and G. Webb. 1970. Tf–idf.

B. Schölkopf and A. Smola. 2018. Learning with kernels: Support vector machines, regularization, optimization, and beyond.

D. Shah, H. Isah, and F. Zulkernine. 2018. Predicting the effects of news sentiments on the stock market. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 4705–4708.

A. Shahi, B. Issac, and J. Modapothala. 2014. Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications*, 13(01):1450006.

A. Sokolov, Mostovoy, J. Ding, and L. Seco. 2021. Building machine learning systems for automated esg scoring.

Morgan Stanley. 2023. Sustainable funds: Returns & demand.

Y.Aiba, T. Ito., and Y. Ibe. 2019. Network structure in esg ratings suggests new corporate strategies: evolving ai technology to quantify qualitative data. *Securities Analysts Journal (SAAJ)*.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019. Predicting the type and target of offensive posts in social media. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

W. Zhao, G. Zhang, G. Yuan, J. Liu, H. Shan, and S. Zhang. 2020. The study on the text classification for financial news based on partial information. *IEEE Access*, PP:1–1.

9. Appendix

A. Topic Modelling

```
[ ]
# Show graph
limit=40; start=2; step=6;
x = range(start, limit, step)
plt.plot(x, coherence_values)
plt.xlabel("Num Topics")
plt.ylabel("Coherence score")
plt.legend(("coherence_values"), loc='best')
plt.show()
```

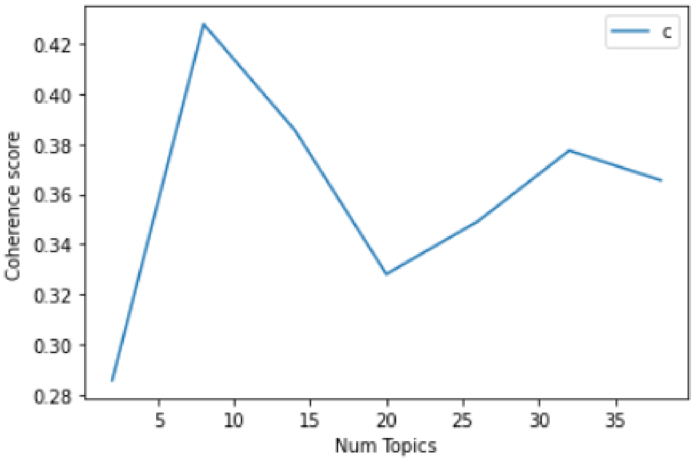


Figure 3: Number of topics.

B. pyLDAvis visualisation tool

Figure 4: Intertopic Distance Map (via multidimensional scaling). Each bubble on the plot represents a topic. The larger the bubble, the more prevalent the topic.

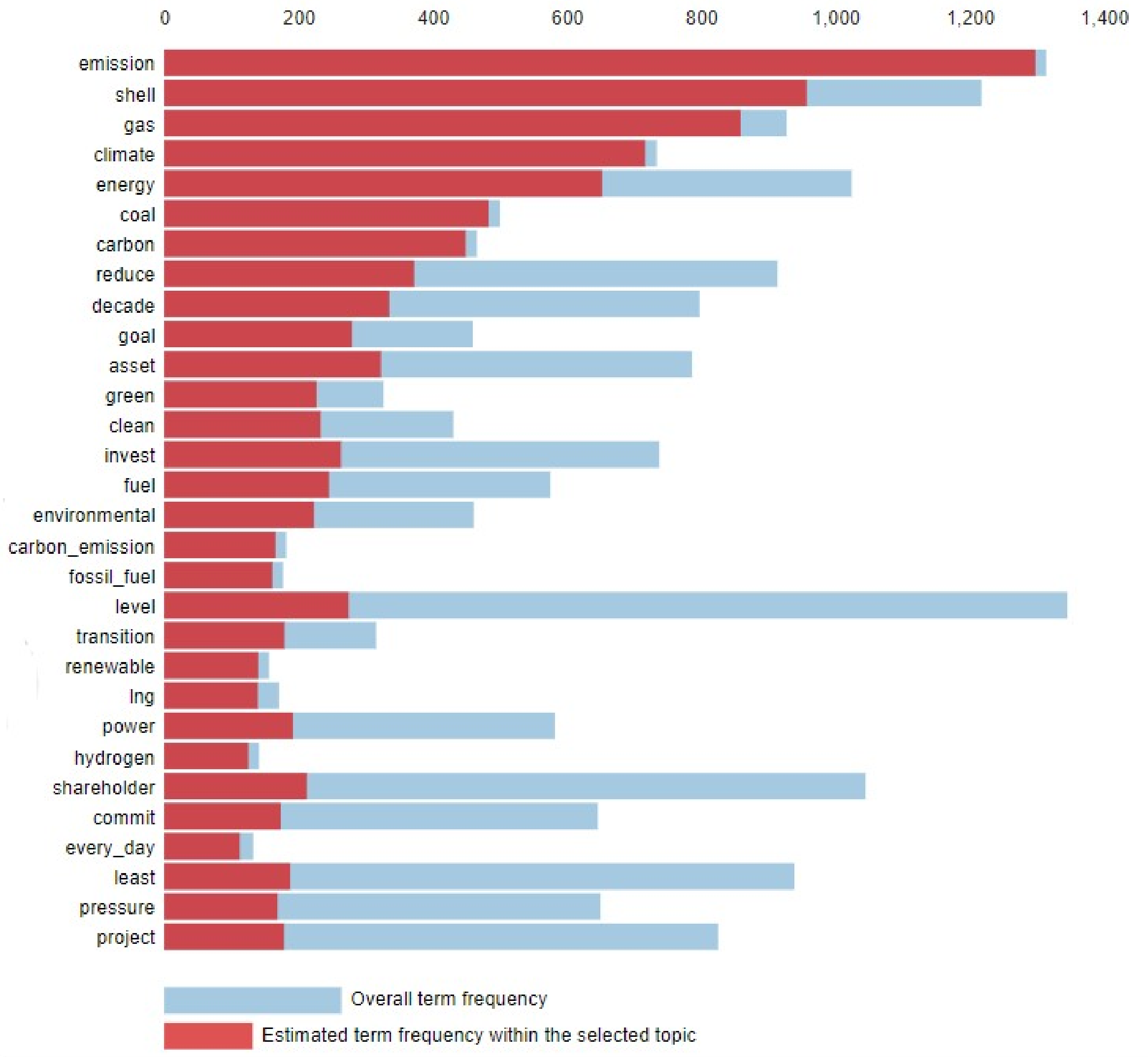


Figure 5: Salient words in an example topic. It has high ESG salience.

BBRC: Brazilian Banking Regulation Corpora

Rafael Faria de Azevedo, Thiago Henrique Eduardo Muniz,
Claudio Pimentel, Guilherme José de Assis Foureaux,
Bárbara Caldeira Macedo, Daniel de Lima Vasconcelos
Banco do Brasil S.A
SAUN Quadra 5, Lote B, s/n, 70.040-912, Asa Norte, Brasília/DF, Brasil
rafael.f.azevedo@outlook.com
{rafael.azevedo, thiagobancodobrasil, clpimentel,
guilhermefx, barbaracaldeira, delima}@bb.com.br

Abstract

We present BBRC, a collection of 25 corpus of banking regulatory risk from different departments of Banco do Brasil (BB). These are individual corpus about investments, insurance, human resources, security, technology, treasury, loans, accounting, fraud, credit cards, payment methods, agribusiness, risks, etc. They were annotated in binary form by experts indicating whether each regulatory document contains regulatory risk that may require changes to products, processes, services, and channels of a bank department or not. The corpora in Portuguese contain documents from 26 Brazilian regulatory authorities in the financial sector. In total, there are 61,650 annotated documents, mostly between half and three pages long. The corpora belong to a Natural Language Processing (NLP) application that has been in production since 2020. In this work, we also performed binary classification benchmarks with some of the corpus. Experiments were carried out with different sampling techniques and in one of them we sought to solve an intraclass imbalance problem present in each corpus of the corpora. For the benchmarks, we used the following classifiers: Multinomial Naive Bayes, Random Forest, SVM, XGBoost, and BERTimbau (a version of BERT for Portuguese). The BBRC can be downloaded through a link in the article.

Keywords: banking, corpus, regulatory risk

1. Introduction

Regulation is part of business activities in any industry, including the financial sector. Considering the sheer volume of regulations companies must follow, it is often a manual and onerous process. However, regulation is beneficial to the bank's customers, the market, and even the company itself, as it can even bring profits to the bank (Pasiouras et al. 2009; Aldasoro et al. 2020; Kim et al. 2013). To manage and automate the regulation it must respond to, Banco do Brasil created a tool to manage the daily publications that financial market regulatory authorities make, which can impact the company's activities. This tool is called Radar Regulatório (Regulatory Radar), which acronym is RR. It has been in production since 2020, classifying regulatory documents as relevant or irrelevant to several of its departments individually, with regard to their potential to impact its activities from a regulatory risk perspective. If a document is classified as relevant, it is forwarded to each department that the publication may impact. Therefore, experts in the area can evaluate the document and make the necessary changes to keep the department in compliance with the regulatory authority that published the document.

Radar Regulatório (RR) serves more than 40 company departments. It classifies between 300 and 1,000 regulatory documents daily published by more than 100 regulatory authorities (municipal, state, and federal levels). It is important to follow city and state regulations, in addition to federal ones, as their needs vary depending on its characteristics (Lastra, 2019), especially in a country as large as Brazil. The application works with a hybrid approach with a pipeline composed of Machine Learning (ML) and rules (regular expressions - regex).

When an expert from a department analyzes a document classified by the tool, he points out the correctness or otherwise of the labeling of the document made by RR, and thus annotates the document that will be part of the department's corpus to be used in retraining of its Artificial Intelligence (AI) model. Each department has created its regulatory risk corpus according to this annotation process. The junction of each corpus of many departments gave birth to the Brazilian Banking Regulation Corpora (BBRC), which is **the main contribution of this work**.

The BBRC is a set of 25 regulatory risk corpus (legal/financial data) from different departments of Banco do Brasil. Furthermore, the corpora contain documents from 26 different Brazilian banking/finance regulatory authorities, which can affect the bank's various activities (products, processes, services, and channels) of the bank. The corpora belongs to various departments such as insurance, investments, treasury, accounting, agribusi-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 150–166

ness, human resources, and others. If on the one hand, BBRC can be useful to explore ML algorithms applied to NLP tasks such as text classification, document analysis, and sentiment analysis, on the other hand, each corpus of BBRC can be used in other areas like sociology, economy and politics, as highlighted in the Section 2 and Section 7.

Our second contribution is a benchmark that compares some models we evaluated for a binary classification task. In one of the benchmarks, we evaluate a strategy to deal with the intraclass imbalance problem present in the entire BBRC corpus (Liu et al., 2021).

The rest of this paper is organized as follows. In Section 2, we introduce related works. In Section 3, we present our corpora. The application that caused the creation of BBRC is presented in Section 4. In Section 5, the experiments performed with some corpus of BBRC and the discussion are presented. Section 6 presents our future work and Section 7 concludes the paper.

2. Related Works

In this section, we mainly present works related to the BBRC, but also some related to Radar Regulatório (Regulatory Radar). We start by presenting some corpus similar to BBRC.

Lima et al. (2020) used machine learning to investigate fraud in the Brazilian public sector. They used a dataset constructed with a source that is also present in the BBRC, which is the Brazilian Official Journal (Diário Oficial da União - DOU). The dataset contains 1,907 annotated risk entries.

Sohn et al. (2021) presented the Global Banking Standards QA dataset (GBS-QA), a banking regulation dataset of questions from market players and answers from the Basel Committee on Banking Supervision (BCBS). The corpus was reorganized and verified by financial regulatory experts. In our search, few banking regulatory corpus were found; however, when we searched for financial corpus, the quantity of corpus increased.

Jiang et al. (2020) introduced an automatic financial news dataset annotation through a weakly-supervised hierarchical multilabel classification for the Chinese language. The event FinCausal 2020 Shared Task on Causality Detection in Financial Documents created the FinCausal Corpus (financial news feed) (Mariko et al., 2020). Lefever and Hoste (2016) presented a supervised machine learning approach to economic events detection in newswire text. To do so, a corpus of Dutch financial news articles with ten types of company-specific economic events was annotated. The work of Zmandar et al. (2022) presented CoFiF Plus, a narrative summarization dataset created from financial reports in French. It is made up of 1,703 reports covering a time period of 1995 to 2021. Jabbari et al. (2020) described an ontology of compliance-related concepts and relationships (annotation schema). They also presented an annotated corpus of financial news articles in French for entity recognition and relation extraction. Chen et al. (2021) introduced FINQA, an expert-annotated dataset containing 8,281 financial QA pairs, along with their numerical reasoning processes. It was built based on the earnings reports of S&P 500 companies. The dataset was tested with algorithms such as BERT, RoBERTa and FinBERT. DoRe is a French and dialectal French corpus for NLP analytics in finance, regulation, and investment. It is composed of 2,350 Annual Reports from 336 companies, covering a time frame from 2009 to 2019 (Masson and Paroubek, 2020). In addition to financial corpus, we also found legal corpus in our research, which may be related to regulation or the financial sector.

The area of NLP has long studied legal texts, as well as texts from the health sector and other areas, whether in Portuguese or other languages. Just as in the area of health, law also has a wealth of specific terms (Thompson et al. 2011; Halder et al. 2017; Quochi et al. 2008; Pardelli et al. 2012; Delfino et al. 2018), the same phenomenon occurs in the area of banking regulation.

De Araujo et al. (2020) described Victor, a dataset of digitized legal documents from the Brazilian Supreme Court. The corpus supports two tasks, document classification and theme assignment. LexGLUE is a benchmark dataset to evaluate the performance of NLP methods, especially Large Language Models (LLMs). It is based on seven existing legal NLP datasets in English (Chalkidis et al., 2021). Au et al. (2022) describe E-NER, a publicly available NER dataset. It is based on legal company filings available from the EDGAR dataset of the US Securities and Exchange Commission. Chalkidis et al. (2023) presented LeXFiles, a diverse multinational English legal corpus that includes 11 distinct subcorpora that cover legislation and case law from six primarily English-speaking legal systems (EU, CoE, Canada, US, UK, and India). The work also introduces LegalLAMA, a new probing benchmark suite inspired by LAnguage Model Analysis (LAMA). The LexFiles are compared to the Pile of Law corpus (Henderson et al., 2022), a large legal corpus (256GB dataset).

In addition to the datasets, we also tried to find applications similar to RR. We found the use of multiple classifiers to detect investment rules in long regulatory documents (Mansar and Ferradans, 2018), a Python library for NLP and machine learning for legal and regulatory texts (Bommarito et al., 2018), a semi-supervised text classification framework for operational risk (Zhou et al., 2020) and approaches with LLMs (Mamakas et al. 2022; Chakravarthy et al. 2020). In the next section, our corpora is presented in detail.

3. The Corpora

This section presents the main subject of this paper, the Brazilian Banking Regulation Corpora (BBRC). The corpora annotation started with the creation of the application Radar Regulatório (Regulatory Radar - RR), which is presented in Section 4. A corpora is a collection of corpus. In Natural Language Processing (NLP), a dataset is called corpus. Each corpus in BBRC belongs to a department of Banco do Brasil and was annotated in a binary way: relevant or irrelevant (in reality, the bank's experts annotate the corpora with scores from 0 to 3, where 0 is irrelevant, 1 is not very relevant, 2 is relevant and 3 is extremely relevant, however, the company decided to share the data in binary format). A corpus, in the context of RR, is a collection of documents from various **regulatory authorities** that could affect the activities of a department. Each document was annotated as belonging to the relevant class or to the irrelevant class. If the document was classified as relevant, it means that the department may have to make changes to its activities to comply with the relevant regulatory document published by the regulatory authority. If the document is classified as irrelevant, no change is needed. All corpora documents are public, as documents published by all regulatory authorities mentioned in this article are valid for all Brazilian banks and financial institutions.

The annotation process for each corpus of each department has been performed by one or more experts from that department since 2020. These experts are responsible for ensuring that the changes demanded in a relevant document are met as required by the regulatory authority. For this reason, the corpora did not pass through an evaluation of the agreement between annotators (inter-annotator agreement). When possible, it is a process that produces a more reliable corpus, where each sample is annotated by different annotators, who follow a rigorous process that helps the annotators make decisions guided by a well-developed guideline (this guideline could not be shared by the company). The agreement between annotators evaluation potentially improves the quality of the corpus and, consequently, the quality of the trained model increases, which can be highly affected by corpus quality, as presented by (Alhamzeh et al. 2022; Artstein 2017; Nowak and Rüger 2010).

However, the quality of the annotation of BBRC is ensured by the consequences that can occur if a mistake is made. Failure can cause expensive fines, restrictions, and sanctions to the bank. No expert wants to live in a situation like this. The correct classification of a regulatory document is the first step that decides whether an action plan must be carried out and executed to change a product, a process, a channel, or a service to keep it in compliance. The BBRC data ranges from June 18, 2020 to August 16, 2023.

The regulatory authorities (regulators) belong to one of these three levels of compliance: federal, state, or municipal (as mentioned by Lima et al. 2020). Examples of regulators are the Brazilian Central Bank (Banco Central do Brasil - BACEN), Brazil's federal revenue (Receita Federal do Brasil - RFB), Legislative Assembly of the State of Mato Grosso (Assembleia Legislativa do Estado do Mato Grosso) and Rio de Janeiro City Council (Câmara Municipal do Rio de Janeiro). Table 1 presents numbers about regulators in BBRC.

From the perspective of all regulatory authorities, BBRC has in total 5,698 unique documents in the relevant class, 20,131 unique documents in the irrelevant class, and 25,829 unique documents considering both classes. To be part of the corpora, each regulator had to have at least five documents classified in the relevant class. Regarding the departments, only those with at least 50 documents of the relevant class were elected to the corpora. The description of all 25 departments (corpus) of the bank in the corpora is given in Table 8, which is in the Appendix A Section at the end of the paper, after the references.

Table 2 presents the description of each column of the corpora. The idea was to offer a wider understanding of the details of the corpora. The corpora[1] is shared with the community in a CSV format file (1.7 GB). Figure 1 presents BBRC data schema. Figures 3, 4, 5, and 6 in the Appendix A present examples of the content of each column of BBRC. Figure 7, also at Appendix A section, presents the text of one sample of the BBRC. Table 3 shows information on the number of samples per class of each corpus in the BBRC.

In total, the corpora has 61,650 document samples, 7,823 in the relevant class, and 53,827 in the other class. The documents are unique in each class and in each corpus, but can be repeated in different corpus. This repetition of documents happens because one document can be relevant or irrelevant for several departments. The most important feature (column) in the BBRC is **text** (as it is an NLP dataset collection).

Table 4 presents the basic statistics of the column text in the relevant class. Character information

[1] Data available at `https://huggingface.co/datasets/bancodobrasil/bbrc_brazilian_banking_regulation_corpora`

Regulatory authority	Relevant	Irrelevant	Total
National Civil Aviation Agency (ANAC)	14	507	521
Brazilian Association of Financial and Capital Market Entities (ANBIMA)	262	381	643
National Data Protection Authority (ANPD)	7	22	29
National Supplementary Health Agency (ANS)	85	346	431
Legislative Assembly of the State of Mato Grosso	9	66	75
Brazil, Stock Exchange, Counter (B3)	886	1,241	2,127
Brazilian Central Bank (BACEN)	1,796	3,455	5,251
Commodities and Futures Exchange & São Paulo Stock Exchange (BM&F BOVESPA)	13	39	52
National Bank for Economic and Social Development (BNDES)	176	202	378
Rio de Janeiro City Council	11	843	854
Securities Custody and Financial Settlement Center (CETIP) (currently B3)	141	47	188
Federal Accounting Council (CFC)	32	99	131
Interbank Payments Chamber (CIP)	266	640	906
Financial Activities Control Board (COAF)	40	42	82
Accounting Pronouncements Committee (CPC)	18	4	22
Securities and Exchange Commission (CVM)	380	1,026	1,406
Brazilian Official Journal (DOU)	534	7,203	7,737
Brazilian Federation of Banks (FEBRABAN)	8	0	8
National Institute of Information Technology (ITI)	23	56	79
Ministry of Labour	11	44	55
Núclea (previous CIP)	39	62	101
Presidency of the Republic (PR)	176	844	1,020
National Supplementary Pension Superintendence (PREVIC)	22	79	101
Brazil's Federal Revenue (RFB)	370	1,117	1,487
National Treasury Secretariat (STN)	227	1,003	1,230
Private Insurance Superintendency (SUSEP)	152	763	915
Total	**5,698**	**20,131**	**25,829**

Table 1: Column "Relevant" presents unique documents in the relevant class for each regulator. The column "Irrelevant" presents unique documents in the irrelevant class for each regulator. The column "Total" shows the unique documents in the whole corpora for each regulator. The URLs of all regulators are in the Appendix A, Table 9. The name of the regulatory authority was translated, but the acronym was kept in Portuguese.

can give an idea of the length of documents. Assuming that a Microsoft Word page holds around 3,600 characters (Arial 11), the median (middle quartile) of a text in the relevant class is longer than a page. So, 50% of the documents in the relevant class are at least one full page long. Similar statistics occur in the irrelevant class. To count the **words** and **unique words** of each document, a function was used to separate words between blank spaces. All texts were analyzed in their original state (without preprocessing or cleaning), and noise such as URLs, HTML, and email addresses could have caused the incorrect number of words in the text. However, the results still give a fairly precise idea of the size of the document.

	Column	Description
1	class	The class of the document is 1 to relevant or 0 to irrelevant
2	department	The department (board, directorate or related company) of Banco do Brasil that uses RR with a corpus
3	entry_date	The date the document was received by RR from the contracted company
4	general_id	The document's unique identifier across the entire corpora
5	normative_identifier	Identifier of the regulatory document given by the regulatory authority
6	publication_date	The date the regulatory document was published
7	regulatory_authority	The regulatory authority (regulator) that published the regulatory document
8	subject	Most regulatory documents usually have a subject, such as a title or summary
9	subject_length	The number of characters in the subject
10	subject_unique_words	The number of unique words in the subject
11	subject_words	The number of words in the subject
12	text	The full text of the regulatory document
13	text_length	The number of characters in the text
14	text_unique_words	The number of unique words in the text
15	text_words	The number of words in the text
16	type	The type of the regulatory document, most regulatory authorities publish several types of documents
17	unique_document_id	Unique identifier of the regulatory document in the corpus (is repeated in different corpus)

Table 2: BBRC columns description.

The main contribution of this work is BBRC, which is fundamental to Radar Regulatório (Regulatory Radar). This application is presented in the next section.

4. The Application

Before Radar Regulatório (Regulatory Radar - RR), the entire regulatory risk process was done manually, without a formal process, and without standards, where most departments acted in isolation. For example, department A could have one regulatory risk expert who searched and read all regulatory documents published every day to evaluate whether a new norm or a new law could impact the businesses of department A. On the other hand, department B could have a team of three experts that checked once a month for possible

```
Data columns (total 17 columns):
 #   Column                 Dtype
---  ------                 -----
 0   class                  int64
 1   department             object
 2   entry_date             object
 3   general_id             int64
 4   normative_identifier   object
 5   publication_date       object
 6   regulatory_authority   object
 7   subject                object
 8   subject_length         int64
 9   subject_unique_words   float64
 10  subject_words          float64
 11  text                   object
 12  text_length            int64
 13  text_unique_words      float64
 14  text_words             float64
 15  type                   object
 16  unique_document_id     int64
dtypes: float64(4), int64(5), object(8)
memory usage: 2.7 GB
```

Figure 1: BBRC data schema

impacting regulatory documents that could affect the businesses of department B. One department could have to check 10 regulators' websites, while another department could have to check 50 regulators' websites. This difference occurs because of the characteristics of the business in which the department is involved.

RR was created to solve all these problems. At first, a pure AI (ML) application was thought to solve the issue. After all, AI is widely used in the financial industry (Wall 2018; Zhang et al. 2018). However, the small amount of initial samples and the overlapping classes showed that the use of rules (regex) would also be necessary. So, what worked was a pipeline made up of ML models (Support Vector Machine - SVM) and deterministic rules for a binary classification challenge. Even if the aim of this article is not to present the application, a brief architecture explanation will help with corpora construction understanding. A detailed overview of the application was presented in the article published by de Azevedo et al. (2022).The application is presented in figure 2.

The application pipeline starts at **step 1**, it represents a hired company that collects daily all documents published by all regulators of interest of all departments of the bank (a little more than 100 regulatory authorities have their publications classified daily by the application). In the preprocessing phase (**step 2**), the numbers, special characters, and Portuguese stop words (NLTK) present in the document are removed. All tokens are turned to ASCII version and lowercased. Vectorization is performed by the TF-IDF algorithm, also in Step 2. In steps 3 to 6, the single regulatory document (norm/law) that entered the pipeline will be evaluated in an iterative manner for all models and rules

Department	Relevant	Irrelevant
BB Seguros	184	2,402
BB Asset	702	5,379
CIB	703	1,828
COGER	320	2,366
COGER GESUB	542	2,614
COGER GETRI	137	737
DICRE	88	3,858
DIGOV	253	1,397
DIMEP	581	3,494
DINED	121	9
DIOPE	361	2,333
DIOPE GEFID	403	2,559
DIPES	79	50
DIRAG	345	3,234
DIRIS	184	3,134
DISEM	439	1,566
DITEC	80	30
TESOU	411	3,093
UAC	53	781
UCF	157	129
UCI	193	4,153
UGE	144	363
UNI	94	1,329
UPB/MERCAP	429	2,933
USI	819	4,056
Total	**7,823**	**53,827**

Table 3: Number of samples per class of each department/corpus

Relevant class	Characters	Words	Unique words
Mean	26,221.44	3,786.43	755.22
Standard deviation	98,636.79	13,273.48	1,864.91
Minimum	7	1	1
25% (lower quartile)	1,612	246	150
50% (middle quartile) (median)	3,872	570.5	289
75% (upper quartile)	14,235	2,142	695
Maximum	1,457,062	190,940	31,196

Table 4: Statistics of the **text column** in the **relevant class** in terms of number of characters, number of words, and number of unique words.

of each department registered in the application.

In **step 3** the ML model of a department predicts whether the regulatory document is relevant or irrelevant for the department's business. In **step 4**, there is a rule that has keywords registered for the department that are searched in the text of the document being evaluated. If there is a match, the document is classified as relevant; otherwise, it is

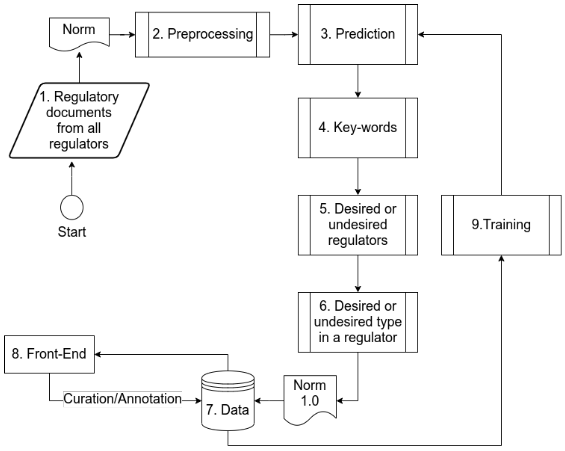


Figure 2: Radar Regulatório architecture. The pipeline each document (norm, law, etc.) passes through for each department.

classified as irrelevant. In the rule of **step 5**, each department can fill 2 lists, one of the desired regulators and another of the undesired ones. If the document evaluated by the rule was published by a regulator in the desired regulators list, the document will be classified as relevant. However, if the document was published by a regulator in the undesired list, it is classified as irrelevant. The same two lists (desired and undesired) in **step 6** are filled only for regulators registered in the desired regulators list of step 5. These lists in step 6 refer to the type of document, as a regulator publishes different types of documents (the type attribute is presented on line 16 of Table 2).

The classification of a previous step can be replaced by the classification done in the current step, except for step 3, which is the first classification. Another point is that step 6 will only activate if the regulator of the document being evaluated is in the desired regulators list of step 5. Once the document is classified, it is saved in the database (**step 7**). The front-end of the application gets all classified documents of each department once a day and presents the relevant ones to the experts of each department (each expert only receives documents of its department). These professionals check the classification of the tool and indicate to the system if it is correct or not (annotation/curation) (**step 8**). From time to time, the ML model of each department is re-trained (**step 9**) with the annotated data stored in the database (step 7).

In summary, Radar Regulatório (Regulatory Radar) classification eases the work of all workers who used to do the same classification process manually. The application prevents errors that could lead to expensive fines and restrictions. In other words, it stops them from having to search for a needle (document) in the haystack once regulators publish far more documents that do not impact the company's businesses (irrelevant documents). The next section presents the benchmarks of the experiments performed with BBRC and the discussion.

5. Experiments and Discussion

This section presents baseline experiments with BBRC using five different algorithms. They are Multinomial Naive Bayes (Kibriya et al., 2005), Random Forest (RF) (Breiman, 2001), Support Vector Machine (SVM) (Cortes and Vapnik 1995; Platt et al. 1999; Chang and Lin 2011), eXtreme Gradient Boosting (XGBoost) (Chen and Guestrin, 2016), and BERTimbau (Base and Large) (Souza et al., 2020) (a variation of BERT (Devlin et al., 2018) for Portuguese). The source code for the Machine Learning (ML) experiments is available on GitHub[2]. Figure 8 and Figure 9 show the hyperparameters used in the BERTimbau experiments, which were made with batch_size equals 20, 512 tokens and 5 epochs. The experiments used the same preprocessing (cleaning) described in step 2 of Section 4. The difference is that only documents with at least 50 words (selected before preprocessing and cleaning) were elected to be part of the train and test. In the case of BERTimbau, the stop words were not removed and the characters were not turned into ASCII ones, but the UTF-8 version was kept. For shallow machine learning and deep learning algorithms (BERTimbau), a 60% train and a 40% test split was used. All experiments were carried out using a fixed seed (random_state), and the data was stratified. We conducted two different experiments: a simple binary classification and, a binary classification dealing with the intraclass imbalancing problem. GridSearchCV optimization (Pedregosa et al., 2011) was applied to shallow machine learning algorithms. The hyperparameters of BERTimbau are presented in the Appendix A Section, the ML used algorithms are in the GitHub mentioned above.

All experiments used only samples from the text column of each corpus used. In the first experiment, we made approaches with and without the undersampling technique to evaluate the effect of the inter-class imbalance problem. The first experiment was carried out with 6 corpus of the corpora. They were USI, CIB, BB Asset, DIMEP, COGER GESUB, and DISEM. The second experiment used 3 corpus, which were USI, CIB and, BB Asset.

The **first experiment (experiment 1)** was a binary classification, its results are presented in Table 5. BERTimbau had the best results in all the

[2] Code available at `https://github.com/bancodobrasil/bbrc`

comparisons in the imbalanced data experiment, which could indicate the superiority of LLMs (deep learning) over shallow machine learning in this scenario. The BERT algorithm and its variations have been successful in many works (Sarkar et al. 2021; Huang et al. 2023; Campiotti et al. 2023). However, when the data was balanced by the undersampling technique, there was no winner algorithm, the results were pretty close, and it showed the impact of undersampling compared to imbalanced data in the results. Furthermore, corpus in Table 5 are sorted in descending order, considering the largest number of samples from the relevant class. The USI corpus is the one with the largest number of samples from the relevant class, and the DISEM corpus is the one with the smallest number of samples from that class (see Table 3). Undersampling was done using the total number of samples from the relevant class in the irrelevant class (819 samples from the USI corpus, 702 samples from the BB Asset corpus, 439 samples from the DISEM corpus, etc. - see Table 3).

In the **second experiment (experiment 2)**, the problem of intraclass imbalance was addressed. This problem exists in all corpus of BBRC. It happens to both classes, relevant and irrelevant. The point is that in the same class there exist more samples from some regulators than samples from other regulators. It happens because some regulators publish far more regulatory documents than others. To carry out the experiment, we chose the four regulators with more documents in the relevant class of each of the three corpus (USI, CIB, and BB Asset). To be part of the evaluation, regulators must have documents in both classes of the corpus. Only regulators with at least 10 documents in the relevant class were chosen. To perform intraclass undersampling, we took the regulator with the fewest documents in the relevant class for each corpus (considering the prerequisites already-mentioned), 26 DOU documents in the USI corpus, for example. Table 6 shows the number of samples available in the corpus chosen for the experiment. Table 7 presents the results of the three corpus evaluated in the second experiment. In this experiment, interclass undersampling was also applied. We observe that shallow machine learning algorithms had better results in corpus such as USI and CIB. However, surprisingly, BERTimbau got the best result in BB Asset, which is exactly the one with fewer samples.

The future work is presented in the next section.

6. Future Work

In future, we intend to expand the corpora with more samples and possibly offer to the scientific community a version of BBRC with scores classification, instead of a binary one. Furthermore,

Corpus	Classifier	F1 score undersampling	F1 score imbalanced
USI	Multin. NB	0.8225	0.7978
	RF	0.8487	0.8097
	SVM	0.8716	0.8456
	XGB	**0.8874**	0.8116
	BERTimbau Base	0.8822	0.8925
	BERTimbau Large	0.8869	**0.8972**
CIB	Multin. NB	0.6691	0.5919
	RF	0.7715	0.595
	SVM	0.7783	0.6485
	XGB	0.7783	0.6758
	BERTimbau Base	**0.782**	**0.7909**
	BERTimbau Large	0.78	0.7634
BB Asset	Multin. NB	0.8532	0.6091
	RF	0.8774	0.4132
	SVM	0.8717	0.6141
	XGB	**0.8884**	0.5465
	BERTimbau Base	0.8829	0.7728
	BERTimbau Large	0.8646	**0.7935**
DIMEP	Multin. NB	**0.8931**	0.538
	RF	0.8742	0.4256
	SVM	0.8919	0.552
	XGB	0.8714	0.508
	BERTimbau Base	0.8749	0.7643
	BERTimbau Large	0.8872	**0.7716**
COGER GESUB	Multin. NB	0.8461	0.5573
	RF	0.8535	0.4368
	SVM	0.8542	0.5155
	XGB	0.8401	0.5014
	BERTimbau Base	**0.855**	0.7155
	BERTimbau Large	0.8117	**0.731**
DISEM	Multin. NB	0.8684	0.6203
	RF	**0.8936**	0.4835
	SVM	0.8557	0.561
	XGB	0.8739	0.5306
	BERTimbau Base	0.8888	0.721
	BERTimbau Large	0.8747	**0.7485**

Table 5: Results of 6 different models trained on 6 corpus of BBRC for the binary classification task.

	Relevant			Irrelevant		
	USI	**CIB**	**BB Asset**	**USI**	**CIB**	**BB Asset**
ANBIMA	-	55	221	-	85	399
B3	-	302	147	-	592	1921
BACEN	666	175	-	467	257	
COAF	28	-	-	48	-	-
CVM	-	49	253		368	783
DOU	26	-	-	2,982	-	-
Presidência da República	29	-	-	32	-	-
RFB	-	-	13	-	-	678

Table 6: Quantity of samples per class, per corpus and per regulator used in the second experiment.

Corpus	Samples per class	Unique words	Classifier	F1 score undersampling
USI	104	13,272	**Multin. NB**	**0.6434**
			RF	0.5542
			SVM	0.617
			XGB	0.619
			BERTimbau Base	0.4034
			BERTimbau Large	0.458
CIB	196	21,236	Multin. NB	0.7234
			RF	**0.7283**
			SVM	0.7261
			XGB	0.6499
			BERTimbau Base	0.5895
			BERTimbau Large	0.6494
BB Asset	52	8,472	Multin. NB	0.5
			RF	0.65
			SVM	0.5142
			XGB	0.65
			BERTimbau Base	0.4986
			BERTimbau Large	**0.6666**

Table 7: Results from 6 different classifiers (inductors) trained on 3 BBRC corpus attacking the intraclass problem (binary classification).

we intend to make experiments with BBRC using Generative AI.
The conclusion section ends the paper.

7. Conclusion

We present BBRC, a corpora that brings together several corpus of regulatory risk documents from the Brazilian banking/financial sector. There are 25 corpus from different areas of banking activities such as insurance, agribusiness, human resources, payment methods, security, investments, among others. In this corpora, 26 regulatory authorities in the financial sector are represented. Each corpus was built for binary classification, as they are used at Banco do Brasil in a tool that has been in production since 2020. In total, the corpora has 61,650 documents, all relevant ones were annotated by experts in the area who built each corpus for the needs of their department. We used some of the BBRC corpus to perform binary classification benchmarks with some shallow and deep learning (LLMs) algorithms. We believe that BBRC can help researchers explore ML applied to the regulatory risk and legal field, document analysis, text classification, sentiment analysis (Nopp and Hanbury 2015; Agarwal et al. 2019) and other tasks.
The contribution of the corpora can easily go beyond AI or computer science (Wu and Salomon 2017; Kim et al. 2013; De Masi et al. 2023), as regulatory texts in banking/finance can be used, for example, to assess whether the rent of natural resources is a blessing or a curse for a country that has its economy based on these resources (Tang et al., 2022). The same corpora can be used to analyze the possibility of regulatory lobbying in favor of consolidated financial companies, as a way to prevent new entrants into the sector (Manish and O'Reilly, 2019). The same data can even allow the study of the impact of regulation on national or foreign banks (Wu and Salomon, 2017). Regulatory datasets are essential sources of study to evolve regulations, which are often not prepared for new events outside its context, such as pandemics, climate change (Le Quang and Scialom, 2022), and other crises (Thiemann et al., 2021).
We also hope that the public sharing of BBRC will encourage the sharing of more corpus of banking regulation and other areas, for Portuguese and other languages. Finally, we hope that our data and benchmarks encourage further exploration of better-performing models and techniques. The link to download BBRC is in Section 3.

8. Acknowledgements

The authors thank Banco do Brasil immensely for sharing such an important set of data to promote NLP research for Portuguese. This action demonstrates its commitment and understanding of the industry's active participation in the evolution of science and technology. Special thanks must be given to the Artificial Intelligence and Analytical Unit (**Unidade de Inteligência Artificial e Analítica - UAN**), the Internal Controls Board (**Diretoria de Controles Internos - DICOI**) and the Technology Board (**Diretoria de Tecnologia - DITEC**). We also would like to thank Tiago Nunes Silva and Leonardo Piccaro Rezende for their participation in the production of this paper and Radar Regulatório (Regulatory Radar).

9. Ethical Considerations

Making BBRC available to the scientific community is a rare opportunity for a company in the financial industry to share annotated data, especially in times of the Brazilian General Data Protection Law (Lei Geral de Proteção de Dados - LGPD). This was only possible because all corpora samples are public, since the regulatory documents that make up the corpora were published by regulatory authorities that make public publications, which affect the entire Brazilian financial sector.

10. Bibliographical References

Arvind Agarwal, Aparna Gupta, Arun Kumar, and Srikanth G Tamilselvam. 2019. Learning risk culture of banks using news analytics. *European Journal of Operational Research*, 277(2):770–783.

Iñaki Aldasoro, Leonardo Gambacorta, Paolo Giudici, and Thomas Leach. 2020. Operational and cyber risks in the financial sector.

Alaa Alhamzeh, Romain Fonck, Erwan Versmée, Elöd Egyed-Zsigmond, Harald Kosch, and Lionel Brunie. 2022. It's time to reason: Annotating argumentation structures in financial earnings calls: The finarg dataset. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 163–169.

Ron Artstein. 2017. Inter-annotator agreement. *Handbook of linguistic annotation*, pages 297–313.

Ting Wai Terence Au, Ingemar J Cox, and Vasileios Lampos. 2022. E-ner–an annotated named entity recognition corpus of legal text. *arXiv preprint arXiv:2212.09306*.

MJ Bommarito, Daniel Martin Katz, and E Detterman. 2018. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. *Research Handbook on Big Data Law*.

Leo Breiman. 2001. Random forests. *Machine learning*, 45:5–32.

Israel Campiotti, Matheus Rodrigues, Yuri Albuquerque, Rafael Azevedo, and Alyson Andrade. 2023. Debertinha: A multistep approach to adapt debertav3 xsmall for brazilian portuguese natural language processing task. *arXiv preprint arXiv:2309.16844*.

Sharanya Chakravarthy, Tushar Kanakagiri, Karthik Radhakrishnan, and Anjana Umapathy. 2020. Domino at fincausal 2020, task 1 and 2: causal extraction system. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 90–94.

Ilias Chalkidis, Nicolas Garneau, Catalina Goanta, Daniel Martin Katz, and Anders Søgaard. 2023. Lexfiles and legallama: Facilitating english multinational legal language model development. *arXiv preprint arXiv:2305.07507*.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. 2021. Lexglue: A benchmark dataset for legal language understanding in english. *arXiv preprint arXiv:2110.00976*.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Pedro Henrique Luz De Araujo, Teófilo Emídio de Campos, Fabricio Ataides Braz, and Nilton Correia da Silva. 2020. Victor: a dataset for brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458.

Rafael Faria de Azevedo, João Pedro Santos Rodrigues, Mayara Regina da Silva Reis, Claudia Maria Cabral Moro, and Emerson Cabrera Paraiso. 2018. Temporal tagging of noisy clinical texts in brazilian portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 231–241. Springer.

Rafael Faria de Azevedo, Tiago Nunes Silva, Henrique Tibério Brandão Vieira Augusto, Paulo Oliveira Sampaio Reis, Isadora Bastos Chaves, Samara Beatriz Naka de Vasconcellos, Liliany Aparecida dos Anjos Pereira, Mauro Melo de Souza Biccas, André Luiz Monteiro, and Alexandre Rodrigues Duarte. 2022. Banking regulation classification in portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pages 137–147. Springer.

Sara De Masi, Kose John, Agnieszka Słomka-Gołębiowska, and Piotr Urbanek. 2023. Regulation and post-crisis pay disclosure strategies of banks. *Review of Quantitative Finance and Accounting*, pages 1–33.

Pedro Delfino, Bruno Cuconato, Guilherme Paulino-Passos, Gerson Zaverucha, and Alexandre Rademaker. 2018. Using openwordnet-pt for question answering on legal domain. In *Proceedings of the 9th Global Wordnet Conference*, pages 105–112.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Mahmoud El-Haj, Marina Litvak, Nikiforos Pittaras, George Giannakopoulos, et al. 2020. The financial narrative summarisation shared task (fns 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 1–12.

Kishaloy Halder, Lahari Poddar, and Min-Yen Kan. 2017. Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 127–135.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A french corpus and annotation schema for named entity recognition and relation extraction of financial news. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2293–2299.

Hang Jiang, Zhongchen Miao, Yuefeng Lin, Chenyu Wang, Mengjun Ni, Jian Gao, Jidong Lu, and Guangwei Shi. 2020. Financial news annotation by weakly-supervised hierarchical multi-label learning. In *Proceedings of the second workshop on financial technology and natural language processing*, pages 1–7.

Ashraf M Kibriya, Eibe Frank, Bernhard Pfahringer, and Geoffrey Holmes. 2005. Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence: 17th Australian Joint Conference on Artificial Intelligence, Cairns, Australia, December 4-6, 2004. Proceedings 17*, pages 488–499. Springer.

Teakdong Kim, Bonwoo Koo, and Minsoo Park. 2013. Role of financial regulation and innovation in the financial crisis. *Journal of Financial stability*, 9(4):662–672.

Rosa M Lastra. 2019. Multilevel governance in banking regulation. *The Palgrave Handbook of European Banking Union Law*, pages 3–17.

Gaëtan Le Quang and Laurence Scialom. 2022. Better safe than sorry: Macroprudential policy, covid 19 and climate change. *International Economics*, 172:403–413.

Els Lefever and Véronique Hoste. 2016. A classification-based approach to economic event detection in dutch news text. In *10th International Conference on Language Resources and Evaluation (LREC)*, pages 330–335. ELRA.

Marcos Lima, Roberta Silva, Felipe Lopes de Souza Mendes, Leonardo R de Carvalho, Aleteia Araujo, and Flavio de Barros Vidal. 2020. Inferring about fraudulent collusion risk on brazilian public works contracts in official texts using a bi-lstm approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1580–1588.

Zhining Liu, Pengfei Wei, Zhepei Wei, Boyang Yu, Jing Jiang, Wei Cao, Jiang Bian, and Yi Chang. 2021. Handling inter-class and intra-class imbalance in class-imbalanced learning. *arXiv preprint arXiv:2111.12791*.

Dimitris Mamakas, Petros Tsotsi, Ion Androutsopoulos, and Ilias Chalkidis. 2022. Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. *arXiv preprint arXiv:2211.00974*.

GP Manish and Colin O'Reilly. 2019. Banking regulation, regulatory capture and inequality. *Public Choice*, 180(1-2):145–164.

Youness Mansar and Sira Ferradans. 2018. Sentence classification for investment rules detection. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 44–48.

Dominique Mariko, Hanna Abi Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. Financial document causality detection shared task (fincausal 2020). *arXiv preprint arXiv:2012.02505*.

Corentin Masson and Patrick Paroubek. 2020. Nlp analytics in finance with dore: a french 250m tokens corpus of corporate annual reports. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2261–2267.

Sérgio Moro, Paulo Cortez, and Paulo Rita. 2016. An automated literature analysis on data mining applications to credit risk assessment. *Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics*, pages 161–177.

Clemens Nopp and Allan Hanbury. 2015. Detecting risks in the banking system by sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 591–600.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, pages 557–566.

Gabriella Pardelli, Manuela Sassi, Sara Goggi, and Stefania Biagioni. 2012. From medical language processing to bionlp domain. In *LREC*, pages 2049–2055.

Fotios Pasiouras, Sailesh Tanna, and Constantin Zopounidis. 2009. The impact of banking regulations on banks' cost and profit efficiency: Cross-country evidence. *International review of financial analysis*, 18(5):294–302.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Valeria Quochi, Monica Monachini, Riccardo Del Gratta, and Nicoletta Calzolari. 2008. A lexicon for biology and bioinformatics: the bootstrep experience. In *LREC*. Citeseer.

Rajdeep Sarkar, Atul Kr Ojha, Jay Megaro, John Mariano, Vall Herard, and John Philip McCrae. 2021. Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 102–106.

Kyunghwan Sohn, Sunjae Kwon, and Jaesik Choi. 2021. The global banking standards qa dataset (gbs-qa). In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 19–25.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Bertimbau: pretrained bert models for brazilian portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 403–417. Springer.

Chang Tang, Muhammad Irfan, Asif Razzaq, and Vishal Dagar. 2022. Natural resources and financial development: Role of business regulations in testing the resource-curse hypothesis in asean countries. *Resources Policy*, 76:102612.

Matthias Thiemann, Carolina Raquel Melches, and Edin Ibrocevic. 2021. Measuring and mitigating systemic risks: how the forging of new alliances between central bank and academic economists legitimize the transnational macroprudential agenda. *Review of international political economy*, 28(6):1433–1458.

Paul Thompson, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo Del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, Valeria Quochi, et al. 2011. The biolexicon: a large-scale terminological resource for biomedical text mining. *BMC bioinformatics*, 12(1):1–29.

Larry D Wall. 2018. Some financial regulatory implications of artificial intelligence. *Journal of Economics and Business*, 100:55–63.

Zheying Wu and Robert Salomon. 2017. Deconstructing the liability of foreignness: Regulatory enforcement actions against foreign banks. *Journal of International Business Studies*, 48:837–861.

Qi Zhang, Jue Wang, Aiguo Lu, Shouyang Wang, and Jian Ma. 2018. An improved smo algorithm for financial credit risk assessment–evidence from china's banking. *Neurocomputing*, 272:314–325.

Fan Zhou, Shengming Zhang, and Yi Yang. 2020. Interpretable operational risk classification with semi-supervised variational autoencoder. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 846–852.

Nadhem Zmandar, Tobias Daudert, Sina Ahmadi, Mahmoud El-Haj, and Paul Rayson. 2022. Cofif plus: A french financial narrative summarisation corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1622–1639.

A. Appendix

	Department	Description
1	BB Seguros	BB Seguridade Participações is a holding company controlled by Banco do Brasil and operates in the insurance business. The group comprises the controlled companies BB Corretora de Seguros e Gestora de Bens and BB Seguros Participações and their subsidiaries.
2	BB Asset	BB ASSET (DTVM): BB Gestão de Recursos - Distribuidora de Títulos e Valores Mobiliários is a company specialized in the management of third-party resources and in the management of investment funds for Banco do Brasil clients
3	CIB	Corporate and Investment Bank (CIB) Board: acts as client, product and channel manager within the Corporate and Large Corporate segments
4	COGER	Accounting Board (COGER): operates within the scope of accounting strategies; standardization, bookkeeping, control and accounting disclosure; accounting statements; tax planning and management; accounting information to the market; and results of availability, integrity, reliability and compliance of accounting information.
5	COGER GESUB	Coger Executive Management that operates within the scope of BB Subsidiaries
6	COGER GETRI	Executive Management at Coger, which operates in the scope of planning, tax management and tax compliance.
7	DICRE	Credit Department (DICRE): Strategic Unit that operates in the management of credit risk, credit portfolios, customer registration, guarantees, parameterization of credit operations and credit limits, as well as the development of solutions for the credit process in the organization.
8	DIGOV	Acts as manager of clients and government products
9	DIMEP	Payment Means and Services Department (DIMEP): acts as product manager and operational support for Business Transactions, within the scope of the cards, vouchers and Instant Payment System (PIX) market
10	DINED	Digital Business Directorate (DINED): acts as strategy, product and channel manager within the scope of new digital business models, covering startups, distribution on BB's digital platforms and digital ecosystems, including Bank as a Service (BaaS)
11	DIOPE	Operations Board (DIOPE): acts as operational support for business transactions, for internal processes and logistics
12	DIOPE GEFID	Executive Management of Fiduciary Services (GEFID), subordinate to DIOPE: operates within the scope of fiduciary services (specialized services, with duties and attributions arising from legislation and market supervisory bodies, which guarantee the security and credibility required by investors)
13	DIPES	Culture and People Management Directorate (DIPES): Strategic unit that operates within the scope of people management, including recruitment and selection, career, training, remuneration and benefits.

14	DIRAG	Agribusiness Board (DIRAG): manages customers, products and operational support for business transactions, within the scope of agribusiness.
15	DIRIS	Risk Management Board (DIRIS): operates in risk management.
16	DISEM	Business Solutions Directorate (DISEM): Strategic unit that acted as product and channel manager within companies (customers) of various sizes (middle, upper middle, and high middle.)
17	DITEC	Technology Directorate (DITEC): Strategic unit that operates within the scope of Information Technology, as well as IT risk management, IT models and projects, etc.
18	TESOU	Global Treasury Unit (TESOU): operates in cash and liquidity management; treasury operations; and financial portfolio management
19	UAC	Service and Channels Unit (UAC): Strategic unit that operates within the scope of channel management, including monitoring the internal and external environment in relation to standards, regulations,and demands for service and relationship channels, standardization of service procedures performed in the customer service and management of banking correspondents, among others.
20	UCF	Cyber and Fraud Prevention Unit (UCF): operates within the scope of document and electronic fraud prevention strategies; digital/cyber security policies, models, methodologies, tools, standards, and instruments; security management in electronic channels; and results of risks and losses incurred in digital/cyber security processes
21	UCI	Fundraising and Investments Unit (UCI): acts as a product and channel manager within the scope of funding and investment products
22	UGE	Related Entities Governance Unit (UGE): Strategic unit that operates within the scope of the governance of related entities and corporate operations.
23	UNI	International Business Unit (UNI): Strategic unit that operates within the scope of product and channel management at the international level.
24	UPB/MERCAP	UCI/MERCAP (UPB/MERCAP): works with fundraising products. It is subordinate to the UCI
25	USI	Institutional Security Unit (USI): operates in managing the security of environments and people, information security, privacy and protection of personal data.

Table 8: Departments (corpus) in the corpora and their description.

	Regulator or Acronym	URL
1	ANAC	https://www.gov.br/anac
2	ANBIMA	https://www.anbima.com.br
3	ANPD	https://www.gov.br/anpd
4	ANS	https://www.gov.br/ans
5	Legislative Assembly of the State of Mato Grosso	https://www.al.mt.gov.br/
6	B3	https://www.b3.com.br
7	BACEN/BCB	https://www.bcb.gov.br/

8	BM&F BOVESPA (currently B3)	https://www.b3.com.br/pt_br/regulacao/oficios-e-comunicados/bm-fbovespa/
9	BNDES	https://www.bndes.gov.br/wps/portal/site/home
10	Rio de Janeiro City Council	https://www.camara.rio/
11	CETIP (currently B3)	https://www.b3.com.br
12	CFC	https://cfc.org.br/
13	CIP	https://www2.cip-bancos.org.br/Paginas/Sobre.aspx
14	COAF	https://www.gov.br/coaf
15	CPC	https://www.cpc.org.br/CPC
16	CVM	https://www.gov.br/cvm
17	DOU	https://www.in.gov.br/servicos/diario-oficial-da-uniao
18	FEBRABAN	https://portal.febraban.org.br/
19	ITI	https://www.gov.br/iti
20	Ministry of Labour	https://www.gov.br/trabalho-e-emprego
21	Núclea	https://www.nuclea.com.br/
22	Presidência da República (PR)	https://www.gov.br/planalto
23	PREVIC	https://www.gov.br/previc
24	RFB	https://www.gov.br/receitafederal
25	STN	https://www.gov.br/tesouronacional
26	SUSEP	https://www.gov.br/susep

Table 9: All regulators present at BBRC (mentioned in Table 1).

class	department	entry_date	general_id	normative_identifier	publication_date	regulatory_authority
1	usi	2021-02-10	8290	0106/2021	2021-02-05	BACEN
1	usi	2021-02-12	8291	4888	2021-02-12	BACEN
1	usi	2021-02-12	8292	72	2021-02-12	BACEN
1	usi	2021-03-19	14875	36912	2021-03-19	BACEN
1	usi	2021-03-24	14879	36935	2021-03-24	BACEN
1	usi	2021-03-25	14880	81	2021-03-25	BACEN
1	usi	2021-04-01	24465	93	2021-04-01	BACEN
1	usi	2021-04-14	24518	99	2021-04-14	BACEN
1	usi	2021-04-14	24519	96	2021-04-14	BACEN
1	usi	2021-04-14	24520	86	2021-04-14	BACEN
1	usi	2021-04-14	24521	95	2021-04-14	BACEN
1	usi	2021-04-14	24522	97	2021-04-14	BACEN
1	usi	2021-04-14	24523	98	2021-04-14	BACEN
1	usi	2021-04-22	24530	89	2021-04-22	BACEN

Figure 3: BBRC columns overview (part 1)

subject	subject_length	subject_unique_words	subject_words
Caracterizado fornecimento intempestivo de informações ao Banco Central do Brasil, sobr	468	57.0	88.0
Altera a Resolução nº 4.734, de 27 de junho de 2019, dispondo sobre a realização de novas	218	29.0	35.0
Altera a Circular nº 3.952, de 27 de junho de 2019, dispondo sobre a realização de novas et	245	31.0	38.0
Divulga a realização de leilão de venda conjugado com leilão de compra pós-fixado Selic no	122	15.0	19.0
Divulga comunicado do Grupo de Ação Financeira contra a Lavagem de Dinheiro e o Financi	120	16.0	18.0
Disciplina os processos de autorização relacionados ao funcionamento das instituições de	224	26.0	32.0
Altera a Instrução Normativa BCB nº 20, que dispõe sobre os limites de valor para as transa	113	21.0	21.0
Divulga a versão 2.0 do Manual de Segurança do Open Banking.	60	10.0	11.0
Divulga a versão 2.0 do Manual de Escopo de Dados e Serviços do Open Banking.	77	13.0	15.0
Altera a Resolução BCB nº 32, de 29 de outubro de 2020, que estabelece os requisitos técr	196	28.0	31.0
Divulga a versão 2.0 do Manual de APIs do Open Banking.	55	10.0	11.0
Divulga a versão 1.0 do Manual de Experiência do Cliente no Open Banking.	73	12.0	13.0
Divulga a versão 2.0 do Manual de Serviços Prestados pela Estrutura Responsável pela Gov	112	15.0	17.0
Altera a Circular nº 3.682, de 4 de novembro de 2013, e seu Regulamento anexo, para disp	449	47.0	72.0

Figure 4: BBRC columns overview (part 2)

text	text_length	text_unique_words	text_words
DEPARTAMENTO DE RESOLUÇÃO E DE AÇÃO SANCIONADORA GERÊNCIA TÉCNICA EM SÃO PAULO DECIS	843	96.0	141.0
O Banco Central do Brasil, na forma do art. 9º da Lei nº 4.595, de 31 de dezembro de 1964, torna público	1808	169.0	330.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão extraordinária realizada em 11 de fevereiro d	3251	243.0	562.0
Aviso de Abertura do Leilão de Câmbio 11/2021 O Departamento das Reservas Internacionais (DEPIN), dc	723	80.0	105.0
Comunicamos, com referência ao previsto no art. 39, alínea "g", inciso I, da Circular nº 3.978, de 23 de jar	1237	92.0	121.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão realizada em 25 de março de 2021, com base	27568	1058.0	4224.0
O Chefe do Departamento de Competição e de Estrutura do Mercado Financeiro (Decem), no uso das atrib	1206	102.0	150.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	22221	1069.0	3214.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	49205	1435.0	7039.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão realizada em 14 de abril de 2021, com base r	7283	395.0	1041.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	19293	1100.0	2785.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	18126	815.0	2657.0
Os Chefes do Departamento de Regulação do Sistema Financeiro (Denor) e do Departamento de Tecnolog	30146	1438.0	4482.0
A Diretoria Colegiada do Banco Central do Brasil, em sessão realizada em 20 de abril de 2021, com base r	9726	532.0	1470.0

Figure 5: BBRC columns overview (part 3)

type	unique_document_id
PROCESSO ADMINISTRATIVO SANCIONADOR	787104
RESOLUÇÃO	787954
RESOLUÇÃO	787955
COMUNICADO	797670
COMUNICADO	798871
RESOLUÇÃO BCB	799306
INSTRUÇÃO NORMATIVA BCB	801512
INSTRUÇÃO NORMATIVA BCB	805258
INSTRUÇÃO NORMATIVA BCB	805257
RESOLUÇÃO BCB	805252
INSTRUÇÃO NORMATIVA BCB	805254
INSTRUÇÃO NORMATIVA BCB	805255
INSTRUÇÃO NORMATIVA BCB	805256
RESOLUÇÃO BCB	808061

Figure 6: BBRC columns overview (part 4)

O Banco Central do Brasil, na forma do art. 9º da Lei nº 4.595, de 31 de dezembro de 1964, torna público que o Conselho Monetário Nacional, em sessão extraordinária realizada em 11 de fevereiro de 2021, com base no disposto nos arts. 4º, incisos VI e VIII, da referida Lei, e 26-A da Lei nº 12.810, de 15 de maio de 2013, R E S O L V E U : Art. 1º A Resolução nº 4.734, de 27 de junho de 2019, passa a vigorar com as seguintes alterações: “Art. 7º-B As instituições financeiras de que trata o art. 7º-A devem estar aptas a cumprir o disposto nesta Resolução a partir da data mencionada no inciso II do art. 11. § 1º A aptidão de que trata o caput será atestada pelo cumprimento, com sucesso, de todas as etapas dos testes homologatórios de integração de que trata o art. 7º-A, conforme cronograma de que trata o inciso I do art. 8º. § 2º O descumprimento de qualquer etapa dos testes homologatórios de que trata o art. 7º-A sujeita as instituições financeiras às sanções e às medidas administrativas previstas na legislação em vigor, bem como, a critério do Banco Central do Brasil, à suspensão provisória da realização das operações de que trata o art. 1º, a partir da data mencionada no inciso II do art. 11. § 3º O Banco Central do Brasil, ao determinar a suspensão de que trata o § 2º, estabelecerá as condições mediante as quais essa suspensão será levantada.” (NR) “Art. 11. .... I - na data de sua publicação, em relação aos arts. 7º-A, 7º-B, 8º e 9º; e II - em 7 de junho de 2021, em relação aos demais dispositivos.” (NR) Art. 2º Ficam revogados os §§ 4º e 5º do art. 7º-A da Resolução nº 4.734, de 2019. Art. 3º Esta Resolução entra em vigor na data de sua publicação. Roberto de Oliveira Campos Neto Presidente do Banco Central do Brasil\r\n

Figure 7: Text sample of ”unique_document_id” number 787954, annotated as ”relevant” by DICRE and USI. The same sample was annotated as ”irrelevant” by CIB and DIMEP. The document was published by BACEN (Brazilian Central Bank)

```python
def tokenizer_model(self, model_path, df, token):
    tokenizer = BertTokenizer.from_pretrained(BERTIMBAU_TOKENIZER, do_lower_case=True)

    # Tokenizing the training data
    encoded_data_train = tokenizer.batch_encode_plus(
        df[df.data_type=='train'].new_content.tolist(),
        add_special_tokens=True,
        return_attention_mask=True,
        padding=True,
        max_length=token,
        return_tensors='pt',
        truncation=True
    )

    # Tokenizing test data
    encoded_data_val = tokenizer.batch_encode_plus(
        df[df.data_type=='val'].new_content.tolist(),
        add_special_tokens=True,
        return_attention_mask=True,
        padding=True,
        max_length=token,
        return_tensors='pt',
        truncation=True
    )


    input_ids_train = encoded_data_train['input_ids']
    attention_masks_train = encoded_data_train['attention_mask']
    labels_train = torch.tensor(df[df.data_type=='train'].label_num.values)

    input_ids_val = encoded_data_val['input_ids']
    attention_masks_val = encoded_data_val['attention_mask']
    labels_val = torch.tensor(df[df.data_type=='val'].label_num.values)

    dataset_train = TensorDataset(input_ids_train, attention_masks_train, labels_train)
    dataset_val = TensorDataset(input_ids_val, attention_masks_val, labels_val)

    return(dataset_train, dataset_val)
```

Figure 8: BERTimbau hyperparameters (part 1)

```python
def setup_model(self, model_path, dataset_train, dataset_val, epochs):
    model = BertForSequenceClassification.from_pretrained(model_path,
                                                          num_labels=len(LABEL_DICT),
                                                          output_attentions=False,
                                                          output_hidden_states=False)

    device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
    model.to(device)
    batch_size = BATCH_SIZE
    dataloader_train = DataLoader(dataset_train,
                                  sampler=RandomSampler(dataset_train),
                                  batch_size=batch_size)
    dataloader_validation = DataLoader(dataset_val,
                                       sampler=SequentialSampler(dataset_val),
                                       batch_size=batch_size)
    optimizer = AdamW(model.parameters(),
                      lr=1e-5,
                      eps=1e-8)

    scheduler = get_linear_schedule_with_warmup(optimizer,
                                                num_warmup_steps=100,
                                                num_training_steps=len(dataloader_train)*epochs)
    return(model, dataloader_train, dataloader_validation, epochs, scheduler, optimizer, device)
```

Figure 9: BERTimbau hyperparameters (part 2)

Stock Price Prediction with Sentiment Analysis for Chinese Market

Yuchen Luan[1], Haiyang Zhang[1*], Chenlei Zhang[1], Yida Mu[2], Wei Wang[1]
[1]Xi'an Jiaotong Liverpool University
[2]The University of Sheffield
{yuchen.luan22, chenlei.zhang}@student.xjtlu.edu.cn
{haiyang.zhang, wei.wang03}@xjtlu.edu.cn
y.mu@sheffield.ac.uk

Abstract

Accurate prediction of stock prices is considered as a significant practical challenge and has been a longstanding topic of debate within the economic domain. In recent years, sentiment analysis on social media comments has been considered an important data source for stock prediction. However, most of these works focus on exploring stocks with high market values or from specific industries. The extent to which sentiments affect a broader range of stocks and their overall performance remains uncertain. In this paper, we study the influence of sentiment analysis on stock price prediction with respect to (1) different market value groups and (2) different Book-to-Market ratio groups in the Chinese stock market. To this end, we create a new dataset that consists of 24 stocks across different market value groups and Book-to-Market ratio categories, along with 12,000 associated comments that have been collected and manually annotated. We then utilized this dataset to train a variety of sentiment classifiers, which were subsequently integrated into sequential neural-based models for stock price prediction. Experimental findings indicate that while sentiment integration generally improve the predictive performance for price prediction, it may not consistently lead to better results for individual stocks. Moreover, these outcomes are notably influenced by varying market values and Book-to-Market ratios, with stocks of higher market values and B/M ratios often exhibiting more accurate predictions. Among all the models tested, the Bi-LSTM model incorporated with the sentiment analysis, achieves the best prediction performance.

Keywords: Stock Price Prediction, Sentiment Analysis, Chinese Stock Market

1. Introduction

Stocks are frequently traded investment products, and accurately forecasting stock prices is regarded as a crucial practical concern. This topic has been a subject of ongoing debate in the field of economics, with numerous scholars proposing various methods to forecast stock market trends. In recent years, the rise of social media has led many investors to express their views and sentiments on stocks in online forums, prompting scholars and practitioners to pay attention to discourse on these investment platforms. Such information has been shown to offer evidence indicating that investor sentiment might play a pivotal role in explaining stock price fluctuations (Dewally, 2003; Sunny et al., 2020).

Most existing works on stock prediction with sentiment analysis follow a two-stage process: the first stage involves using sentiment classification methods to compute sentiment values, which are subsequently integrated into conventional time series stock price prediction models. (Jing et al., 2021; Tashiro et al., 2019; Sirignano and Cont, 2021; Hiew et al., 2019; Sidogi et al., 2021). Common models employed for sentiment analysis include Convolutional Neural Networks (CNNs) and BERT models. Sentiment analysis often employs various models, including Convolutional Neural Networks (CNNs) and BERT-based models, to interpret and classify emotions within text data effectively. Specifically, when analyzing sentiment in Chinese text, a significant number of studies prefer the Bert-base-Chinese model (BBC) for its general applicability. However, a smaller yet noteworthy body of research opts for the Erlangshen-MegatronBert-1.3B-Sentiment model (EMB-1.3B-S), which has been shown to outperform others in classification tasks, as highlighted in the literature (Zhang et al., 2022). As for the stock prediction task, the majority of studies aim to predict the future direction of stock movements as a classification task. In contrast, a lesser-explored avenue is to predict the exact stock price based on historical data, treating it as a regression task. For this latter task, Long-Short Term Memory (LSTM) networks are frequently chosen due to their proficiency in processing and analyzing time series data (Hiew et al., 2019; Sidogi et al., 2021).

In the realm of stock prediction research, a prevalent trend involves selecting stocks based on criteria such as market capitalization (Zhang et al., 2017; Liu and Chen, 2019) or industry sector (Huang et al., 2018; Wu et al., 2018). However, such methods introduce a selection bias where the chosen stocks often share similar features, leading to a lack of diversity within the analyzed portfolio. Even when

* denotes corresponding author.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 167–177

considering both market capitalization and industry factors together, it remains challenging to avoid the concentration of market capitalization within specific industries (Jing et al., 2021). For instance, stocks in the banking and food and beverage industries typically have high market capitalization, while those in the chemical and communication equipment industries tend to have lower market capitalization. This leads to an issue of similarity among the stocks to be predicted within the portfolio. An innovative approach to counteract this bias involves incorporating the Book-to-Market (B/M) ratio (Pontiff and Schall, 1998), a pivotal metric in value investment strategies indicating company valuation. Considering both the B/M ratio and market capitalization for stock selection can effectively mitigate this selection bias. In this paper, we examine the influence of sentiment analysis on stock price prediction with respect to (1) different market value groups and (2) different Book-to-Market ratio groups in the Chinese stock market. We train a set of sentiment classifiers, which are then incorporated with sequence-based deep learning models for price prediction. The contributions of this work are as follows:

- We construct a new dataset comprising 24 stocks from various market value and book-to-market ratio groups in the Chinese stock market, along with 12,000 corresponding comments that were collected and manually annotated.
- We employ various combinations of sentiment analysis models and sequence-based price prediction models to assess the impact of sentiment information on stock prediction.
- Experimental results suggest that while incorporating sentiment generally improves predictive performance, it may not consistently lead to superior results for individual stocks. Furthermore, the results are significantly influenced by different market values and Book-to-Market ratios. Among all the models tested, the Bi-LSTM model integrated with a sentiment factor demonstrates the highest prediction performance.

2. Datasets

2.1. Stock Selection

Considering the diverse market attributes of stocks in different market value portfolios in the Chinese market, we selected stocks from four market indexes from the Shanghai Stock Exchange (SSE)[1], namely `CSI 100`, `CSI 200`, `CSI 500`, and `CSI 1000`, representing portfolios of stocks with different market capitalizations and liquidity in the Chinese stock market. The `CSI 100` comprises the top 100 stocks with the largest market capitalization and best liquidity from the Shanghai and Shenzhen 300 indices, representing mega-cap stocks in the Chinese market; the `CSI 200` consists of 200 stocks excluding the constituents of the `CSI 100` index, representing large-cap stocks; the `CSI 500` and `CSI 1000` represent mid-cap and small-cap stocks, respectively. Subsequently, we constructed a 3×4 table by combining the three market capitalization portfolios with four B/M ratio portfolios. Six stocks meeting the selection criteria were randomly chosen from each cell of the table. For the 24 selected stocks, technical indicators including the opening price, closing price, highest price, lowest price, and trading volume have been collected from the China Stock Market & Accounting Research Database (CSMAR)[2]. Regarding technical indicators, we employ the Lagrange interpolation method to rectify missing and outlier values, subsequently arranging the data chronologically (de Resende et al., 2016).

Time Span To capture highly diverse price fluctuations and to alleviate concerns about data snooping, we selected data spanning from January 1, 2017, to December 31, 2022, covering 1,459 trading days. This interval has been deemed adequate by prior research for stock price prediction purposes, capturing essential fluctuations in market sentiment (Jiang, 2021). This selection ensures a comprehensive analysis period that incorporates significant market events and trends, providing a robust foundation for evaluating the impact of market sentiment on stock price movements.

2.2. Stock Comments Collection

For the experiments, we collected over 1.2 million stock comments related to the 24 selected stocks from the stock forum on the Financial Website (East Money)[3] for the corresponding 24 stocks. Given East Money's reputation as a leading financial information platform in China, the discussions on this forum are indicative of the broader sentiment among Chinese investors (Wang et al., 2018).

Data Filtering To ensure adherence to the fundamental requirements and standards of this experiment, we systematically excluded stocks previously categorized under *ST* or **ST* status[4], elimi-

[1] `http://english.sse.com.cn/`

[2] `http://www.data.csmar.com`

[3] `http://www.guba.eastmoney.com`

[4] In the Chinese stock market, *ST* represents "Special Treatment," indicating companies facing the risk of delisting due to financial distress or other issues, while

Average B/M \ Market Value	CSI 100	CSI 200	CSI 500	CSI 1000
High (33%)	601818.SH 601998.SH	000783.SH 600741.SH	000488.SH 600657.SH	000797.SH 601588.SH
Medium (33%)	600999.SH 002736.SH	600085.SH 601021.SH	000685.SH 300244.SH	002138.SH 002542.SH
Low (33%)	600585.SH 601336.SH	300144.SH 300033.SH	603355.SH 600259.SH	603989.SH 300377.SH

Table 1: Stocks selected based on the Fama-French three-factor model.

nated those with less than 50 trading weeks annually, and removed entries marked by missing data. This process retained stocks that consistently maintained their status as index component constituents throughout the designated trading period. The selected 24 stocks are listed in Table 1.

Text Pre-processing For the stock comments, we remove all garbled text, web links, and irrelevant short phrases.

Data Annotation Three annotators independently assigned sentiment labels: positive, neutral, or negative, to a set of 12,000 stock comments. To evaluate the inter-annotator reliability concerning the sentiment classification of these comments, we employed the Fleiss' Kappa (Fleiss, 1971) statistic as our chosen metric. This approach facilitates a quantifiable assessment of agreement levels among the three annotators, ensuring the consistency and accuracy of the sentiment labels assigned to the dataset. The Fleiss' Kappa is computed as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where $p_o = \frac{1}{N}\sum_{i=1}^{N} p_i$ is the average observed agreement probability across all raters for all samples, and $p_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1)$represents the degree of agreement observed among raters for each sample. n_{ij} is the number of raters who classified sample i into category j, n is the total number of raters (in this study, there are 3 raters), and k is the number of categories (in this study, there are 3 categories: positive, negative, and neutral). $p_e = \sum_{j=1}^{k} p_j^2$ represents the expected average agreement probability that raters can achieve when assigning ratings. $p_j = \frac{1}{Nn}\sum_{i=1}^{N} n_{ij}$ represents the mean number of raters assigned to each category, where N is the total number of samples. The Fleiss Kappa value for our annotation is **0.883**, indicating excellent agreement and demonstrating good classification consistency.

**ST* denotes a more severe level of “Special Treatment.”

3. Methodology

We propose a hybrid predictive pipeline that combines 1) a sentiment analysis model to predict the sentiment score based on the daily comments for each stock, and 2) a sequence model to predict time series stock price that includes the sentiment factor. The architecture of the proposed method is depicted in Figure 1.

3.1. Sentiment Analysis on Stock Comments

We explore a number of text classification methods for predicting the sentiment of stock comments, including traditional machine learning models (e.g., Support Vector Machine (SVM)) and neural-based models, such as Convolutional Neural Networks (CNN) (Luan and Lin, 2019) and Transformer-based models (Vaswani et al., 2017; Devlin et al., 2019).

SVM SVM is used as a baseline for our sentiment classification. It utilizes unigram and bigram bag-of-words, weighted using TF-IDF, as inputs. These are implemented using the default settings of scikit-learn (Pedregosa et al., 2011).

CNN CNN arpproaches leverage multiple convolutional kernels of varied granularities to meticulously extract text features. . This process begins with the generation of feature matrices, followed by the execution of one-dimensional convolution and pooling operations to distill and condense the information. The culmination of this process involves the application of the Softmax function for sentiment classification, which computes a probability distribution across the possible sentiment categories for a given text. Following (Kim, 2014),our approach integrates pre-trained word embeddings through two distinct embedding layers: static and non-static. The filter size is set to 3, where each type of filter comprises 100 filters. Then, max-pooling operations are employed to extract critical information, ultimately yielding output results in the fully-connected layer.

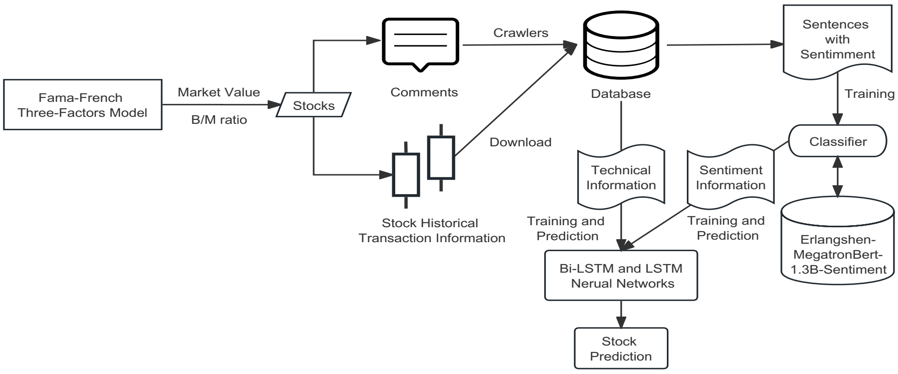


Figure 1: The design of the stock price prediction model in this study based on sentiment analysis.

CBERT & EMB-1.3B-S BERT (Devlin et al., 2019; Zhang et al., 2022), the pre-trained deep bidirectional Transformer, has shown strong performance on many NLP tasks (Devlin et al., 2019). Conventionally, it is pre-trained using two self-supervised tasks (masked language modeling and next sentence prediction) on a large corpus and fine-tuned for downstream tasks. In this paper, we fine-tune two pre-trained BERT models for Chinese for the sentiment classification task: Chinese Bidirectional Encoder Transformers[5] (CBERT)(Cui et al., 2021) and Erlangshen-MegatronBert-1.3B-Sentiment[6] (EMB-1.3B-S).

CBERT is pre-trained on an extra Chinese corpus (e.g., news articles and social media posts), based on the pre-existing checkpoint of the BertBase-Chinese model (Devlin et al., 2019), maintaining an identical structure (e.g., 12 layers and 110M parameters) to the vanilla BERT-base model. It achieves comparable predictive performance on multiple Chinese NLP downstream tasks compared to traditional machine learning approaches. EMB-1.3B-S, one of the largest open-source Chinese BERT models to date with 1.3 billion parameters, surpasses human performance on downstream tasks such as the TNEWS[7] Subtask.

We employ CBERT and EMB-1.3B-S in our task by incorporating an additional linear layer on top of the 12-layer transformer blocks with a Sigmoid activation, following the standard model fine-tuning pipeline introduced by (Devlin et al., 2019). For both transformer-based models, we set the maximum input length to 512 tokens. Additionally, to maintain consistency with the time input of the stock price prediction model, we computed the daily sentiment value (SV_t) for each trading day using the following equation:

$$SV_t = \frac{num_t^+ \cdot \mathrm{TScores}_t^+ - num_t^- \cdot \mathrm{TScores}_t^-}{num_t} \quad (2)$$

where $\mathrm{TScores}_t^+$ and $\mathrm{TScores}_t^-$ represent the sum of sentiment probability scores for all positive and negative labels corresponding to a stock on the t-th trading day, respectively; num_t denotes the total number of comments on the t-th trading day. The sentiment value ranges from -1 to 1, indicating the overall investor sentiment towards a particular stock on that day: a positive value suggests a predominance of positive sentiments, and a negative value indicates the opposite.

3.2. Stock Prediction with Sentiment Analysis

In our study, we deploy both Long Short-Term Memory (LSTM) and Bidirectional Long Short-Term Memory (Bi-LSTM) models, synergistically combined with sentiment analysis, to forecast stock prices. Specifically, we adopt a sliding window technique for predicting stock prices for the subsequent day. This method involves progressively moving the input window over the data series to generate predictions for each new time step. This approach allows for dynamic analysis of time-series data, where the LSTM and Bi-LSTM frameworks leverage both historical stock prices and sentiment indicators within each window to make informed predictions about future stock price movements.

LSTM Long Short-Term Memory (LSTM) networks, a subclass of recurrent neural networks (RNNs), enhance the RNN framework by effectively managing sequential data while overcoming the notorious gradient vanishing and exploding issues commonly associated with traditional RNNs (Hochreiter and Schmidhuber, 1997). LSTMs introduce a unique mechanism for long-term memory retention, enabling the model to make judicious

[5]https://huggingface.co/hfl/chinese-bert-wwm

[6]https://huggingface.co/IDEA-CCNL/Erlangshen-TCBert-1.3B-Sentence-Embedding-Chinese

[7]Toutiao News Classification Dataset

use of relevant historical information without being overly dependent on distant past data. This feature ensures a more balanced consideration of both recent and older inputs, significantly improving the network's ability to learn from sequences over extended periods.

The calcluation of the LSTM are is shown as follows:

$$\begin{cases} i_t = \sigma(w_i \cdot [H_{t-1}, X_t] + b_i) \\ f_t = \sigma(w_f \cdot [H_{t-1}, X_t] + b_f) \\ \tilde{C}_t = \tanh(w_c \cdot [H_{t-1}, X_t] + b_c) \\ o_t = \sigma(w_o \cdot [H_{t-1}, X_t] + b_o) \\ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \\ h_t = o_t \cdot \tanh(C_t) \end{cases} \tag{3}$$

where t represents the time point, X_t signifies the input value at the cell, and H_t represents the output state of the cell at the same time point. The symbols f_t, i_t, and o_t correspond to the formulas for the forget gate, input gate, and output gate, respectively. C_t denotes the cell state update. Matrices w_i, w_f, w_c, and w_o are the weight matrices for the input gate, forget gate, update gate, and output gate, respectively. Biases b_i, b_f, b_c, and b_o represent the respective biases. The activation function σ is applied to each gate unit, generating values between 0 and 1. This activation function is also applied to the cell state and output, constraining their values to a range between -1 and 1.

To enhance the LSTM model's capability for stock price prediction, we integrate sentiment values as supplementary features. More precisely, we concatenate the sentiment value SV_t as an additional feature of the input data, forming an augmented input vector, as shown in Equation 4.

$$i_t = \sigma(w_i \cdot [H_{t-1}, X_t, SV_t] + b_i) \tag{4}$$

By incorporating these sentiment values, they directly influence the operations of the input gate, forget gate, and the calculation of the input candidate value. This strategic integration empowers the model to adeptly leverage sentiment information, refining its ability to predict stock prices by learning from the nuanced interplay between market sentiment and stock price movements during the training phase.

Bi-LSTM The Bi-LSTM model, initially proposed by (Graves and Schmidhuber, 2005), consists of two LSTM layers that enable bidirectional processing of stock price information around time t. By leveraging historical data from both forward and backward directions, it jointly predicts the stock's closing price at time t. Bi-LSTM structure consists of two distinct LSTM layers aligned in parallel, each processing the temporal data sequence in opposite directions: one forward and the other backward. This setup allows for the comprehensive assimilation of contextual information, both preceding and following the target time t, thereby enriching the model's understanding and predictive accuracy of stock price movements by leveraging insights from both past and future contexts. The calculation of Bi-LSTM is represented as:

$$\begin{cases} \overrightarrow{h_t} = LSTM(\overrightarrow{h_{t-1}}, x_t) \\ \overleftarrow{h_t} = LSTM(\overleftarrow{h_{t+1}}, x_t) \\ h_t = (\overrightarrow{h_t}, \overleftarrow{h_t}) \end{cases} \tag{5}$$

Similarly, we integrate sentiment factors into the computation, where at each time step t, the sentiment factor is included in Bi-LSTM as part of the input x_t:

$$h_t = LSTM(h_{t-1}, x_t, SV_t) \tag{6}$$

This approach allows sentiment factors to influence the input gate, forget gate, and input candidate value computations, enabling the model to learn how to effectively use sentiment information for stock price prediction during the training process.

4. Experiment and Results

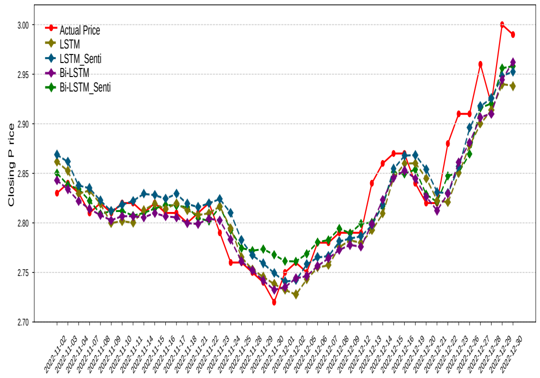
4.1. Experiments on Sentiment Analysis

To evaluate the predictive performance of various classifiers on sentiment classification, we use comments collected from January 1, 2017, to October 30, 2022, as the training set, and comments from November and December 2022 as the test set. We report precision, recall, and the F1 measure to assess their performance.

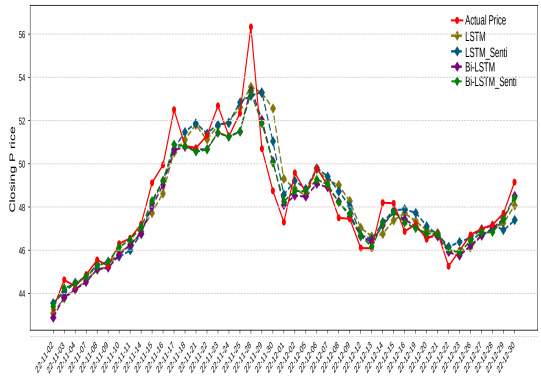
The average evaluation results are presented in Table 2. Considering the presence of data imbalance within the dataset, we employ a micro-average method for calculating the F-measure. As indicated in Table 2, the EMB-1.3B-S model achieves the best overall performance. Given that the number of comments collected from forums exceeds 1.2 million, this level of improvement can significantly enhance the accuracy of sentiment judgment. Therefore, employing this classifier for analyzing the hidden sentiments in text data collected from forums is feasible. In the stock price prediction phrase, we utilize the results from the EMB-1.3B-S model as one of the input features.

Metric	Precision	Recall	F-1
SVM	0.823	0.764	0.792
CNN	0.875	0.823	0.848
CBERT	0.947	0.946	0.946
EMB-1.3B-S	**0.970**	**0.969**	**0.969**

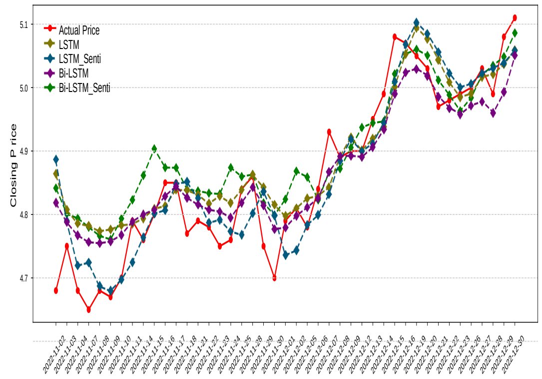
Table 2: Performance Comparison of Various Classifiers in Sentiment Analysis.



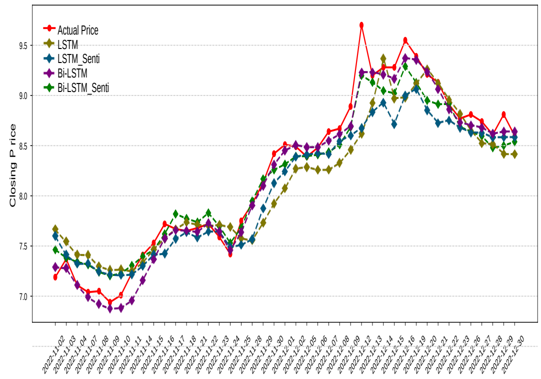
(a) 601818.SH. from CSI100, b/M ratio = high



(b) 600085.SH. from CSI200, b/M ratio = medium.



(c) 000488.SH. from CSI500, b/M ratio = high



(d) 300377.SH. from CSI1000, b/M ratio = low.

Figure 2: The actual closing price and the predicted prices across the four model combinations on one stock from each stock market index.

4.2. Experiments on Stock Price Prediction

In our experimental setup for forecasting stock prices, we merge sentiment scores obtained from sentiment analysis with technical indicators related to the stock market to predict the closing prices for the following day. This integration approach combines qualitative insights from investor sentiment with quantitative stock technical factors, providing a comprehensive view that enhances the accuracy of our predictive model for next-day closing prices. Same train/test data split are used as that of sentiment analysis. Two metrics are employed: Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) to evaluate the performance. Smaller values of these two metrics indicate that the model's predictions are closer to the actual values.

For both LSTM and Bi-LSTM, we set the input length to 3, and use 64D-3 layer neural networks. The batch size is 32. Figure 2 demonstrates the performance comparison using different methods: 1) LSTM, 2) LSTM with sentiment factor, 3) Bi-LSTM and 4) Bi-LSTM with sentiment factor, against the actual stock prices for one stock (600818.SH, 600085.SH, 000488.SH, 300377.SH) from each market index. It demonstrates that all prediction models accurately forecast the stock price trends.

Influence of Sentiment Factor Table 3 presents the aggregated performance of stocks from different market value groups using various models, with the best performance highlighted in bold. Performance for individual stock within each market value group are provided in Appendices. It is observed that incorporating sentiment information does not uniformly enhance prediction accuracy for every stock. This observation suggests that the effectiveness of sentiment data integration varies across different stocks, indicating a nuanced relationship between sentiment analysis and stock performance forecasting.

Table 4 aggregates the performance metrics for all stocks analyzed through various models, highlighting the comparative results. Notably, the Bi-LSTM model, augmented with sentiment data, demonstrates the best results, achieving a RMSE of 41.1603 and a MAPEof 145.5350. In contrast, the LSTM model that does not incorporate sentiment factors registers the least favorable outcomes, with an RMSE of 41.3073 and a MAPE of 148.6382. These findings indicate that integrating sentiment information for stock prediction can generated overall better performance.

Performance on different Market Value and B/M ratio groups Furthermore, by segmenting the results according to various market capitalizations and Book-to-Market (B/M) ratios, we noted marked variations in model performance across different segments, as detailed in Table 5. Particularly, the `CSI 100` group exhibited the best performance, with an RMSE of 24.4222 and a MAPE of 102.4448. Conversely, the `CSI 200` group recorded the highest RMSE at 73.8258, while the `CSI 1000` group had the highest MAPE at 186.7908, indicating that the model performs excellently in predicting stocks with higher market values.

We also conducted an analysis to evaluate the impact of sentiment factor on different Book-to-Market (B/M) ratios, with the results detailed in Table 6. The findings indicate that stocks categorized within the High B/M ratio group exhibited the most accurate predictions, with an RMSE around 3.8, showcasing their robustness in predictive accuracy. In contrast, stocks within the Low B/M ratio group displayed the least favorable performance. It also reveals a trend where the overall RMSE progressively increases as the B/M ratio shifts from High to Low.

4.3. Ablation Study

Influence of B/M ratio To assess the influence of the Book-to-Market (B/M) ratio and the effect of integrating sentiment analysis on the models' overall efficacy, we embarked on a detailed ablation study. Specifically, we investigated the relationship between daily sentiment values and actual closing prices for stocks grouped by their B/M ratios. For this purpose, we employed the Pearson correlation coefficient (Asuero et al., 2006) as our primary metric. This coefficient is determined by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{7}$$

Table 7 demonstrates the Pearson correlation coefficients between sentiment scores and closing prices within different B/M ratio groups. It is observed that the highest correlation is observed within the high B/M ratio group, suggesting a pronounced link between sentiment scores and closing prices in this group. This finding aligns with the superior performance of our predictive model within the high B/M ratio group.

Influence of market index We noted that the proposed model combination performed best in the CSI 100 portfolio, consistent with the characteristics of large-cap companies, which typically possess advantages such as high stability, high liquidity, and comprehensive information disclosure. High stability and liquidity often manifest as relatively stable technical indicators, favoring predictions from single models. Moreover, comprehensive information disclosure implies richer information about high market value stocks, making them focal points for investors' attention, naturally accompanied by more stock comments. To verify this, We also investigated the relationship between the number of stock reviews and predictive results within different market value groups. It was observed that in the `CSI 100` high market value group, there were the most stock reviews (280,583 records), while in the `CSI 1000` low market value group, there were the fewest stock reviews (203,526 records). This finding aligns with the focus of public attention, as stocks with higher market values are typically associated with larger companies and enjoy greater exposure, thus attracting more stock review information.

Drawing from the insights garnered in this study, investors and analysts looking to leverage time series models for forecasting stock prices in the Chinese market might benefit from focusing on stocks characterized by high Book-to-Market (B/M) ratios and exceptionally large market values, specifically those within the CSI 100 category. These segments have shown to yield more accurate predictive outcomes. Additionally, for models that incorporate sentiment analysis into the stock price forecasting process, the Bi-Long Short-Term Memory (Bi-LSTM) model emerges as a more effective option compared to the Long Short-Term Memory (LSTM) model. This recommendation is based on the Bi-LSTM model's superior performance, especially when analyzing stocks with high B/M ratios, where the integration of sentiment factors enhances prediction accuracy.

5. Conclusion and Future Work

This study introduces a novel hybrid model for stock price prediction, alongside the creation of a comprehensive Chinese stock sentiment classification dataset. Experimental results show that the performance of machine learning models on stock prediction varies on different market index groups , with best performance on high market values (CSI 100). It also suggest that the integration of sentiment analysis into stock price prediction models generally leads to improved accuracy, although the extent of this improvement varies. The impact of incorporating sentiment analysis is not uniform across all stocks, with noticeable differences based on market value and Book-to-Market (B/M) ratio segments and different market index groups. Intriguingly, for some stocks, the addition of sentiment data has been observed to diminish predictive performance, with such effects being especially marked within the low B/M ratio category.

Market Ind.	eval.	LSTM		Bi-LSTM	
		without	with senti	without	with senti
CSI 100	Total RMSE	6.1054	6.1112	6.1156	**6.0800**
	Total MAPE	25.4933	25.5460	25.8655	**25.4900**
CSI 200	Total RMSE	18.8044	18.2662	**18.2409**	18.5133
	Total MAPE	35.8019	**35.1904**	35.7643	35.9698
CSI 500	Total RMSE	9.5629	**9.3874**	9.5060	9.4294
	Total MAPE	39.5894	**38.9313**	39.3167	39.1492
CSI 1000	Total RMSE	7.3246	7.3893	7.3096	**7.0375**
	Total MAPE	48.7534	47.5120	46.1122	**44.4132**

Table 3: Aggregated performance for stocks from each market index group using different models.

Combination	LSTM	LSTM with senti	Bi-LSTM	Bi-LSTM with senti
Total RMSE	41.3073	41.1638	41.2541	**41.1603**
Total MAPE	148.6382	147.1796	147.0567	**145.5350**

Table 4: The total RMSE and MAPE aggregated by the combined predictive model.

Market value	CSI 100	CSI 200	CSI 500	CSI 1000
Total RMSE	**24.4222**	73.8258	37.8857	28.0510
Total MAPE	**102.4448**	142.7264	157.9866	186.7908

Table 5: Total RMSE and MAPE aggregated by market value.

B/M	LSTM			LSTM with senti		
	High	Medium	Low	High	Medium	Low
Total RMSE	3.8537	13.3996	24.0540	3.8124	12.8506	24.4802
Total MAPE	48.2670	43.9803	56.3907	47.2058	43.2707	56.7031

B/M	Bi-LSTM			Bi-LSTM with senti		
	High	Medium	Low	High	Medium	Low
Total RMSE	3.8493	13.4192	23.9037	3.7799	13.1928	24.2056
Total MAPE	46.5322	43.4357	57.0888	45.3580	43.9159	56.4611

Table 6: RMSE and MAPE for different models across different Book-to-Market ratio groups.

B/M ratio	High	Medium	Low
Pearson	**0.7104**	0.4705	0.1259

Table 7: The correlation coefficients between sentiment values and stock closing prices across different B/M ratio groups.

This research, while providing valuable insights, is subject to certain limitations. The predictive outcomes detailed in this study are derived solely from the context of the Chinese stock market and have not been tested across diverse market environments. The specific attributes of China's market, such as the absence of same-day buying and selling (T+0 trading), could potentially skew the applicability of our findings to other financial contexts. In our forthcoming efforts, we plan to broaden the scope of our investigation by integrating a wider array of sentiment analysis methodologies and including additional external market variables. This expansion aims to enhance the robustness and generalizability of our results, ensuring that our conclusions hold weight across varying global market dynamics.

6. Acknowledgements

We would like to acknowledge the support provided by the XJTLU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (No. YZCXPT2022103) and is also supported by the Research Development Funding (RDF) (No. RDF-21-02-044) at Xi'an Jiaotong-Liverpool University. Additionally, this research has received partial funding from the Jiangsu Science and Technology Programme (contract number BK20221260).

7. Bibliographical References

Agustin Garcia Asuero, Ana Sayago, and AG González. 2006. The correlation coefficient: An overview. *Critical reviews in analytical chemistry*, 36(1):41–59.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Damares CO de Resende, Ádamo Lima de Santana, and Fábio Manoel França Lobato. 2016. Time series imputation using genetic programming and lagrange interpolation. In *2016 5th Brazilian conference on intelligent systems (BRACIS)*, pages 169–174. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Dewally. 2003. Internet investment advice: Investing with a rock of salt. *Financial Analysts Journal*, 59(4):65–77.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Joshua Zoen Git Hiew, Xin Huang, Hao Mou, Duan Li, Qi Wu, and Yabo Xu. 2019. Bert-based financial sentiment index and lstm-based stock return predictability. *arXiv preprint arXiv:1906.09024*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jieyun Huang, Yunjia Zhang, Jialai Zhang, and Xi Zhang. 2018. A tensor-based sub-mode coordinate algorithm for stock prediction. In *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*, pages 716–721. IEEE.

Weiwei Jiang. 2021. Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications*, 184:115537.

Nan Jing, Zhao Wu, and Hefei Wang. 2021. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. *Expert Systems with Applications*, 178:115019.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Jiexi Liu and Songcan Chen. 2019. Non-stationary multivariate time series prediction with selective recurrent neural networks. In *Pacific rim international conference on artificial intelligence*, pages 636–649. Springer.

Yuandong Luan and Shaofu Lin. 2019. Research on text classification based on cnn and lstm. In *2019 IEEE international conference on artificial intelligence and computer applications (ICAICA)*, pages 352–355. IEEE.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jeffrey Pontiff and Lawrence D Schall. 1998. Book-to-market ratios as predictors of market returns. *Journal of financial economics*, 49(2):141–160.

Thendo Sidogi, Rendani Mbuvha, and Tshilidzi Marwala. 2021. Stock price prediction using sentiment analysis. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 46–51. IEEE.

Justin Sirignano and Rama Cont. 2021. Universal features of price formation in financial markets: perspectives from deep learning. In *Machine Learning and AI in Finance*, pages 5–15. Routledge.

Md Arif Istiake Sunny, Mirza Mohd Shahriar Maswood, and Abdullah G Alharbi. 2020. Deep learning-based stock price prediction using lstm and bi-directional lstm model. In *2020 2nd novel intelligent and leading emerging sciences conference (NILES)*, pages 87–92. IEEE.

Daigo Tashiro, Hiroyasu Matsushima, Kiyoshi Izumi, and Hiroki Sakaji. 2019. Encoding of high-frequency order information and prediction of short-term stock price by deep learning. *Quantitative Finance*, 19(9):1499–1506.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qili Wang, Wei Xu, and Han Zheng. 2018. Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. *Neurocomputing*, 299:51–61.

Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. 2018. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1627–1630.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Liheng Zhang, Charu Aggarwal, and Guo-Jun Qi. 2017. Stock price prediction via discovering multi-frequency trading patterns. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2141–2149.

A. Appendix

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
601818.SH	0.0723	2.0003	0.0644	1.7618	0.0717	1.9621	0.0597	1.6164
601998.SH	0.1733	3.1535	0.2213	3.5069	0.1948	3.4679	0.2287	3.6645
600999.SH	0.4904	3.0828	0.4966	3.1254	0.5142	3.2260	0.5008	3.1497
002736.SH	0.3146	2.8587	0.3118	2.8250	0.3113	2.8343	0.3041	2.7722
600585.SH	3.0708	8.6040	3.0450	8.5696	2.9508	8.3421	2.9175	8.2399
601336.SH	1.9840	5.7940	1.9721	5.7572	2.0728	6.0331	2.0792	6.0473

Table 8: RMSE and MAPE of the predicted results for stocks selected in CSI 100.

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
000783.SH	0.2051	3.0481	0.2018	3.0007	0.2098	3.1069	0.1990	2.9555
600741.SH	1.7092	7.5582	1.6745	7.3673	1.7176	7.5955	1.7031	7.5349
600085.SH	3.5355	5.8289	3.4708	5.7213	3.4208	5.4714	3.3259	5.6429
601021.SH	3.2210	5.0601	2.8503	4.5878	3.4136	5.3472	3.3309	5.2134
300144.SH	0.9565	6.2332	0.9268	6.0339	0.9763	6.3349	0.9507	6.1798
300033.SH	8.6771	8.0734	9.1420	8.4794	8.5029	7.9084	9.0037	8.4033

Table 9: RMSE and MAPE of the predicted results for stocks selected in CSI 200.

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
000488.SH	0.3031	4.9292	0.2988	4.8500	0.2899	4.7048	0.2989	4.8091
600657.SH	0.7743	12.3015	0.7668	12.1422	0.7774	12.3377	0.7628	12.0320
000685.SH	0.4096	4.8377	0.4098	4.8191	0.4163	4.8653	0.4263	5.0726
300244.SH	1.7188	4.6935	1.6800	4.5999	1.7034	4.6744	1.7232	4.6904
603355.SH	1.8167	4.4031	1.7429	4.2319	1.8246	4.4261	1.7819	4.3314
600259.SH	4.5404	8.4244	4.4691	8.2882	4.4944	8.3084	4.4364	8.2137

Table 10: RMSE and MAPE of the predicted results for stocks selected in CSI 500.

Stock ID	LSTM		LSTM with senti		Bi-LSTM		Bi-LSTM with senti	
	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
000797.SH	0.4968	10.7736	0.4784	10.4666	0.4758	9.0313	0.4157	8.6981
601588.SH	0.1196	4.5026	0.1064	4.1103	0.1123	4.3260	0.1040	3.8075
002138.SH	3.4924	12.1558	3.4245	11.9604	3.4415	12.0183	3.3665	11.7765
002542.SH	0.2173	5.4628	0.2274	5.6318	0.1981	4.9988	0.2051	5.0834
603989.SH	2.1708	6.5909	2.3174	7.0524	2.1710	6.6277	2.2018	6.6794
300377.SH	0.8377	8.2677	0.8352	8.2905	0.9109	9.1081	0.8344	8.3663

Table 11: RMSE and MAPE of the predicted results for stocks selected in CSI 1000.

B/M	LSTM			Bi-LSTM		
	High	Medium	Low	High	Medium	Low
Total RMSE	7.7366	26.2708	48.5342	7.6661	26.6120	48.1093
Total MAPE	95.4728	87.2510	113.0938	91.8902	87.3516	113.5499

Table 12: Total RMSE and MAPE for different models across different Book-to-Market ratio groups.

Topic Taxonomy Construction from ESG Reports

Saif Alnajjar[1], **Xinyu Wang**[2], **Yulan He**[1,3]
[1]King's College London [2]University of Warwick [3]The Alan Turing Institute
saif.alnajjar@kcl.ac.uk, Xinyu.Wang.11@warwick.ac.uk, yulan.he@kcl.ac.uk

Abstract

The surge in Environmental, Societal, and Governance (ESG) reports, essential for corporate transparency and modern investments, presents a challenge for investors due to their varying lengths and sheer volume. We present a novel methodology, called MultiTaxoGen, for creating topic taxonomies designed specifically for analysing the ESG reports. Topic taxonomies serve to illustrate topics covered in a corpus of ESG reports while also highlighting the hierarchical relationships between them. Unfortunately, current state-of-the-art approaches for constructing topic taxonomies are designed for more general datasets, resulting in ambiguous topics and the omission of many latent topics presented in ESG-focused corpora. This makes them unsuitable for the specificity required by investors. Our method instead adapts topic modelling techniques by employing them recursively on each topic's local neighbourhood, the subcorpus of documents assigned to that topic. This iterative approach allows us to identify the children topics and offers a better understanding of topic hierarchies in a fine-grained paradigm. Our findings reveal that our method captures more latent topics in our ESG report corpus than the leading method and provides more coherent topics with comparable relational accuracy.

Keywords: Text Mining, Text Analytics, Document Classification, Text categorisation, Knowledge Discovery/Representation, Topic Detection and Tracking

1. Introduction

Environmental, Societal, and Governance (ESG) reports are a type of report that companies release to discuss their plans and performance in, as the name suggests, *environmental*, *societal*, and *governance* issues. As the world shifts towards transparency and accountability, ESG reports serve as an indispensable resource for stakeholders, especially given the dramatic 27-fold increase in socially responsible investing (SRI) assets over 25 years (Christiansen et al., 2023).

However, with the rising importance of ESG reporting, as well as a recent EU directive that mandates larger European companies to publish ESG reports, there has been a significant upswing of companies issuing such reports, which can vary in lengths, spanning from a few pages to several hundred pages. The proliferation of ESG reports poses a challenge for investors who need to review them when making investment decisions.

As such, one analytical approach that can help investors and consumers is the creation of a topic taxonomy for a collection of ESG reports. A topic taxonomy is a hierarchical structure that displays the relationship between topics within a corpus. Each topic could serve as a parent to one or more subtopics, forming a structured hierarchy. Figure 1 shows an example of what a topic taxonomy looks like. Within each topic, a list of relevant terms represents the overarching concept, and a primary term is selected from that list to represent the topic in the taxonomy.

However, current state-of-the-art taxonomy methods, namely TaxoCom (Lee et al., 2022a), are often tailored for more general datasets, and as such falter with the distinct nuances of ESG reports. As depicted in Table 4, their extracted topics often emerge ambiguous or overly broad, missing many of the latent topics in the corpus, making the result barely usable for investors, who usually prefer a much deeper level of information. Not only that, these methods also use a phrase mining tool, like AutoPhrase (Shang et al., 2017), to get a list of potential terms, and thus some terms that are relevant but in lower frequency are missed, while at the same time non-ESG terms are also mined, creating some noise and worsening the results.

Recognising these limitations, we propose a novel method, called MultiTaxoGen, that leverages topic modeling techniques to better capture the intricacies of ESG reports, and heavily adapt and optimise them for building a topic taxonomy for our corpus. The main idea is to recursively run the topic modeling technique on every topic's local neighbourhoods based on the idea of local embeddings used in previous topic taxonomy works (Lee et al., 2022a; Shang et al., 2020; Zhang et al., 2018), to find its subtopics. Local neighbourhoods refer to the subcorpus of documents that were assigned to the current topic.

We modify the topic modeling technique to suit each level of the taxonomy to find more generalised topics in the second-level, and more specific and focused topics in the bottom-level. Unfortunately, these topic modeling techniques, in general, have no hierarchical understanding of our topics, so we create embeddings for the topics and compare them in the taxonomy and remove any deemed as outliers or redundant. We also improve the efficacy of assigning documents at the top-level by using an ESG classifier, giving better results downstream

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 178–187
20 May, 2024.

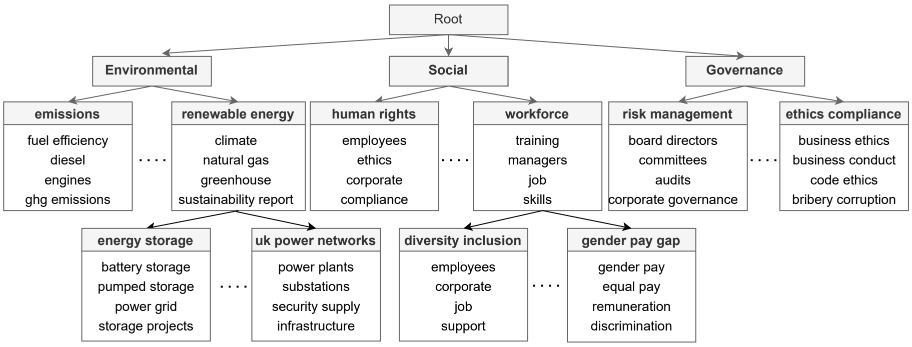


Figure 1: A sample from our constructed three-level topic taxonomy, featuring the top-level topics along with their respective subtopics, and the terms associated with those subtopics.

due to less documents being misassigned to the incorrect local neighbourhood.

Our main contributions are two-fold:

- We introduce a three-level framework for ESG reporting taxonomy. At each level, we employ tailored strategies adapted to the specific text and topic granularity.
- We conduct comprehensive experiments and evaluations of our method, including human assessments. The experimental results show that our method captures more latent topics than the leading method and provides more coherent topics.

2. Related Work

Topic Modeling Topic taxonomy construction and topic modeling are, naturally, very similar, and so a lot could be learned from topic modeling, especially since it is a widely studied field with many methods being researched (Blei et al., 2003; Angelov, 2020; Grootendorst, 2022; Bianchi et al., 2021). The two most common types of methods are statistical and neural topic models. The most widely used method, Latent Dirichlet Allocation (LDA) (Blei et al., 2003), is one such example of a statistical method. Early hierarchical topic modelling approaches built on LDA (Blei et al., 2010; Kim et al., 2012) were proposed for the discovery of topical hierarchies within the abstracts of scientific papers. In such models, each document is presumed to be linked to a path where each level represents a topic. The assignment of paths adheres to an nested Chinese Restaurant Process (nCRP) or recurrent CRP prior. Additionally, Paisley et al. (2014) proposed a nonparametric model called the nested hierarchical Dirichlet process, enabling the incorporation of shared groups among clusters, thus extending the capabilities of the nCRP model through the incorporation of a hierarchical Dirichlet process. The primary challenges associated with hierarchical topic modeling methods include the complexity of incorporating prior knowledge about topics and their dependency on having access to the complete vocabulary of the corpus.

More recent literature, however, suggests that neural topic models, namely those that use embedding techniques like BERT (Devlin et al., 2019) or Word2Vec (Mikolov et al., 2013), can outperform the classic topic modeling techniques, with BERTopic (Grootendorst, 2022) and CTM (Bianchi et al., 2021) being some examples.

Topic Taxonomy Generation Topic taxonomy generation primarily follows two approaches: from scratch and seed-guided. The former constructs taxonomies without any prior knowledge of the taxonomy and just relying on the corpus. Seed-guided, a more weakly supervised approach, uses an initial seed taxonomy in addition to the corpus to nudge the generated topics towards that seed. Currently, most of the highest-performing methods in either approach rely on what they call “local embeddings” (Lee et al., 2022a; Shang et al., 2020; Zhang et al., 2018). To improve the granularity of the embedding space when adding children topics to a parent topic, we create a subcorpus of documents that are related to that parent topic, and train an embedding, like Word2Vec (Mikolov et al., 2013), on that subcorpus, instead of using a global embedding that was trained on the entire corpus for all the children to be added (Lee et al., 2022a). Since the documents in the subcorpus are clustered to find the new subtopics, having different subcorpora for each of the topics can make the embeddings more discriminative and ultimately improve results. One other promising seed-guided approach is TopicExpan (Lee et al., 2022b), which out-performs all the other taxonomy generation method, but is a supervised method that requires all the documents in the corpus to be labelled with a term and topic related to that document.

ESG Baier et al. (2020) develops a word list for ESG topics and a corresponding taxonomy, then analyzes the distribution of these topics in ESG reports to determine their prevalence. This expert-curated taxonomy is valuable as it gives us a good starting point for the seed we will be using. Meanwhile, FinBERT (Huang et al., 2023) further pretrains BERT on a corpus of financial documents, improving its performance in the financial domain. Most relevant to us though, the authors fine-tune FinBERT for classifying a document as Environmental, Social, and Governance, achieving state-of-the-art performance for ESG classification.

3. Preliminary - BERTopic

Our proposed approach is built on BERTopic (Grootendorst, 2022). Throughout this paper, we choose the topic modeling method of BERTopic (Grootendorst, 2022) as our primary focus, alongside corresponding experiments. In this section, we give an overview of the BERTopic method, which consists of three steps: document embedding generation, document clustering, and topic term extraction.

Document Embedding Generation First, document embeddings are generated using a language model. A common choice for this task is the Sentence Transformers (Reimers and Gurevych, 2019), which have been fine-tuned for document embedding generation. Following this, dimensionality reduction is performed on the embeddings using techniques such as Principal Component Analysis (PCA) or Uniform Manifold Approximation Projection (UMAP) (McInnes et al., 2020). Dimensionality reduction accelerates the model and also mitigates the curse of dimensionality (Keogh and Mueen, 2017), prior to proceeding with the subsequent step of the pipeline, clustering.

Document Clustering BERTopic then clusters the reduced document embeddings, and each cluster would thus count as a topic. Clustering is of particular importance for the topic taxonomy generation, as changes in the cluster size and clustering algorithm can allow for either more specific or more general topics.

Topic Term Extraction The final step in the pipeline is to extract the top terms of each topic based on the class-specific TF-IDF scores, or c-TF-IDF. To do this, all the documents in a cluster are combined to form a single document and a term-document matrix is formed for all the newly created documents. Then, the c-TF-IDF score of a term **w** in a cluster **c** is calculated using Equation (1).

$$score_{w,c} = tf_{w,c} \times \log(1 + \frac{n\_avg\_words}{tf_w}) \quad (1)$$

The highest scoring words/terms, usually the top 10, are thus used to represent the topic. The scoring mechanism naturally favour terms that appear frequently in a certain cluster while being less common in others. Thus, in the case of larger clusters that encompass more documents, the scoring tends to emphasise more general or overarching terms, as one would anticipate in higher levels of a taxonomy. On the other hand, when the clustering algorithm is forced to generate as many clusters as possible, leading to smaller clusters that ultimately represent all potential topics, the highest scored terms tend to be more specific and focused.

4. Methodology

We propose a multi-level topic taxonomy generation approach, named as MultiTaxoGen, as shown in Figure 2. At the first level, documents are segregated into three main topics: *Environmental*, *Social*, and *Governance*. We deploy a classifier to partition all documents to each of those topics and split the corpus into three distinct subcorpora. Next, on each subcorpus, we utilise a topic modeling technique, specifically BERTopic (Grootendorst, 2022), to search for a small number of topics. However, it is pertinent to note that alternative neural topic modelling techniques are also available, such as Top2Vec (Angelov, 2020).

This methodology echoes the ideas propounded by preceding studies (Zhang et al., 2018; Shang et al., 2020; Lee et al., 2022a), where a local embedding is trained on a topic-specific subcorpus. Our approach, instead, involves operating BERTopic on what can instead be called a local neighbourhood of documents rather than training a local embedding. The topics found from each of the subcorpora would thus constitute the second level of our taxonomy. Then, BERTopic is rerun on the documents under each newly discovered topic, allowing BERTopic to find as many topics as possible. Finally, redundant or unrelated topics are then merged or removed respectively. This would thus establish the third and bottom level of the topic taxonomy.

4.1. Local Neighbourhoods

Previous works (Zhang et al., 2018; Shang et al., 2020; Lee et al., 2022a) on creating topic taxonomies have found great success in training a local embedding for each sub-corpus of documents in a topic, rather than using one global embedding training on the entire corpus, allowing for better granularity and discriminativeness between embeddings, and ultimately improved performance when finding subtopics.

Rather than training our own embeddings, which would require a massive corpus for training transformers, we simply run BERTopic separately for each topic's subcorpus, and the child topics found would be more tuned towards the parent topic with

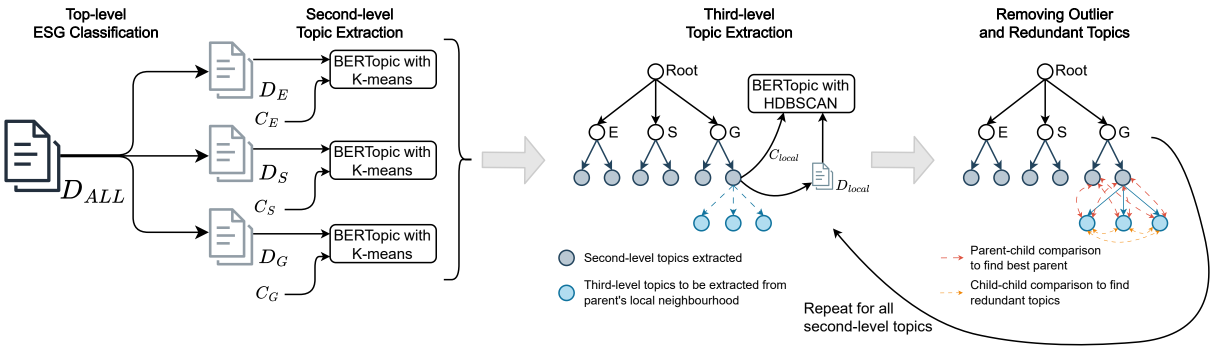


Figure 2: An illustration of MultiTaxoGen, where D refers to a corpus of documents and C is a set of topics. In the context of this illustration, C is the seed topics. Initially, an ESG Classifier split documents in our full corpus into three topics. Subsequently, the second-level topics are extracted on the corpus of each of them. Finally, we use BERTopic on each second-level topic with HDBSCAN to find their children topic. We further remove outlier topics with the parent-child comparisons to find the correct parent, and merge redundant topics by comparing them with the other children and then merging them.

the c-TF-IDF scoring, along with some modifications and optimizations.

4.2. First Level

The top-level topics of any ESG report will, naturally, be *Environmental*, *Social*, and *Governance*. Splitting our corpus into three separate subcorpora gives us the advantage of having more focus on subcorpus for each of the topics when we find the subtopics in the subsequent steps. To facilitate this division, we employ the FinBERT-ESG (Huang et al., 2023) classifier to assign all the documents into either one of the three subtopics or a"none" class if they lack relevance to any of the primary topics. It was reported in (Huang et al., 2023) that the classifier achieves an accuracy of 89.5% on a small set of ESG-related discussions. We then filter documents which have a probability of less than a threshold $\tau_c = 0.7$. The majority of the filtered documents consist of tables and numerical data found in report appendices, which fall outside the scope of our primary focus, or irrelevant documents that can lead to non-ESG related topics being extracted.

4.3. Second Level

To find the second-level topics, i.e. the children of the top-level topics, *Environmental*, *Social*, and *Governance*, we run BERTopic on each of the top-level topic's subcorpus, while using k-means as the clustering algorithm to guide BERTopic in identifying a limited set of clusters by setting the number of clusters k to a small value.

The goal of this is to create large clusters with many documents that discuss many different topics, but all share a certain high-level topic in each cluster. Thus, the highest scoring terms will be those that match that high-level topic and will typically be more general and less focused, while also being inherently related to the parent topic, since they are derived from documents assigned to their parent.

To nudge the generated topics towards our seed, BERTopic takes in a seed of topics with their potential terms and then steers the c-TF-IDF scoring of the terms in the clusters towards those seed topics by applying a multiplier to the score if a term is related any of the seed topics. As a seed, we use the curated ESG topic taxonomy created by Baier et al. (2020), and also remove any topics that rarely appear in their corpus. This approach ensures that the top terms chosen have a strong relevance to ESG topics.

A main term will also need to be selected to represent that topic in the taxonomy, and for our case, we simply select the highest scoring term in the cluster as the main term, as that term typically represents the topic.Table 1 shows some second-level extracted topics, showcasing how the topics clearly relate to the parent topic, *Environmental*.

	Topic's Top 3 Terms	Main Term
T1	sustainability report, environmental management, ghg emissions	sustainability report
T2	water consumption, wastewater, groundwater	water consumption
T3	waste management, recycling, hazardous waste	waste management
T4	greenhouse gas emissions, scope emissions, energy consumption	greenhouse gas emissions

Table 1: Sample second-level topics, their terms, and their main terms for the parent and top-level topic *Environmental*

.

4.4. Third Level

Topics in the third or bottom level are expected to be more specific compared to their parent's. As such, rather than forcing it to find a certain arbitrary number of clusters like in the previous level, finding as many topics as possible is a more optimal alternative, as all the latent topics need to captured at this level. Therefore, we employ Hierarchical DBSCAN (HDBSCAN) (Malzer and Baum, 2020), which in itself is an extension of the popular DBSCAN (Ester et al., 1996), to find all potential clusters of all different sizes. We set the minimum cluster size as the minimum document count required for a topic to be formed in order to modify the number of topics that are found, and this number is based on the size of the second-level topic's local neighbourhood.

Compared to the topics in the previous level, we cannot predetermine a list of potential seed topics for the third level due to the variability inherent in the second-level topics. Instead, we take the top 5 scoring terms of the parent topic as the seed topics when identifying the subtopics of that parent. The highest scoring terms of each topic's top terms are then again selected as the main term. However, due to the high number of topics in the bottom level, many topics may share the same highest-scoring term. In this case, the next highest-scoring term that is not the main term of any other topic is selected as the main term.

One issue in extracting a large number of topics, especially when using HDBSCAN and considering misassigned documents, is the emergence of irrelevant and redundant (where two or more topics can be very similar or even exactly the same as each other) topics. Additionally, topic modeling methods such as BERTopic do not take any hierarchy into account, other than us applying it to a local neighbourhood of documents of a parent topic. It consequently cannot discern when an irrelevant topic is extracted. To address these issues, further optimisations are required to remove or merge the unnecessary topics.

As a final note, we could, in theory, repeat this same process again with the third-level to get a fourth-level, however we opted to stop at three levels for several reasons. At the third level, topics become exceedingly specific, making it challenging to extract meaningful latent topics from the documents associated with third-level topics, as they often revolve around very similar subject matter. Additionally, for the sake of consistency in comparisons with other methods, a three-level taxonomy appears to be more suitable, as the majority of related papers on topic taxonomy construction primarily employ two or three-level hierarchies

Topic Embeddings Generation To determine the necessity of a topic, we convert the topics into embeddings to properly compare different topics. We represent the top 5 terms of a topic as its embedding by using any word embeddings methods. However, using context-independent word embeddings such as GloVe (Pennington et al., 2014), leads to the out-of-vocabulary problem, and multi-word terms would need to be found using the less-than-ideal workaround of calculating the average embedding of their words. Therefore, we instead employ the Sentence Transformers (Reimers and Gurevych, 2019), the same embedding model used to represent our documents. Even though not explicitly trained for this task, the embeddings generated by it are still fairly good and better than the ones found using GloVe embeddings. It also brings the added benefit of being able to directly compare documents with the topic embedding. when assigning relevant topics to documents. We average the embeddings of the topic's top 5 terms, generated by the embedding model, based on the term's score, where the highest-scoring terms hold a higher weight.

Removing Outlier Topics To minimize the number of unrelated topics, we initially check the generated topic embeddings for the third-level topic, as well as embeddings for the second-level topics, including its parent, by comparing the cosine similarity of the third-level topic with each of the second-level topics. If the second-level topic most similar to a third-level topic does not correspond to its parent, the third-level topic is removed, and its associated documents are temporarily marked as outliers. This procedure is reiterated for all the third-level topics.

Merging Redundant Topics Similarly, the redundant topics should be removed as well by merging all the redundant topics into one topic. We first generate the topic embeddings with the embedding model for each of the third-level topics of one of the second-level topics. Next, each third-level topic embedding is compared with all the other third-level topic embeddings by their cosine similarity, creating a similarity matrix. If the similarity between one topic embedding and another is greater than a threshold τ_r, then those topics are merged, meaning the topics are combined by putting the documents of each of the two topics into one topic. In our case, we set $\tau_r = 0.8$ as that was found to be the optimal value in the experiments. The threshold in our case was set to a high value because most of the topics found at the bottom level will have a high similarity score between them since those topics are inherently similar as they share the same parent topic. Finally, we repeat this process for the third-level topics of all the second-level topics separately.

4.5. Assigning Documents to Multiple Topics

Normally, with BERTopic, documents are assigned to only one single topic, rather than all relevant topics. As such, after finding all the topics and creating the topic taxonomy, we attempt to identify each topic within the taxonomy that holds relevance to a specific document.

The first step is the identification of which top-level topic a document is associated with via the ESG classifier, where we assume that only one top-level topic is in each document, as almost documents are focused on only one of the top-level topics. For documents categorized under top-level topics, we identify their corresponding second-level topics using the BERTopic models that have been previously trained for the respective top-level topics.

Then, the remaining potential topics are found by calculating the cosine similarity between the embeddings of documents in the top-level topics and their children (second-level) topic embeddings. Each document's topic assignment is determined by a threshold τ_t. However, a single universal threshold may not be optimal for all topics. Therefore, we designate distinct thresholds τ_t tailored to each specific topic by calculating:

$$\tau_t = \mu_s + (1.5 \times \sigma_s) \tag{2}$$

where μ_s and σ_s denotes the mean of similarities of the second-level topics and is their standard deviation, respectively. Finally, we classify the documents to the third-level topics. For the documents assigned to the parent (second-level) topics, we again compute the cosine similarity with the third-level topic embeddings and find the relative thresholds as described above. To be noticed that a document can also be assigned to more than one second-level topic and thus be compared and checked multiple times. Third-level topics differ slightly from second-level topics in that a document could potentially not be assigned any third-level topics, as HDBSCAN may mark a document as an outlier if it does not match with any topics.

5. Experiments

5.1. Experimental Setup

Dataset We collect 10,645 publicly available ESG reports released from 1992 to 2022 across 2,001 companies from ResponsibilityReports.com[1]. The reports are in PDF format, and we extract text content from ESG reports using PyMuPDF[2]. Next, we split, as best as we can, the reports into paragraphs of a maximum length of 256 words to constitute a total of 1,208,546 documents after splitting, and these would be considered the documents of our corpus.

Splitting them into paragraphs shorter than 256 words is necessary, as BERT models typically have a maximum length of 512 tokens. Also, while a document may encompass various subjects, when we assign a document to a specific topic (as a part of its local neighborhood), we assign it to only the topic most relevant to it, as was done in previous works (Lee et al., 2022a). Naturally, shorter documents will end up having less topics being discussed in them.

Baselines We will be comparing our method with the current state-of-the-art weakly-supervised method, TaxoCom (Lee et al., 2022a). We split our corpus into three subcorpora for each of the top-level topics, and run TaxoCom separately for each of those.

Hyperparameter Setup The document embedding model we use is MiniLM (Wang et al., 2020) comprising 6 layers trained in accordance with the Sentence Transformers (Reimers and Gurevych, 2019) paradigm. The number of clusters k topics, when extracting the second-level topics, for the three top-level topics are $k_{environmental} = 8$, $k_{social} = 8$, and $k_{governance} = 5$. For TaxoCom (Lee et al., 2022a), we set $\beta_1 = 3.5$ and $\beta_2 = 6.0$ for the second and third levels respectively, which controls how many novel topics are found, and keep the other parameters the same as in the paper.

5.2. Evaluation

Considerable research has been done to try and automatically evaluate topic coherence. Some measures have been shown to correlate with humans quite well (Lau et al., 2014) and are commonly used when evaluating topic models, namely NPMI (Bouma, 2009) and C_v (Röder et al., 2015). Other works, however, suggest that although classical topic models like LDA (Blei et al., 2003) do correlate, they may not do so with neural topic models (Hoyle et al., 2021). Ultimately, we opted for human evaluation to get the most accurate results, but have included the C_v scores in the results as well. In particular, we employ two metrics to compare the methods, *topic coherence* and *relation accuracy*. The evaluations of these metrics has been carried out by 3 computer science graduates, who were paid hourly rate of £20 for their evaluation of the methods. We then average out their results to minimize human bias.

Topic Coherence The first metric tries to measure how "coherent" a topic is by how clearly the

[1] https://www.responsibilityreports.com/

[2] https://pymupdf.readthedocs.io/

	Total Number of Topics		Topic Coherence		C_v		Relation Accuracy	
	MultiTaxoGen	TaxoCom	MultiTaxoGen	TaxoCom	MultiTaxoGen	TaxoCom	MultiTaxoGen	TaxoCom
Environmental	174	12	**0.902**	0.731	0.521	**0.526**	0.865	**0.910**
Social	272	23	**0.910**	0.732	0.533	**0.715**	0.831	**0.917**
Governance	51	22	**0.917**	0.790	0.407	**0.711**	**0.895**	0.883
All	500	60	**0.908**	0.754	0.487	**0.651**	0.850	**0.899**

Table 2: Results of our method, MultiTaxoGen, and TaxoCom. Best results in each taxonomy are in bold.
.

Parent Topic	Outlier Topic's Top 3 Terms	Most Similar Parent
renewable energy	**air filters**, indoor air quality, air filtration	greenhouse gas emissions
charity	**consumer credit**, credit history, experian	local communities
corporate governance	**auditing standards**, auditor report, statutory sustainability	audit committee

Table 3: Examples of outlier topics found, their parent topic, and the parent topic that found to be most similar to them. Main term of the outlier topic is highlighted in bold.
.

set of terms in a topic represents a recognisable overarching topic or category (Lund et al., 2019). By definition, this is inherently a subjective measure, as one person may see a certain set of terms as more coherent compared to another person that may see those set of terms to have a different meaning. For the human evaluation, the topic coherence score of each topic is calculated by counting the number of terms in the topic that do belong in that topic, and are then averaged. Next, all the topic coherence scores are averaged as well to get the average topic coherence of the method.

Relation accuracy Relation accuracy tries to evaluate the accuracy of the relationships among the child, parent, and grandparent topics. This is also human evaluated. To find the relation accuracy of a topic, the topic is compared to its parent. If the parent-child relationship is correct, it is given a score of 1. However, if they do not match but the child matches the grandparent, it is given a score of 0.25 instead. If it does not match any of them, then the relation accuracy of that topic is 0. The final relation accuracy of the method is found by averaging the accuracy of all the topics.

5.3. Quantitative Results

The results have been split into the three taxonomies for each of the top-level topics so we provide a deeper look into the results. Table 2 shows a comparison of the results. We can observe that our method found a much larger number of topics, almost ten-fold, compared to TaxoCom, which managed to only find a total of 60 topics. Considering our corpus of more than a million documents, 60 topics do seem to be considerably lower than expected. As we show later, the topics extracted by TaxoCom also are more vague, even in the lower levels. In contrast, our method manages to extract more specific topics at the bottom level.

Our method gives significantly more coherent topic across the board, achieving an average topic coherence of above 0.9. Conversely, TaxoCom achieves a higher relation accuracy, though in all taxonomies, the results are still close, with only a small difference between the two methods. TaxoCom also achieves a higher C_v score, though it is important to be noticed the limitations of automatic topic coherence measures, as described in Section 5.2

5.4. Case Study

We present examples to explore the effects of different parts of our method, as well as the results from both our final taxonomy and TaxoCom's.

Outlier Topics One can anticipate the emergence of irrelevant topics when employing a topic modeling technique, particularly due to their lack of capability to identify hierarchical relationships within a taxonomy. Consequently, if documents unrelated to the parent topic are found in its subcorpus, it could result in the formation of a cluster for those documents, thus generating an outlier topic. Our approach, described in Section 4.4, enhances the model's understanding of the hierarchy, enabling it to detect and remove any outlier topics that do not align with their parent topic. Approximately 54% the initial topics were subsequently removed. Table 3 showcases some of these outlier topics, which were subsequently flagged and removed.

Parent Topic	Sub-topics Generated by MultiTaxoGen	Sub-topics generated by TaxoCom
Renewable Energy	natural gas, greenhouse gas (⊗), sustainable development, red electrica, uk power networks, oil sands, mining (⊗), cenovus, demand response programs, energy storage systems...	vehicles, environmental (⊗)
labor	diversity inclusion, benefits, social responsibility (⊗), innovation, employability, gender paygap, employees diversity board (✂), skills (⊗), best employers...	sony group (⊗), diversity, rights, forced labor
ethics	business ethics, anti money laundering, business integrity, concerning marketing communications, financial industry laws, anti competitive behavior, ownes curning sustainability (⊗), modern slavery act	corruption

Table 4: Comparison between MultiTaxoGen and TaxoCom. Redundant topics are marked with (✂) and incorrect topics are marked with (⊗)

.

Redundant Topics When delving deeper into a taxonomy, topics become more specific when using a subcorpus derived from their parent topic. Due to the shared parent topic, documents within this subcorpus are closely related, posing challenges in distinguishing between child topics and resulting in the emergence of many similar child topics that may need merging. Table 5 highlights some redundant topics that are merged once they exceed the threshold τ_r. This led to a further reduction of 48% in the number of topics, thereby indicating that after removing the redundant and outlier topics, the total number of initial topics found by our method was reduced by approximately 75%.

Parent Topic		Child Topic's Top 3 Terms	Similarity
greenhouse gas emissions	T1	**emission reductions**, emissions kt, kyoto protocol	0.882
	T2	**emissions reducing**, reduce carbon, ghg emissions	
human rights	T1	**human rights assessment**, training human rights, rights policies	0.850
	T2	**rights assessment**, supplier human rights, grievance	

Table 5: Examples of redundant topics found (denoted as T1 and T2), their parent topic, and the similarities between the topics to be merged. Main term of child topic is highlighted in bold.

.

Generated Topics Comparison A major issue that we've observed with TaxoCom and other state-of-the-art topic taxonomy construction methods is that they have been designed to work on much more general datasets and are hard to optimise for specific tasks, such as analyzing ESG reports. Also, TaxoCom struggles to identify the majority of topics within larger corpora, and those it does identify often prove ambiguous or overly broad, even after tuning the novelty parameter β to enhance topic discovery. In contrast, our approach, as demonstrated in Table 4, notably discovers more topics in our corpus. Figure 1 shows a small sample of a taxonomy constructed by our method.

5.5. Discussion

We have introduced a novel method called MultiTaxoGen for creating a topic taxonomy from ESG reports. Our compiled corpus consists of ESG reports with varying styles, formats, and lengths, ranging from a few pages to several hundred pages, all originally in PDF format. The use of an existing PDF parser introduced some text inaccuracies, adding complexity to the data. To enhance the effectiveness of constructing a topic taxonomy, we divided the text documents into 256-word segments. Ideally, the segmentation of ESG reports should be based on their actual content. Future research could explore discourse relations and topic transitions to improve document segmentation. Our preprocessed corpus comprises more than 1 million documents, making it challenging to directly apply existing hierarchical topic models for taxonomy construction due to the extensive computational time required and the difficulty in controlling the topic quality. The current leading approach, TaxoCom, only managed to identify a total of 60 topics, missing many salient ones. In contrast, our approach has the capability to uncover more nuanced topics.

6. Conclusions

We have proposed a novel method optimized for analyzing ESG reports though topic taxonomy construction, addressing limitations in existing methods for complex ESG reports. For future work, we can explore the use of Large Language Models (LLMs) to generate more suitable main term for the topics. Although we tested ChatGPT (Brown et al., 2020) for this purpose, the results were unsatisfactory and the generated main term were generally incorrect. However, it is worth noting that LLMs are rapidly improving, with new models constantly being produced. We anticipate that new LLMs equipped with enhanced capabilities may yield more accurate and contextually relevant main terms for topics, making them valuable tools for future research in topic taxonomy construction.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council through a Turing AI Fellowship (EP/V020579/1, EP/V020579/2).

7. Bibliographical References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Philipp Baier, Marc Berninger, and Florian Kiesel. 2020. Environmental, social and governance reporting in annual reports: A textual analysis. *Financial Markets, Institutions & Instruments*, 29(3):93–118.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M Blei, Thomas L Griffiths, and Michael I Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Charlotte Christiansen, Thomas Jansson, Malene Kallestrup-Lamb, and Vicke Noren. 2023. Households' investments in socially responsible mutual funds. *The Quarterly Review of Economics and Finance*, 87:46–67.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033. Curran Associates, Inc.

Allen H. Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text*. *Contemporary Accounting Research*, 40(2):806–841.

Eamonn Keogh and Abdullah Mueen. 2017. *Curse of Dimensionality*, pages 314–315. Springer US, Boston, MA.

Joon Hee Kim, Dongwoo Kim, Suin Kim, and Alice Oh. 2012. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 783–792.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022a. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. *CoRR*, abs/2201.06771.

Dongha Lee, Jiaming Shen, Seonghyeon Lee, Susik Yoon, Hwanjo Yu, and Jiawei Han. 2022b. Topic taxonomy expansion via hierarchy-aware topic phrase generation.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. IEEE.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

John Paisley, Chong Wang, David M Blei, and Michael I Jordan. 2014. Nested hierarchical dirichlet processes. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):256–270.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2017. Automated phrase mining from massive text corpora.

Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of The Web Conference 2020*, WWW '20, page 1908–1919, New York, NY, USA. Association for Computing Machinery.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *ArXiv*, abs/2002.10957.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. *CoRR*, abs/1812.09551.

Duration Dynamics: Fin-Turbo's Rapid Route to ESG Impact Insight

Xinyun Rong, Weijie Yang
roxyrong@berkeley.edu, raphaelyang1998@berkeley.edu

Abstract

This study introduces "Duration Dynamics: Fin-Turbo's Rapid Route to ESG Impact Insight", an innovative approach employing advanced Natural Language Processing (NLP) techniques to assess the impact duration of ESG events on corporations. Leveraging a unique dataset comprising multilingual news articles, the research explores the utility of machine translation for language uniformity, text segmentation for contextual understanding, data augmentation for dataset balance, and an ensemble learning method integrating models like ESG-BERT, RoBERTa, DeBERTa, and Flan-T5 for nuanced analysis. Yielding excellent results, our research showcases the potential of using language models to improve ESG-oriented decision-making, contributing valuable insights to the FinNLP community.

Keywords: ESG Impact Analysis, Financial NLP, Multilingual Data Pipeline

1. Introduction

The growing emphasis on Environmental, Social and Corporate Governance (ESG) within the financial sector underscores the necessity for better understanding and analysis of ESG-centric information. To address this need, the FinNLP community has been at the forefront of crafting natural language processing tasks on ESG-related news. Previous efforts encompassed taxonomy enrichment, semantic representation[1], ESG-issue identification and classification[2] in a variety of languages.

Building on previous work, ML-ESG-3[3] introduces a new task aimed to evaluate the potential impact duration of ESG events reported in news articles on corporations. This task challenges NLP models to pinpoint the impact timeline of ESG events on a company's performance and sustainability. Gaining this insight is vital for investment decisions, corporate strategies, and policy-making.

This paper delves into the intricacies of leveraging large language models to tackle the ML-ESG-3 challenge. By harnessing a blend of machine translation for multilingual coherence, data processing and augmentation for content uniformity, and a mix of language models like ESG-BERT, RoBERTa, DeBERTa, and Flan-T5, we aim to quantify the temporal effects of ESG-related news on companies across multiple languages. Our work not only enriches the domain's academic discourse but also offers practical insights for stakeholders in the financial sector.

[1]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg

[2]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp2023/shared-task-esg-impact

[3]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3

2. Related Work

Despite the clear definition of Environmental, Social, and Governance principles following years of evolution, systematically identifying and assessing ESG-related news presents persistent challenges, drawing the attention of scholars aiming to address it. The annotation work by Kannan and Seki laid a comprehensive framework for categorizing ESG themes and assessing their sentiment through the analysis of Japanese corporate CSR (Corporate Social Responsibility) reports. Further advancing the field, the DynamicESG project, led by Tseng et al., compiled and analyzed an extensive collection of news articles over a twelve-year period, drawing upon MSCI ESG ratings and SASB standards to categorize news by impact type, level, and duration. The temporal dimension enables a more nuanced analysis of how news coverage could align with ESG criteria.

Domain-specific models have also played a significant role in enhancing the analysis of ESG information. Among these, FinBERT-ESG stands out as a specialized adaptation of the FinBERT model, which has been fine-tuned on 2,000 manually annotated sentences extracted from firms' ESG reports and annual reports (Huang et al., 2022). This allows FinBERT-ESG to efficiently tackle ESG classification tasks. Similarly, ESG-BERT is an environment-focused variant of BERT, initially trained through a Masked Language Model (MLM) task on Accounting for Sustainability corpus, and subsequently fine-tuned for sequence classification tasks (Mehra et al., 2022). These models have proven to be invaluable tools for both researchers and industry practitioners.

In light of these advancements, the ESG task series has been launched, starting with FinSim4-ESG at FinNLP-2022 to expand the taxonomy for semantic analysis of sustainability reports. In 2023, the ML-ESG-1 task focuses on the identification and classification of ESG-related news into 35 key is-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 188–196
20 May, 2024.

sues as per MSCI ESG rating guidelines. ML-ESG-2 further explores the determination of whether the news signifies an opportunity or a risk from an ESG perspective.

3. Dataset & Task Setting

Multilingual ESG Impact Duration Inference (ML-ESG-3) is the latest task which seeks to evaluate the duration or length of impact that an event reported in a news article might have on a company. Specifically, the duration of the impact is classified into three groups: short-term (under 2 years), medium-term (2 to 5 years), and long-term (over 5 years).

The dataset includes 545 English articles from ESGToday[4], 352 Chinese articles from ESG Sustainable Taiwan[5], 661 French articles from Novethic[6], and 800 Korean articles from ESGEconomy[7], 2,358 news articles in total. Each article is provided with a title and the content of the news. As shown in Table 1, the distribution of labels in the training set is fairly even. However, the distribution of these labels under each language is not uniform and varies significantly across the different languages. Generally, news articles categorized with an impact duration of 2 to 5 years form the smallest group, which might require a more detailed examination.

Label	$<$ 2 yr	2 - 5 yr	$>$ 5 yr	Total
Chinese	87 (24.7%)	62 (17.6%)	203 (57.7%)	352
English	82 (15%)	198 (36.3%)	265 (48.6%)	545
French	131 (19.8%)	231 (34.9%)	299 (45.2%)	661
Korean	446 (55.8%)	142 (17.8%)	212 (26.5%)	800
Total	746 (31.6%)	633 (26.9%)	979 (41.5%)	2,358

Table 1: Validation Set Label Distribution

The test set provided later comprises datasets in English, French, and Korean, with a total of 482 news articles. The test set shows a varied distribution of impact duration in English from the training set, as shown in Table 2.

[4]https://www.esgtoday.com/
[5]https://esg.businesstoday.com.tw/
[6]https://www.novethic.fr/
[7]https://www.esgeconomy.com/

Label	$<$ 2 yr	2 - 5 yr	$>$ 5 yr	Total
English	6 (4.4%)	47 (34.6%)	83 (61%)	136
French	31 (21.2%)	32 (21.9%)	83 (56.8%)	146
Korean	96 (48%)	40 (20%)	64 (32%)	200
Total	133 (27.6%)	119 (24.7%)	230 (47.7%)	482

Table 2: Test Set Label Distribution

4. Methodology

This section describes our approach, covering data pre-processing, model selection, and the application of ensemble learning techniques. We detail the different strategies considered at each stage and provide the reasoning for our choice of methodologies.

4.1. Data Pre-processing

4.1.1. Translation

Initially, the bert-base-multilingual-cased model (Devlin et al., 2018) was used as a baseline for our multilingual dataset but achieved a low accuracy of 0.31. Pivoted away from further multilingual adaptations, we adopted machine translation to convert the dataset into English, following a strategy noted in previous research (Lee et al., 2023).

For the translation task, we experimented with the Facebook M2M model (Fan et al., 2020), Google Translation API [8], and DeepL API [9] to translate titles and content from Chinese, French and Korean into English, including a preliminary step of converting traditional Chinese to simplified Chinese to enhance API compatibility. To assess the translation quality and stability, the BLEU metric was introduced, aiming to compare the machine-generated translations with original content by evaluating the precision of n-gram matches (Papineni et al., 2002). This process involved selecting samples from the Chinese, Korean, and French datasets for translation to English, and then back-translating these English texts into the original languages. Table 3 displays the average BLEU scores for all translators under all language translation tasks.

Our sampling review found Google Translate occasionally repeated sentences, and Facebook's M2M underperformed significantly, especially with long articles. DeepL API, however, showed consistent quality without these issues, achieving the high-

[8]https://cloud.google.com/translate
[9]https://www.deepl.com/en/docs-api

Method	Chinese	French	Korean
DeepL API	0.48	0.82	0.62
Google API	0.15	0.56	0.37
FB M2M	0	0	0

Table 3: BLEU Score Across Translation Methods

est BLEU scores across all tested languages. In addition, it showed a strong capability in accurately translating both simplified and traditional Chinese characters. Therefore, DeepL API was selected for all non-English translations.

4.1.2. Article Segmentation

After exploratory data analysis of the translated dataset, we observed that English and French news articles were much shorter, averaging word counts of 73 and 96, respectively. These samples typically highlight the most pertinent sentences—often 2 to 3—from a given news piece, as selected by their annotators. In contrast, articles in Chinese and Korean exhibit substantially higher word counts, averaging 922 and 555 words respectively. Notably, the Chinese articles were extracted and cleaned from their original HTML by our team.

Since a more uniform distribution of content length is generally preferred for model training, it is necessary to employ segmentation techniques, which divide the longer articles into smaller, more manageable paragraphs. Specifically, each article was divided at intervals of every five sentences. Following this, each segment was fed into the FinBERT-ESG classification model, where it was assigned to one of four categories: Environmental, Social, Governance, or None, accompanied by a respective probability score. Paragraphs categorized as None with a probability of 0.9 or higher were excluded. This step not only prevents the datasets in Chinese and Korean from disproportionately influencing the overall training set, but also ensures the integrity and ESG-related quality of the new samples, balancing both the quantity and the quality of data across languages.

Following this segmentation and filtering process, our dataset expanded from 2,358 to 6,115 samples, while reducing the Chinese and Korean average word count to 163 and 119 respectively. A group shuffle split is applied to separate the training and validation sets, ensuring that samples with identical titles are not present in both sets to prevent data leakage. With a training-validation split of 0.2, our dataset was divided into 4,887 samples for training and 1,238 for validation.

4.1.3. Data Augmentation

The step of segmenting articles helps to standardize the length of the content but also introduces the problem of class imbalances, particularly noticeable in the category of impacts lasting 2 to 5 years, which constitutes only 22.3% of the training data. This imbalance makes it challenging for the model to accurately predict medium-term impacts, resulting in a prediction accuracy of less than 0.2 in our baseline.

To address this issue, we leveraged the widely recognized Reuters dataset[10] for news to augment the existing dataset. From the initial pool of 17,712 unique news articles, we adopted a similar methodology as the training set, using FinBERT-ESG for classification and segmentation, which narrowed down the dataset to 2,741 samples. Given the general nature of Reuters news, the selection process was much more stringent, filtering out news to only include those with an E/S/G label probability of 0.5 or higher.

Prior to annotating the Reuter dataset, a preliminary evaluation was conducted on chat-based models, including both commercially available models like GPT-4 (OpenAI, 2024) and Gemini-Pro (Team, 2023), and open-source alternatives such as GPT-NeoXT-Chat-Base-20B (Black et al., 2022) and Pythia-Chat-Base-7B-v0.16 (Biderman et al., 2023), as referenced by Lee et al. (2023). Our evaluation involved a random selection of 100 samples from the train set. GPT-4 and Gemini-Pro recorded accuracies of 50% and 48%, respectively, showing comparable outcomes albeit with noticeable differences in their label distribution. Specifically, GPT-4 categorizes 65% of its labels as more than five years of impact duration, whereas Gemini-Pro identified 53% of impact duration with a two to five-year range. Conversely, the other two models displayed inferior performance, yielding predictions that lacked generalizability. According to Table 4, GPT-NeoXT-Chat-Base-20B frequently predicted an impact duration of more than five years in 90% of cases, while Pythia-7B achieved a mere 29% accuracy rate.

Model	$<$ 2 yr	2 - 5yr	$>$ 5 yr	Acc
GPT-4	5	30	65	0.50
Gemini-Pro	10	53	37	0.48
NeoXT-20B	9	1	90	0.46
Pythia-7B	63	26	0	0.29

Table 4: Label Distribution and Prediction Accuracy across Language Models

From the preliminary study, GPT-4 and Gemini-pro were selected, with GPT-4 acting as the base

[10] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

model and its classifications being cross-verified with those from Gemini-pro. Despite the significant biases and variances in label distribution across these two models, each offers distinct insights into the classification process. When both models agree on a label, the combined accuracy of their predictions can reach 56%. To further refine the selection of ESG-related news and reduce noise in the augmentation set, we also introduced a new classification, "Not have an ESG impact," into the prompt, detailed in the appendix. After GPT-4 filtered the unrelated news, any article receiving the same classification was included in the augmentation set.

The final augmented dataset thus comprised 1,221 samples, with 858 labeled as having an impact duration of 2 to 5 years. Following augmentation, the training set expanded to 6,098 samples, achieving a more balanced distribution across categories, as shown in Table 5.

Model	$<$ **2 yr**	**2 - 5 yr**	$>$ **5 yr**
Non-Augmented	1,681 (34.5%)	1,092 (22.4%)	2,104 (43.1%)
Augmented	1,867 (30.6%)	1,950 (32.0%)	2,281 (37.4%)

Table 5: Label Distribution after Augmentation

4.2. Model Selection

Considering the diverse topics and professional terminology found in ESG news, often presented in long articles, we selected ESG-BERT (Mehra et al., 2022), RoBERTa (Zhuang et al., 2021), DeBERTa (He et al., 2021), and Flan-T5 (et. al, 2022) as our pre-trained models to perform finetuning, for their advanced generalization capabilities, semantic understanding, and popularity in the prior ESG task series. Here is a brief overview of each model's strengths:

- ESG-BERT is highly effective in extracting and classifying information pertinent to sustainable investing and ESG themes. This effectiveness is largely due to its tailored training on ESG-specific text, enhancing its capacity for ESG task performance and semantic extraction, as highlighted by Lacoste et al. (2019).

- RoBERTa is an enhanced variant of BERT with a dynamic masking mechanism, 10x training corpus, and an improved training strategy, resulting in superior text comprehension and model generalization. Notably, a study conducted by Pontes et al. (2023) underscores RoBERTa's proficiency and accuracy in classifying news documents into specific ESG issue labels within an English-language dataset.

- DeBERTa is designed for processing lengthy articles, as its disentangled attention mechanism is key for analyzing long-distance sentence dependencies, essential for understanding context and handling complex sentence structures. In the realm of ESG, DeBERTa has demonstrated commendable efficacy and precision in identifying fraudulent ESG news, attributed to its advanced attention mechanism (Suryavardan et al., 2023).

- Flan-T5 is designed to generalize better to new tasks with minimal examples. Its few-shot learning capability allows it to understand and perform tasks that it might not have been explicitly trained for, using only a few examples to guide its predictions. This versatility makes Flan-T5 an excellent choice for ESG impact duration inference. Specifically, its encoder part is used for extracting semantic meanings, aiding in the inference.

During this stage, each model underwent finetuning on the training dataset, which was either the original or augmented version, through adjustments to its structure, such as selecting layers to unfreeze and incorporating extra layers before the softmax layer, as well as tweaking hyperparameters including batch size, dropout rate, and learning rate. The goal was to determine the optimal settings for each model under consideration.

4.3. Ensemble Learning

Ensemble learning can effectively reduce overfitting by averaging out biases and variances across diverse models (Opitz and Maclin, 1999) and correcting errors of weak learners, further improving model performance (Schapire, 1990). Several experiments were conducted to integrate classification results of the four distinct models by taking their softmax layers of probabilities as the inputs.

Prior to applying ensemble learning, to manage scenarios where multiple segments from a single news article yield divergent predictions, we calculated the mean softmax value for segments sharing the same Group ID. For example, Figure 1 demonstrates how a news article is partitioned into three segments, with each outputting a softmax layer configured as of 1 x 3 (corresponding to the probabilities of the 3 classes), and these segments are collectively averaged.

Figure 2 displays the entire model architecture for ensemble learning. After aggregating softmax results for each article, we applied averaging for each class across all four finetuned models as one of the

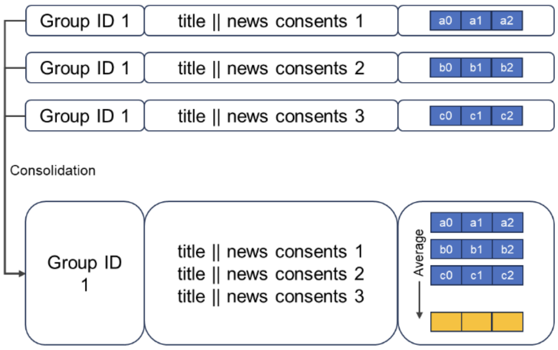


Figure 1: Consolidated prediction via averaging softmax outputs with same Group ID.

ensemble approaches. For the rest of the experiments, correlation analysis was conducted first to eliminate pairs exhibiting a correlation higher than 0.7. The remaining probabilities were input into various classifiers, including K-Nearest Neighbor, Decision Tree, Random Forest, and Multiple Linear Perceptrons, to produce the predicted labels.

5. Experiment Results

The experimental process is divided into three key stages: 1) evaluating the performance of the augmentation set, 2) tuning the hyperparameters for all models, and 3) identifying the most effective ensemble techniques to determine the optimal model. Results for each stage are presented and analyzed in this section.

5.1. Augmentation Effectiveness

During the first phase of the experiments, a learning rate of 1e-5, a batch size of 32, and the strategy of unfreezing the last three layers for each model were adopted. For the experiments that utilized non-augmented data, resampling techniques were employed to address issues of data imbalance.

As shown in Table 6 the comparison between models using non-augmented and augmented data reveals varying degrees of performance improvement. ESG-BERT and DeBERTa-base saw improvements in both Micro and Macro F1 score; RoBERTa-base experienced mixed results as Macro F1 dropped slightly.

The analysis suggests that while augmentation can lead to a more balanced dataset and potentially better model performance, the effectiveness of these techniques can differ between models. We use Macro F1 as the criteria for determining whether the augmented dataset is used for model training in subsequent training phases.

5.2. Hyperparameter Tuning Results

In the second phase of the experiments, we focused on hyperparameter tuning for each model, mainly on the number of un-freezed layers, batch size and learning rate. Table 7 illustrates the parameters we searched, best parameters for each model, and their Micro F1 and Macro F1 scores. Among all the models, RoBERTa-large, ESG-BERT, DeBERTa-base, and Flan-T5-large emerged as the top performers, making them prime candidates for integration into ensemble models due to their relatively high prediction accuracy.

5.3. Ensemble Learning Results

In the previous step, RoBERTa-base recorded a Micro F1 score of 0.5998 and a Macro F1 score of 0.4899, setting a benchmark for the ensemble learning phase. Figure 3 illustrates the Macro F1 scores for various ensemble learning methods and their performance across different languages. Techniques such as MLP, K-Nearest Neighbors, and Averaging all surpassed the RoBERTa-base baseline in terms of score.

5.4. Test Results

Finally, we presented the test outcomes for four individual models and five ensemble models, including the three submitted in the ML-ESG-3 task, as displayed in Table 8. The Flan-T5 model exhibited superior results on the English dataset, while the Averaging ensemble model outperformed others on the French dataset, and Random Forest emerged as the top-performing model for the Korean dataset.

6. Discussions

In this section, we dive into prediction results and discuss potential future work to improve model robustness. This entails analyzing the effect on word count, augmentation set and validation set.

6.1. Effect of Word Count

We observed a notable trade-off in model performance between English/French datasets and Korean datasets, likely due to the large disparity in article word count and label distribution. To understand the impact of word count, we performed a logistic regression analysis, revealing a significant positive correlation between word count and model performance, significant at a 10% level (p-value: 0.074).

Given the analysis, two principal methodologies are applicable to achieve this aim in the future: extending the text length per input and adopting ad-

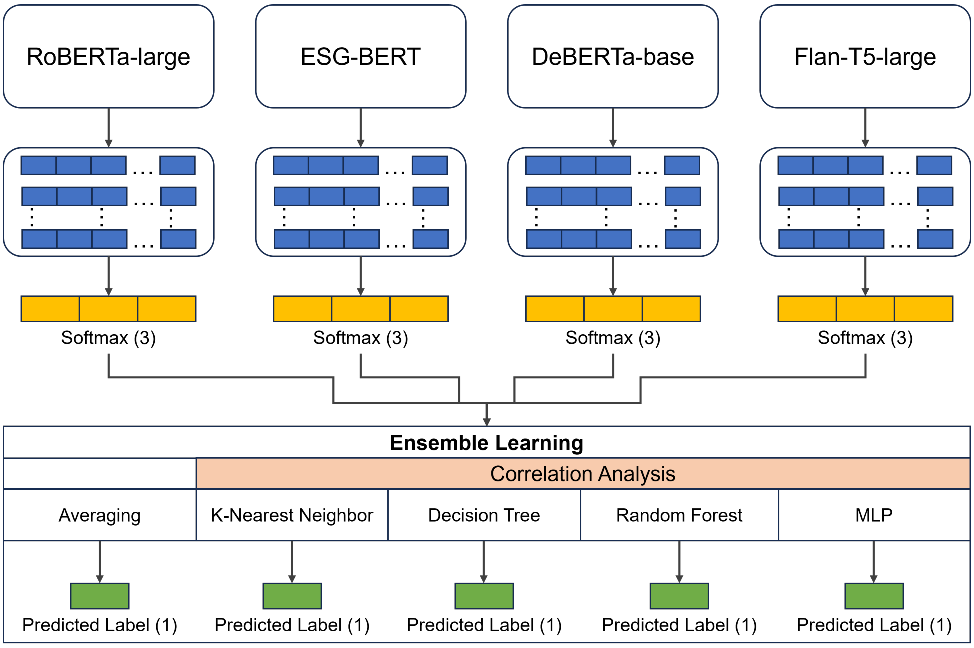


Figure 2: The Complete Ensemble Model Architecture

Model	Non-augmented		Augmented	
	Micro F1	Macro F1	Micro F1	Macro F1
ESG-BERT	0.5521	0.4823	0.5645	0.5050
RoBERTa-base	0.5721	0.5147	0.5807	0.4869
DeBERTa-base	0.5654	0.4823	0.5922	0.4881

Table 6: Augmentation Set Performance across Models

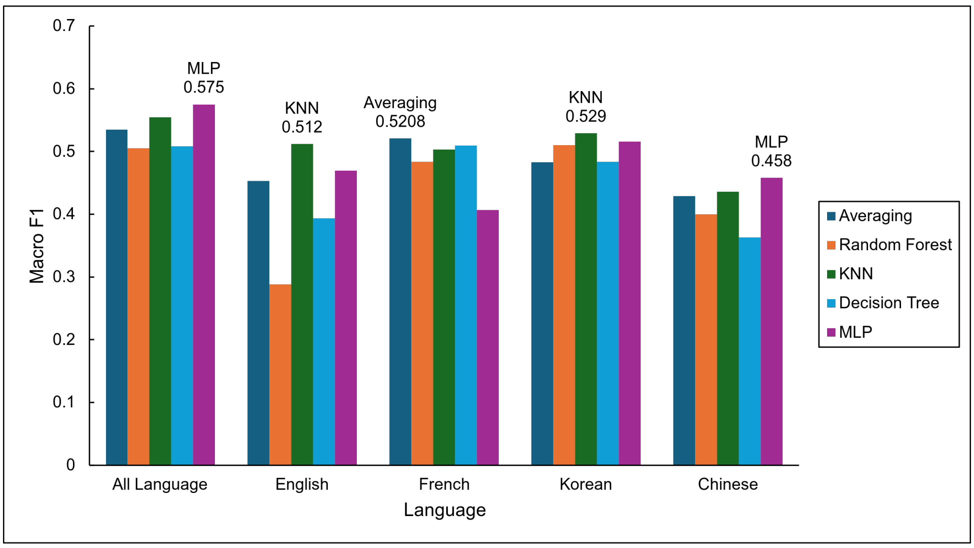


Figure 3: Ensemble Learning Macro F1 Score across Languages

vanced segmentation techniques beyond simple segmentation.

Regarding input length, directly extracting English and French news from original URLs could be beneficial, mirroring the success seen with Korean articles, which typically contains more information. In addition, adjusting segmentation sizes could be advantageous, especially considering the current

Model	Parameter Search Value	Best Parameter	Micro F1	Macro F1
ESG-BERT	num_layers = 1, 3, 6 batch_size = [16, 32, 64] lr = [1e-4, 5e-4, 1e-5]	num_layers = 3 batch_size = 32 lr = 1e-5	0.5645	0.5050
RoBERTa-base	num_layers = 1, 3, 6 batch_size = [16, 32] lr = [1e-5, 1.5e-5, 5e–5]	num_layers = 6 batch_size = 32 lr = 1.5e-5	0.5998	0.4899
RoBERTa-large	num_layers = 1, 3, 6 batch_size = [16, 32] lr = [1e-5, 1.5e-5, 5e–5]	num_layers = 6 batch_size = 16 lr = 1.5e-5	0.5874	0.5057
DeBERTa-v3-base	num_layers = 1, 3, 6 batch_size = [16, 32] lr = [1e-5, 5e-5]	num_layers = 3 batch_size = 16 lr = 1.5e-5	0.5702	0.4596
DeBERTa-v3-large	num_layers = 1, 3, 6 batch_size = [16, 32] lr = [1e-5, 5e-5]	num_layers = 3 batch_size = 16 lr = 1.5e-5	0.5683	0.4549
Flan-T5-base	num_layers = 1, 3 batch_size = [16, 32] lr = [1e-5, 2e-5, 5e-5]	num_layers = 3 lr = 5e-05	0.5711	0.4050
Flan-T5-large	num_layers = 1, 3 batch_size = [16, 32] lr = [1e-5, 2e-5, 5e-5]	num_layers = 3 lr = 2e-5	0.6050	0.4293

Table 7: Hyperparameters Tuning Results on Validation Set

	English		French		Korean	
Model	Micro F1	Macro F1	Micro F1	Macro F1	Micro F1	Macro F1
ESG-BERT	0.5735	0.4120	0.4315	0.3844	0.6450	0.5148
RoBERTa	0.4559	0.3705	0.4521	0.4357	0.6500	0.5637
DeBERTa	0.5000	0.4063	0.5000	0.4654	0.6150	0.4630
Flan-T5-Large*	**0.6912**	**0.4689**	0.5753	0.4335	0.6450	0.4730
MLP (ES)*	0.6176	0.4035	0.4863	0.4488	0.6550	0.5126
Averaging (ES)*	0.5882	0.3983	**0.5616**	**0.4622**	0.6500	0.4732
KNN (ES)	0.4706	0.3820	0.4041	0.3965	0.6450	0.5881
RF (ES)	0.4632	0.3789	0.4247	0.4206	**0.6600**	**0.6214**
DT (ES)	0.4412	0.3546	0.3562	0.3312	0.6100	0.5378

Table 8: Performance Metrics by Language and Model on Testset. Model marked with asterisk(*) were submitted to the committee.

average segmentation size is about 100 tokens, while the capacity of all models extends to 512 tokens.

In terms of segmentation techniques, the implementation of sliding window segmentation could improve contextual flow and semantic continuity. Cross-segment attention integrates full-article context, potentially improving the model's ability to understand long-distance dependencies and intricate relationships (Lukasik et al., 2020). Hierarchical BERT serves to bridge local and global contexts within an article, amalgamating both detailed and overarching semantic information (Lu et al., 2021).

6.2. Effect of Augmentation Set

Despite our effort in handling imbalanced dataset and low prediction accuracy in impact duration between 2 and 5 years, our model still struggles with the out-of-sample distribution issue in the test set, indicating a potential over-fitting to the training data. A potential enhancement in our process could involve adopting Gemini-pro as the primary model for labeling augmentation dataset, given its superior F1 score of 0.5144 in the medium-term duration inference, in contrast to GPT4's F1 score of 0.4179.

In addition, the inclusion of ESG-related news rather generic business news for the augmentation set would likely boost model performance. For instance, utilizing the Global Database of Events, Language, and Tone (GDELT) project enabled Aue et al. to collect 8,000 ESG-related ratings derived from 3 million articles pertaining to 3,000 US corporations throughout the period of 2018 to 2020. We could use this database rather than the filtered Reuters dataset for data augmentation.

7. Conclusion

This study undertakes the ML-ESG-3 shared task, with the goal of predicting the ESG impacts duration across datasets in English, French, Korean, and Chinese. We finetuned BERT-based and T5-based classifiers in conjunction with techniques such as machine translation, text segmentation, data augmentation, and ensemble learning. Our findings indicate the performance enhancement from data augmentation and strategic segmentation while mitigate issues of class imbalance. Through experimentation, we identified the optimal model configurations that significantly enhanced our predictions' precision and reliability. Our research contributes valuable insights and methodologies to the FinNLP community, providing a robust framework for assessing the temporal effects of ESG-related news on corporations. Future directions include enhancing our approach by extending text inputs length, employing advanced segmentation techniques for better contextual understanding, and concentrating on ESG-related news to enrich our data augmentation process.

8. Availability

The code is available at https://github.com/roxyrong/ml-esg-3.

9. Bibliographical References & Language Resources

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting companies' esg ratings from news articles using multivariate timeseries analysis.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Hyung Won Chung et. al. 2022. Scaling instruction-finetuned language models.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR 2021*.

Allen H. Huang, Hui Wang, and Yi Yang. 2022. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. In *Proceedings of The 5th Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. EaSyGuide: ESG Issue Identification Framework leveraging Abilities of Generative Large Language Models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 127–132, Macao.

Jinghui Lu, Maeve Henchion, Ivan Bacher, and Brian Mac Namee. 2021. A sentence-level hierarchical bert model for document classification with limited labelled data.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Sim oes. 2020. Text segmentation by cross segment attention. In *Proceedings of* the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4707–4716, Online. Association for Computational Linguistics.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *Embedded Systems and Applications*, EMSA 2022. Academy and Industry Research Collaboration Center (AIRCC).

OpenAI. 2024. Gpt-4 technical report.

D. Opitz and R. Maclin. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198.

3DS Outscale. 2024. Ml-esg 2024 for social good (esg) - 3rd edition guidelines. Guidelines by 3DS Outscale.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging bert language models for multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 121–126, Macao.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

R.E. Schapire. 1990. The strength of weak learnability. *Machine Learning*, 5(2):197–227.

S. Suryavardan, S. Mishra, M. Chakraborty, P. Patwa, A. Rani, A. Chadha, A.N. Reganti, A. Das, A.P. Sheth, M.K. Chinnakotla, A. Ekbal, and S. Kumar. 2023. Findings of factify 2: Multimodal fake news detection. *ArXiv*, abs/2307.10475.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of The 32nd ACM International Conference on Information and Knowledge Management (CIKM'23)*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized bert pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

10. Appendix

10.1. Template for Chat-based Language Model for Labeling

Below is our one shot learning template for labeling augmented using various chat-based language models.

TEMPLATE =
"""
Label the ESG impact duration for the following news:
Options:

- 0 - below 2 years
- 1 - between 2 and 5 years
- 2 - more than 5 years

You should only output the number and have no explanations.

Example: The ways to practice self-care with a fitness watch are almost limitless, but here are six easy-to-implement tips to start today | Dubai-based airline Emirates announced plans to conduct its first experimental flight using 100% sustainable aviation fuel (SAF) in one engine this week, in a test aimed at supporting expanded use of SAF for commercial flights.
Output: 0

News: {$news}
Output:
"""

Multilingual ESG News Impact Identification using an Augmented Ensemble Approach

Harika Abburi[1], Ajay Kumar[2], Edward Bowen[2], Balaji Veeramani[2]

[1]Deloitte & Touche Assurance & Enterprise Risk Services India Private Limited India
[2]Deloitte & Touche LLP, USA

{abharika, ajaykumar35, edbowen, bveeramani}@deloitte.com

Abstract

Determining the duration and length of a news event's impact on a company's performance remains elusive for financial analysts. The complexity arises from the fact that the effects of these news articles are influenced by various extraneous factors and can change over time. As a result, in this work, we investigate our ability to predict 1) the duration (length) of a news event's impact, and 2) level of impact on companies. The datasets used in this study are provided as part of the Multi-Lingual ESG Impact Duration Inference (ML-ESG-3) shared task. To handle the data scarcity, we explored data augmentation techniques to augment our training data. To address each of the research objectives stated above, we employ an ensemble approach combining transformer model, a variant of Convolutional Neural Networks (CNNs), specifically the KimCNN model and contextual embeddings. The model's performance is assessed across a multilingual dataset encompassing English, French, Japanese, and Korean news articles. For the first task of determining impact duration, our model ranked in first, fifth, seventh, and eight place for Japanese, French, Korean and English texts respectively (with respective macro F1 scores of 0.256, 0.458, 0.552, 0.441). For the second task of assessing impact level, our model ranked in sixth, and eight place for French and English texts, respectively (with respective macro F1 scores of 0.488 and 0.550).

Keywords: impact, data augmentation, transformers, CNN

1. Introduction

The surge in Environmental, Social, and Governance (ESG) research over the past few years is a testament to the growing importance of these issues in the corporate world (Zumente and Bistrova, 2021). Companies are increasingly recognizing that ESG-related matters can pose significant risks if not addressed properly (Aue et al., 2022). This rising awareness, and importance of the analyzing large volumes of ESG related documents has necessitated the use of language technologies in this area.

The rapid advancements in deep learning, and Natural Language Processing (NLP) technologies have enabled the research in development of systems designed to extract relevant information from ESG reports. Language models have been used for various financial tasks such as sentiment analysis, named-entity recognition, and document classification (Araci, 2019; Wu et al., 2023). However, their application to ESG-specific tasks remains relatively limited. Existing works have begun to explore this area, demonstrating the potential of language models for ESG analysis. (Raman et al., 2020) evaluate the impact of language model embeddings on the classification of sentences concerning their relevance to the ESG domain. Similarly, (Mehra et al., 2022) pre-train a Bidirectional Encoder Representations from Transformers (BERT) model on ESG-related text to show improvement on classification tasks. Furthermore, (Wang et al., 2023) explore the potential of combining contrastive learning with BERT language model for the task of identifying environmental, social, and governance issues in news articles.

Despite these promising initial explorations, the scarcity of publicly available ESG data, particularly for low-resource languages, remains a significant challenge that hinders further advancements in this field. To address this issue, various data augmentation techniques have been explored to expand and enrich the training data, including Easy Data Augmentation, translation, zero-shot classification, contextual augmentation (Lee et al., 2023; Kobayashi, 2018). (Nugent et al., 2021) leverage back-translation technique to generate additional training data to perform ESG document classification. The generated data is then used to fine-tune the BERT model to further enhance its performance. Furthermore (Glenn et al., 2023), generated synthetic data with LLMs in zero-shot and few-shot settings effectively bridging the gaps in data availability for low-resource languages.

These efforts have paved the way for the development of advanced multilingual solutions. (Mashkin and Chersoni, 2023) highlights the usefulness of using Transformer-based representations and cross-lingual models for multilingual ESG analysis. (Jør-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 197–202

gensen et al., 2023) extend the concept of pre-training on financial text to multilingual data in seven languages. Additionally, (Pontes et al., 2023) investigates the use of BERT and its variants for classifying news articles into different ESG categories. They also explores the effectiveness of these models in multiple languages, offering insights into the potential of this approach for expanding the scope of ESG issue identification.

Motivated by these developments, our team participated in the Financial Technology and Natural Language Processing, the Knowledge Discovery from Unstructured Data in Financial Services, and Economics and Natural Language Processing (FinNLP-KDF-ECONLP-2024) shared task on ML-ESG-3 (Chen et al., 2024). This task aimed to predict the duration and level of a news article's impact on a company. Towards this task, we adopted data augmentation techniques such as translation, paraphrasing, and Generative Pre-training Transformer (GPT) mix to augment the training data. Furthermore, we trained an ensemble model combining transformers, KimCNN architecture (Kim, 2014), and Voyage AI embeddings[1] and assessed their performance across various languages and subtasks. Our model achieved top rankings (ranging from 1st to 8th) across different subtasks, demonstrating the effectiveness of our approach in furthering the capabilities of NLP for identifying ESG impact level and duration.

2. Dataset

This section describes the dataset used for exploring the ML-ESG-3 shared task (Chen et al., 2024). The dataset consists of ESG new articles annotated with one or more annotations to each news article. The data was provided by the task organizers and the task is slightly different across the language subsets.

- English and French: This dataset includes two annotations: "Impact Level" and "Impact Length". Impact Level qualifies the opportunity or risk as being of "low", "medium" or "high." Impact Length annotations of "Less than 2 years", "2 to 5 years", and "More than 5 years".
- Japanese: For this language, only the annotations of 'Impact Length' are provided which are similar to the English and French datasets.
- Korean: In this dataset, there are two annotations: 'Impact Length' and 'Impact Type'. Impact Length annotations are same as the English and French datasets where as 'Impact Type' is categorized as 'opportunity,' 'risk,' or 'cannot distinguish' (Tseng et al., 2023). In Korean language, we participated only in Impact length.

Table 1 shows the data statistics for the languages and subtasks. More detailed information about the dataset can be found in the shared task overview paper (Chen et al., 2024).

Language	Impact Length		Impact Level	
	Train	*Test*	*Train*	*Test*
English	545	136	545	136
French	661	146	661	146
Japanese	52	1500	—	—
Korean	800	200	—	—

Table 1: Dataset statistics (number of samples) of Original data

Data Type	Impact Length	Impact Level
Original	545	545
Original + tr	1835	1054
Original + tr +pp	3670	2108
Original + tr+ pp + GPT-mix	6670	5108

Table 2: Original and augmented training data statistics (number of samples) for English subtasks. Augmentations were performed using translation (tr), paraphrasing (pp) and GPT-mix.

2.1. Synthetic Data Generation

Due to the limited data available in each language task, we employed various data augmentation techniques to enrich the training set: we used translation (tr), paraphrasing (pp), and GPT-mix.

Translation: To augment the training data, we translated the French, Japanese, and Korean datasets into English using the widely recognized DeepL[2] translation service. We used English data as it is and converted the other languages to English.

Paraphrase: After translation, we employed the Pre-training with Extracted Gap-sentences for Abstractive Summarization (PEGASUS) transformer model (Marceau et al., 2022) for paraphrasing the text. While this model was originally designed for abstractive summarization, its ability to leverage large amounts of text and understand semantic relationships between words makes it suitable for paraphrasing tasks.

GPT-mix: We further augmented the data using

[1] https://www.voyageai.com/

[2] https://www.deepl.com/translator

GPT-mix (Yoo et al., 2021), a technique that leverages large language models to generate summary of text samples. GPT-mix effectively captures human language nuances by blending two real samples. We selected two samples with identical labels from the original dataset and used GPT-mix to generate summary of these pairs, creating new data points. This process yielded 3,000 additional samples.

Table 2 shows the number of samples augmented for the English subtasks using the aforementioned augmentation techniques. We then translated this augmented data into other languages using DeepL translator, standardizing the number of training samples across language subtasks.

Our use of translation of data across languages, and the use of transformation (pp) and mixing samples for augmentations helps create better datasets for data sparse tasks.

3. Proposed approach

In this section, we describe our approach for detecting the duration and length of a news event's impact on a company's performance. Our text classification architecture builds upon a modified KimCNN framework (Kim, 2014), with carefully incorporated transformer-based representations and Voyage AI embeddings. The proposed framework comprises five specific layers: embedding layer, CNN layer, pooling layer, enriched representation layer and output layer. The detailed description of each layer is given as follows.

Embedding layer: Our approach begins with a pre-trained transformer model, which has been trained on a massive corpus of text data. This allows our model to capture rich contextual information about the meaning of words and their relationships within the text. Instead of using the standard output of the transformer model, we specifically focus on the final four hidden layers of the transformer model as these layers effectively captures the relevant information from input data.

Convolutional layer: Following the embedding layer, the extracted representations then undergo a series of convolutional operations. We build upon the KimCNN architecture, which is known for its effectiveness in text classification tasks. This architecture utilizes multiple convolutional layers with varying filter sizes (specifically 3, 4, and 5) with padding enabling the model to learn patterns from different n-gram combinations within the text. This allows the model to focus on different n-gram lengths, potentially capturing both short and long-range dependencies that contribute to the overall meaning. To improve efficiency, we use depthwise separable convolutions. After each convolutional layer, we apply a Rectified Linear activation Unit (ReLU) activation function to introduce non-linearity. This allows the model to learn more complex relationships between features. Additionally, we use dropout as a regularization technique to prevent overfitting to the training data.

Pooling layer: Following the convolutional layer, max-over-time pooling is applied to each convolutional layer's output. This operation extracts the prominent feature from each sequence captured by the convolution, focusing on the relevant information within each n-gram length. The features from the convolutional layers are then concatenated into a single representation, effectively combining the information learned from different n-gram lengths.

Enriched representation layer: To further enhance the model's understanding, the single representation is again concatenated with Voyage AI embeddings. These state-of-the-art pre-trained text embedding models capture semantic meaning from text data, effectively injecting external knowledge into the model.

Output layer: The final concatenated representation is then fed into a fully connected layer. This layer performs the final classification task, assigning probabilities to each possible class the text belongs to, enabling the model to predict the impact duration of the news event.

3.1. Different transformer based models

We explored various state-of-the-art large language models (Kalyan et al., 2021) to extract the features from the embedding layer. These include prominent models like such as; Bidirectional Encoder Representations from Transformers (BERT),Robustly optimized BERT approach (RoBERTa), and its cross-lingual language model RoBERTa (XLM-RoBERTa) along with their variants. However, we recognized that a single set of models might not perform equally well across diverse datasets and languages within the task. Therefore, we fine-tuned different LLM variants for each subtask and language and pick the top performing models based on heldout data. Table 3 lists the different LLMs that we explored for embedding layer of each subtask: English impact length (English-len), English impact level (English-lev),French impact length (French-len), French im-

Subtask	Transformer model
English-len	EnvRoBERTa-base[3] (Schimanski et al., 2023)
English-lev	ESG-BERT[4]
French-len	xlm-roberta[5] (Reimers and Gurevych, 2019)
French-lev	bert-base-multilingual[6]
Japanese-len	xlm-roberta (Reimers and Gurevych, 2019)
Korean-len	multilingual-mpnet-base[7]

Table 3: Transformer models used for different subtasks

Data Type	Impact Length				Impact Level			
	Acc	F_{macro}	$Prec$	Rec	Acc	F_{macro}	$Prec$	Rec
Original	0.476	0.345	0.317	0.378	0.524	0.462	0.546	0.468
Original + tr	0.451	0.376	0.405	0.383	0.500	0.438	0.466	0.445
Original + tr +pp	0.451	0.386	0.390	0.388	0.561	0.525	0.578	0.521
Original + tr+pp + GPT-mix	**0.524**	**0.477**	**0.478**	**0.476**	**0.585**	**0.530**	**0.583**	**0.533**

Table 4: Results on held out English dataset using various data augmentations

pact level (French-lev), Japanese impact length (Japanese-len), Korean impact length (Korean-len).

4. Experiments

This section provides the experimental evaluation of our proposed methods. For each of the tasks we report accuracy (Acc), macro F1 score (F_{macro}), precision ($Prec$) and recall (Rec).

4.1. Implementation details

We set aside 15% from the training data for performance evaluation. For the testing phase, the held-out set is merged with the training set. The experiments were conducted using the same hyperparameters: batchsize of 64, learning rate of 2e-5, epoch of 10, and optimizer of AdamW. The experiments were run on two A100 GPUs.

4.2. Results

For each task and language, we submitted three runs to the leaderboard (team name *Drocks*). These runs correspond to the different approaches on the heldout data. In this paper, we show only the top run results for each of the tasks. The full leaderboard is available at *https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3*.

4.3. Effect of data augmentation techniques

To evaluate the effectiveness of data augmentation, we conducted experiments with incremental addition of data through translation, paraphrasing, and GPT-mix on the English dataset using the ESG-BERT model. Table 4 shows the performance metrics for models trained on augmented data, as detailed in Table 2. Notably, the results demonstrate that combining data augmentation with translation, paraphrasing and GPT-mix techniques improves the model's performance on both English subtasks with good margin of F_{macro} score. Therefore, for subsequent experiments, we utilize the "Original+tr+pp+GPT-mix" training data for training the model, and report the results on the held-out data for various sub-tasks.

Subtask	Acc	F_{macro}	$Prec$	Rec
English-len	0.596	0.441	0.439	0.451
English-lev	0.574	0.550	0.583	0.535
French-len	0.500	0.458	0.469	0.470
French-lev	0.486	0.488	0.508	0.483
Japanese-len	0.363	0.256	0.220	0.370
Korean-len	0.625	0.552	0.580	0.549

Table 5: Results on the subtasks on testing data

4.4. Results of proposed approach

Table 5 presents the performance of our proposed architecture across the subtasks using the augmented datasets. On English subtasks, the model achieved F_{macro} scores of 0.441 and 0.550 for the English-len and English-lev subtasks, respectively.

[3] https://huggingface.co/ESGBERT/EnvRoBERTa-base

[4] https://huggingface.co/nbroad/ESG-BERT

[5] https://huggingface.co/sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens

[6] https://huggingface.co/Tiamz/bert-base-multilingual-uncased-finetuned-news

[7] https://huggingface.co/sosoai/multilingual-mpnet-base-v2-embedding-all-safetensor

Utilizing the augmented French data, the model achieved F_{macro} scores of 0.458 and 0.488 for the French-len and French-lev subtasks respectively. Similar to English, the slightly lower score for French-len as compared to French-lev suggests identifying impact length might be more challenging as compared to impact level. For the Japanese subtask, the model only achieved an F_{macro} score of 0.256. While this score is lower than other subtasks, it secured the first rank in the competition, highlighting the potential of the approach in this specific task. However, the model achieved a higher F_{macro} score of 0.552 on the Korean subtask.These variations showcase the complexities of applying the model to diverse languages with varying volumes of data, potentially pointing towards areas of future investigation that will be of interest.

5. Conclusion

In this paper, we described our submission to the FinNLP-KDF shared task which consists of multiple sub tasks in determining the duration and level of the impact an event in the news article might have on the company. Our experiments demonstrated that data augmentation techniques effectively improved model performance. Furthermore, the proposed approach ranked within the top 10 for several languages (English, French, Korean) and securing first place based on F_{macro} score for the Japanese language subtask. These findings highlight the potential of the approach for multilingual ESG impact duration inference. However, the variations in performance across languages and subtasks underscore the inherent challenge of this domain. Future work will focus on further enhancing and adapting the model to address these complexities and improve performance across languages and tasks.

6. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting companies' esg ratings from news articles using multivariate timeseries analysis.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the finnlp-2023: Using synthetic data and transfer learning for multilingual esg issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. MultiFin: A dataset for multilingual financial NLP.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mDAPT: Multilingual domain adaptive pretraining in a single model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3404–3418, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2021. Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. Easyguide: Esg issue identification framework leveraging abilities of generative large language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*.

Louis Marceau, Raouf Belbahar, Marc Queudot, Nada Naji, Eric Charton, and Marie-Jean Meurs. 2022. Quick starting dialog systems with paraphrase generation. *arXiv preprint arXiv:2204.02546*.

Ivan Mashkin and Emmanuele Chersoni. 2023. Hkesg at the ml-esg task: Exploring transformer representations for multilingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 140–145.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies' environmental, social, and governance practices. In *CS & IT Conference Proceedings*, volume 12. CS & IT Conference Proceedings.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging bert language models for multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 121–126.

Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, 2(4):453–468.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Bridiging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication. *Available on SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract*$_id = 4622514$.

Hanna Schramm-Klein, Joachim Zentes, Sascha Steinmann, Bernhard Swoboda, and Dirk Morschett. 2016. Retailer corporate social responsibility is relevant to consumer behavior. *Business & Society*, 55(4):550–575.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2335.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging contrastive learning with bert for esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 116–120.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilze Zumente and Jūlija Bistrova. 2021. Esg importance for long-term shareholder value creation: Literature vs. practice. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(2):127.

Cheap Talk: Topic Analysis of CSR Themes on Corporate Twitter

Nile Phillips[1], Sathvika Anand[1], Michelle Lum[1], Manisha Goel[2], Michelle Zemel[2], and Alexandra Schofield[1]
[1]Harvey Mudd College, Claremont, CA, USA
{nphillips, sanand, milum}@hmc.edu, xanda@cs.hmc.edu
[2]Pomona College, Claremont, CA, USA
{manisha.goel, michelle.zemel}@pomona.edu

Abstract

Numerous firms advertise action around corporate social responsibility (CSR) on social media. Using a Twitter corpus from S&P 500 companies and topic modeling, we investigate how companies talk about their social and sustainability efforts and whether CSR-related speech predicts Environmental, Social, and Governance (ESG) risk scores. As part of our work in progress, we present early findings suggesting a possible distinction in language between authentic discussion of positive practices and corporate posturing.

Keywords: corporate social responsibility, ESG, topic models, Twitter, social media, sustainability

1. Introduction

The last two decades have witnessed an urgent recognition by investors, and in response, firms, of a role for corporations in *corporate social responsibility* (CSR) (Bowen, 2013) and environmental stewardship. CSR integrates societal goals into firms' objectives, potentially channeling private investment towards a public good such as combating climate change and addressing inequality. Investors' demand for CSR activities has increased dramatically over this time period: the total market value of US assets managed with ESG strategies in 2020 totaled $17.1 trillion, a 33% increase from 2018's value and ake 25-fold increase relative to 1995 (US SIF Foundation, 2020). While company approaches to CSR may not impact their bottom line (McWilliams and Siegel, 2000), there are political, ethical, and social implications for why a company may build a CSR-focused strategy, with or without a profit motive (Garriga and Melé, 2004).

Social media platforms like Twitter[1] allow companies to communicate publicly about CSR to improve brand awareness and perception (Pilgrim and Bohnet-Joschko, December 2022; Araujo and Kollat, 2018). The embrace of Twitter as a platform to reach shareholders has included the creation of distinct corporate accounts, such as @KelloggsCompany or @CocaColaCo, focused not on products but corporate actions. Our ongoing project explores whether corporate Twitter messaging describes genuine commitments to social goals or instead is an example of "cheap talk" to paint companies in a positive light. We examine the behavior of S&P 500 English-language Twitter accounts, using a topic model to characterize themes in how they communicate about CSR. We present our work in progress, in which we find both concrete, action-oriented CSR-focused topics and more abstract topics highlighting sustainability and social good. We also compare our behavioral findings with Sustainalytics ESG risk scores to demonstrate that less concrete topics can correlate with increased risk.

[1]Our dataset predates renaming Twitter to "X" in 2023.

2. Background

2.1. Corporate Social Responsibility

Investor demand for firm CSR commitments can be explained by two dominant competing theories. Under the "doing well by doing good" theory, investor demand for integration of CSR stems from a belief that CSR activities lead to increased financial benefit to shareholders (McWilliams and Siegel, 2000; Orlitzky et al., 2003). In this theory, shifting to cleaner technologies, employing a diverse workforce, or partnering with local communities, for example, are long-term profit maximizing decisions. In contrast, an alternate theory suggests that demand for CSR is driven by non-pecuniary benefits to investors, such as cleaner air and social equality (Garriga and Melé, 2004). Under either theory, CSR creates value for investors and may drive engagement with current or prospective investors.

Through channels such as financial reports, shareholder calls, and more recently, social media, firms can signal their commitments to CSR to current shareholders, potential investors, consumers, and employees (Araujo and Kollat, 2018). To the extent that signaling a CSR commitment is less costly than executing on the commitment, especially in the less-regulated landscape of social media, the conditions for "cheap talk", or in the case of environmental initiatives, "greenwashing", exist. There is growing evidence that this phenomenon of "cheap talk" is present in social media discus-

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 203–211

sion of ESG commitments. Crowley et al. (2019) show that firms strategically use Twitter communications to "greenwash," i.e., exaggerate their CSR activities. In fact, those that are rated worse on ESG rankings talk more about their initiatives to build a more positive reputation even if this talk is only cheap and not consistent with their actions in reality. Baker et al. (2023) demonstrate that firms similarly use voluntary disclosures to make strong statements about their commitment to diversity initiatives but significantly lag in their actions. This helps build their reputations with customers and investors, and also improve their ESG ratings. Attig and Boshanna (2023) show, however, that such cheap talk worsens firms' market performance.

2.2. Twitter Analysis for CSR

Twitter data has been used for a variety of corporate analyses, including predicting stock behavior (Si et al., 2013, 2014) and financial stance detection (Conforti et al., 2022). Recent existing work also suggests that CSR communication is present on Twitter, including work from Pilgrim and Bohnet-Joschko (December 2022) surveying existing reported-on CSR strategies in digital media and Johnson and Greenwell (2022) analyze 200+ UK companies and the practice of greenwashing (defined as when a company presents itself as environmentally-friendly, even when its actions actually say otherwise), yielding no evidence for greenwashing across UK companies, but signs that environmental messaging occurs with low frequency on company Twitter accounts.

Rybalko and Seltzer (2010) and Okazaki et al. (2020) took a dialogic approach to analyzing CSR communications on Twitter by focusing on dialogue between brands and Twitter users, as encouraged in public relations literature (Kent and Taylor, 1998). Rybalko and Seltzer (2010) found that *Fortune 500* companies tend to underuse dialogue to engage their stakeholders, while Okazaki et al. (2020) found that companies mostly were not explicitly using CSR to engage on Twitter. These two works inform our strategy for examining our own corpus: we focus on company tweets that are not replies or retweets, and we use a many-topic topic model to try to access more diffused CSR-related themes.

Salvatore et al. (2022) use a structural topic model (a weakly supervised approach) to explore how businesses used social media to communicate CSR, specifically in relation to the Sustainable Development Goals set by the United Nations' 2030 agenda, using tweets from the 30 largest firms according to the Dow Jones Industrial Average in August 2020. While our findings echo the focus on social and environmental issues for these companies, we broaden our focus to S&P 500 companies and use an unsupervised topic model.

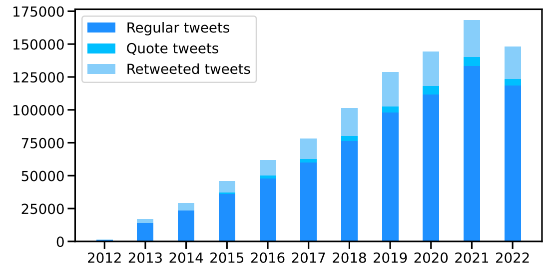


Figure 1: The composition of the initial dataset, broken down by tweet type and year, showing how API limits of 3,200 tweets per company reduced data availability for earlier years of our dataset.

3. Data

3.1. Collection Process

To gather Twitter handles for our companies, we scraped the websites of S&P 500 companies as listed on Wikipedia for all front-page links to Twitter handles. We augmented these Twitter handles with those listed in Twitter profiles for these companies. After manual vetting, we added obvious missing firm accounts, e.g. Match Group's subsidiaries. We excluded customer support Twitter accounts as well as regional accounts that were not immediately listed by companies on their website. With our list of S&P 500 Twitter handles, we used the Twitter API to retrieve as many tweets as possible from each company's Twitter account, going back at most 10 years from November 2022. Only data from more recent years was available for more prolific accounts due to the 3,200-tweet API limit for account history. Tweets in languages other than English were filtered out using the `fasttext-langdetect` library (Joulin et al., 2016b,a).

3.2. Composition

The initial dataset included 1,009,703 tweets from 536 distinct Twitter accounts. The dates of the tweets range from December 2012 to November 2022, with the breakdown of tweets by year shown in Figure 1. The parent companies of the Twitter accounts represent 11 distinct GICS Sectors, including Financials, Information Technology, Energy, Industrials, Consumer Staples, Health Care, Utilities, and Real Estate.

Tweets were tokenized using the Tweet Tokenizer from the Natural Language Toolkit (Bird et al., 2009), and tags to other users were replaced with "@TAG@" to prevent the formation of topics purely centered around tags. All terms were lowercased and stripped of trailing whitespace. Terms from the NLTK built-in English stoplist were filtered out in training.

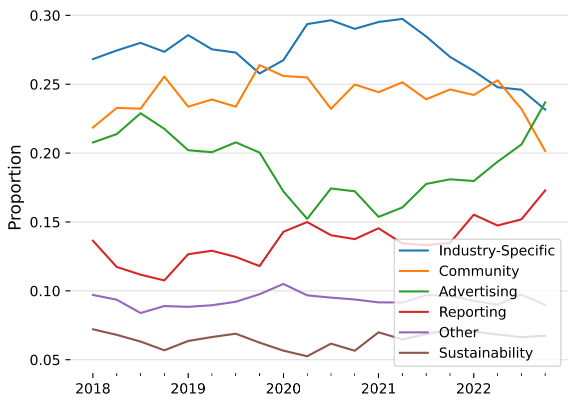


Figure 2: In 2018-2022, trends in the relative proportion of tweet categories vary, but suggest a possible shift of emphasis in recent years to include more corporate reporting (R) in addition to substantial discussion of community (C), with a pandemic-era dip in more straightforward advertising (A).

To focus on language from company accounts, we filtered out replies and retweets that were not "quote tweets", i.e., tweets that comment on an existing tweet. Finally, we pruned our vocabulary to terms (delimited by whitespace or punctuation) used by at least two companies. Without this processing, company-specific hashtags and terms overwhelmed our model. This brought the resulting vocabulary size from ∼1.36M to ∼52.5K across ∼827K tweets. The final dataset contains 827,403 tweets from 525 distinct Twitter accounts.

4. Analysis

4.1. Topic Model

To understand themes in our data, we wanted to find an unsupervised representation of similarity in our documents. We explored different models to identify themes in our corpora, including LDA topic modeling (Blei et al., 2003) and Sentence-BERT (Reimers and Gurevych, 2019) tweet embeddings. From our initial analysis, we decided to focus our work on a Biterm Topic Model (BTM) (Yan et al., 2013), which replaces the use of term-document frequencies with word co-occurrences in a 15-word window to respond better to shorter texts than LDA. The model outputs topics, or probability distributions over our vocabulary, with terms being allowed to have nonzero probability across multiple topics, and can be used to represent tweets as mixtures of topics. We trained our model on the full corpus of ∼827K tweets with 50 topics.

4.2. Topics

For each topic in our model, we inspected the 50 words of highest probability, the 10 tweets with the highest proportion of that topic, and top 10 Twitter accounts (by proportion of the account's tweets which were over a threshold for that topic). To develop themes across the topics, three authors manually labeled each topic, meeting in-person to resolve disputes on individual labels. Since the topic model captures all tweets in this period, we expect some topics to be less coherent; however, the authors did their best to understand why these words may have been grouped together using sample tweets. After labeling was established, we inductively developed six high-level topic categories to group related topics: **Industry-Specific Speech**, **Advertising**, **Corporate News and Reporting**, **Community and People**, **Sustainability**, and **Other** (which includes unclear or incoherent topics). We summarize these topics in Table 2 in the appendix. We plot data from years where we observed at least 100K total tweets before filtering. In this time period (2018-2022), we observe that there is a growth in corporate- and socially-focused speech, as shown in Figure 2.

We verify the existence of expected CSR themes anticipated by Stanislavská et al. (2023) around the environment, including (i) Sustainability (Topic 33 and 36), (ii) Climate (Topic 8), and (iii) Waste (Topics 12 and 40). We also see that keywords alone can be somewhat confusing for analysis: both topics 8 and 40 contain the words "sustainable" or "sustainability" 3 times within their top 50 words, but the difference is in how they use the word. Topic 8 focuses more on company announcements related to their sustainability efforts and goals (e.g. top document 10: "See our sustainability goals and progress achieved: https://t.co/GUUZEhJA26", @PPG), while topic 40 focuses on information and promotion of healthy sustainable practices (e.g. top document 9: "What's the wastewater and recycling connection? #WorldWaterDay https://t.co/qT8JU24eJR", @amwater). While the top companies in Topic 8 focus on sustainability (including Trane Technologies, Sempra, and NextEra Energy), we find The Coca-Cola Company (@CocaColaCo) ranked 7th for the topic, a company with a documented history of both a strong public CSR strategy and a record of significant environmental and social harm (Karnani, 2014). Similar overlaps occur in vocabulary for discussions of energy: Topics 33 and 36 focus on clean energy, while Topic 41 is focusing on energy production and Topic 35 mentions energy in the context of powering electronics.

The *Community and People* category also includes both internally-focused speech on excellent workplaces (e.g. topic 9, which revolves around

Predictors of High Risk		Predictors of Low Risk	
Topic 41 (58.67)	Topic 3 (27.81)	Topic 12 (-28.87)	Topic 13 (-27.97)
new energy gas million	proud support employees communities	new make help packaging	new industry learn latest

(a) Total Risk

Predictors of High Risk		Predictors of Low Risk	
Topic 41 (101.01)	Topic 8 (44.72)	Topic 10 (-33.07)	Topic 24 (-31.77)
new energy gas million	energy sustainable climate global	health help access care	culture work inclusive diversity

(b) Environmental Risk

Predictors of High Risk		Predictors of Low Risk	
Topic 16 (70.53)	Topic 28 (45.86)	Topic 12 (-80.04)	Topic 13 (-66.94)
people help work world	patients help treatment disease	new make help packaging	new industry learn latest

(c) Social Risk

Predictors of High Risk		Predictors of Low Risk	
Topic 30 (68.05)	Topic 16 (39.93)	Topic 13 (-37.41)	Topic 12 (-32.37)
risk help global companies	people help work world	new industry learn latest	new make help packages

(d) Governance Risk

Table 1: The strongest predictors of risk scores. This includes the top two topics that are the best predictors of high risk and the top two topics that are the best predictors of low risk, along with their respective top words and (parenthetical) regression coefficients.

highlighting workplace recognition and achievement) and external-focused communication (e.g. topic 47, which highlights supporting, donating to, and volunteering work). Prior work by Pilgrim and Bohnet-Joschko (December 2022) highlights social themes in CSR as a particular focus in digital media in ways that echo our topics, including the categories of (i) employee relations (Topics 21 and 26), (ii) diversity and inclusion (Topic 24), (iii) local community engagement (Topics 3 and 47), and (iv) philanthropy (Topic 43). We also see less specific socially-oriented topics like Topic 37. With simpler terms including "new", "help," "customers," and later "world", Topic 37 is led by McDonald's, and then immediately followed by multiple defense contractors and energy companies including Raytheon, HII, Lockheed Martin, and General Dynamics. This language connection between companies in seemingly unrelated industries suggests a possible trend of broad tweets about "helping the world", perhaps distinguishing a public CSR posture from concrete action and investment.

4.3. ESG Correlation

To understand how the learned topics from the Twitter corpus reflect corporate actions, we test whether topic proportions are predictive of 2022 Environmental, Social, and Governance (ESG) scores. These scores quantitatively describe Sustainalytics' assessment of companies based on exposure and management approaches to ESG risks. While Berg et al. (2022) show that ESG scores can disagree between sources, they highlight Sustainalytics as having the highest average correlation across other ESG metrics considered in their study. A low combined ESG risk score (<20) indicates positive work done towards managing ESG risks, while a higher risk score (>30) indicates greater concern. We use both the combined ESG score for each firm and three separate scores for Environment, Social, and Governance. We rescale each of these scores to a 0-100 scale for clarity of comparison. We obtained these scores for 453 of our companies via Yahoo Finance.

We represent each company using a 50-dimensional vector, where the i^{th} element is the proportion of the company's tweets in topic i. We then used ridge regression and Leave-One-Out (LOO) cross validation to try to predict both combined ESG and separate E, S, and G scores for each company. When computing regression, the ESG scores were all scaled to be from 0-100, by multiplying the environmental risk scores by 2 and the social and governance risk scores by 4. While fit was strongest for the environmental risk scores ($R^2 = 0.5$, RMSE $= 7.8$), it was weaker for the other two components, social ($R^2 = 0.16$, RMSE $= 13.4$) and governance ($R^2 = 0.16$, RMSE $= 7.94$), as well as for total risk ($R^2 = 0.22$, RMSE $= 6.2$). When compared to a baseline of predicting the risk as the averaging risk scores across the sector in our data, we see that only environmental scores are better predicted by our topic model than by the baseline

($R^2 = 0.4$, RMSE $= 8.9$).

However, even with low correlation, we still can find some meaningful trends in some of our topic features. We used coefficients from the regression model to find which topics were most predictive of high risk scores (most positive coefficients), as well as which predicted lower risk scores (most negative). The top words in these topics for prediction on each E/S/G score and their respective regression coefficients are presented in Table 1. While these are the most extreme, many more topics were also significant; from a permutation test, we determined that coefficients above 0.5 or below -1 were unlikely to be a result of random variation.

Unsurprisingly, we found the strongest predictor of high total ESG risk score was Topic 41, related to gas and energy companies, with a coefficient of 58.67. However, the next three highest predictors were topics that highlighted community “support” and “help” in abstract terms (Topic 3, 16, and 47, with coefficients 27.81, 21.26, 19.26). In contrast, we found that topics that related to concrete sustainable development practices and financial transparency correlated to lower risk scores.

Following Topic 41, the highest predictors of environmental risk was Topic 8, which contained speech relating to sustainability efforts, including #sustainability. The fact that a sustainability-focused topic indicates higher, not lower, risk, suggests that topic 8 is actually capturing “greenwashing” by companies to combat concerns about their climate practices. The other highest predictors also overlapped with those of total risk (Topics 3, and 47). The topics that indicated low environmental risk, Topics 10 (-33.07), 24 (-31.77), 4 (-30.18), and 30 (-27.96), contained more concrete words relating to healthcare, diversity and inclusion, and transparency about company finances.

Surprisingly, one of the highest predictors of social risk was relating to medical treatment and disease, potentially pointing to the complexity of intersecting profit motives with life-saving interventions. Topics that predicted a lower social risk included discussion of sustainable development (Topic 12) and technology reporting (Topic 13), with words inviting information exchange like "discuss", "opportunities", "solution", and "learn".

Finally, high governance risk was predicted by Topic 30 (68.05), containing words about corporate financial risk and economic impacts, as well as topics about community recognition (Topics 16, 43). Conversely, topics about technology solutions and sustainability that contained explicit references to environmental issues ("reduce", "carbon", "clean", "air", "emissions", "renewable") indicated a lower corporate governance risk.

5. Conclusion

In our work so far, we have collected a large corpus of corporate speech across 10 years of Twitter accounts for S&P 500 companies and trained a topic model to find patterns of discussion around CSR-focused themes. We found signs of both genuine reporting on CSR action from companies and cheap talk. The less explicit CSR focused topics correlated with increased ESG risk, especially those related to environmental concerns. These findings suggest that firms might be using communications about CSR as marketing strategies without fully investing in sustainability. We hope in our continuing work to reason further about individual variation in company language and concreteness/vagueness over time, as well as to compare Twitter behavior with spending data to show what distinguishes messaging of firms that invest funds towards sustainability and social good.

6. Acknowledgments

We would like to thank Eloise Burtis and the Pomona ITS team for their computing and data support. We thank our ECONLP 2024 reviewers for their kind reviews.

7. Bibliographical References

Theo B. Araujo and Jana Kollat. 2018. Communicating effectively about CSR on Twitter: The power of engaging strategies and storytelling elements. *Internet Research*, 28:419–431.

Najah Attig and Abdlmutaleb Boshanna. 2023. ESG washing: when cheap talk is not cheap! *Available at SSRN: https://ssrn.com/abstract=4672511*.

Andrew Baker, David F Larcker, Charles McClure, Durgesh Saraph, and Edward M Watts. 2023. Diversity washing. *Chicago Booth Research Paper*.

Florian Berg, Julian F Kölbel, and Roberto Rigobon. 2022. Aggregate Confusion: The Divergence of ESG Ratings. *Review of Finance*, 26(6):1315–1344.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.".

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Howard R. Bowen. 2013. *Social Responsibilities of the Businessman*. University of Iowa Press.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2022. Incorporating stock market signals for Twitter stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4074–4091.

Richard M Crowley, Wenli Huang, Hai Lu, and Wei Luo. 2019. Do firms manage their CSR reputation? Evidence from Twitter. *Singapore Management University School of Accountancy Research Collection*, pages 1–44.

Elisabet Garriga and Domènec Melé. 2004. Corporate social responsibility theories: Mapping the territory. *Journal of Business Ethics*, 53:51–71.

Thomas F Johnson and Matthew P Greenwell. 2022. Are companies using Twitter to greenwash and hide bad environmental performance? *Energy, Ecology and Environment*, 7(3):213–226.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016a. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Aneel Karnani. 2014. Corporate social responsibility does not avert the tragedy of the commons. case study: Coca-Cola India. *Economics, Management, and Financial Markets*, 9(3):11–23.

Michael L Kent and Maureen Taylor. 1998. Building dialogic relationships through the world wide web. *Public Relations Review*, 24(3):321–334.

Abagail McWilliams and Donald S. Siegel. 2000. Corporate social responsibility and financial performance: correlation or misspecification? *Strategic Management Journal*, 21:603–609.

Shintaro Okazaki, Kirk Plangger, Douglas West, and Héctor D. Menéndez. 2020. Exploring digital corporate social responsibility communications on Twitter. *Journal of Business Research*, 117:675–682.

Marc Orlitzky, Frank L. Schmidt, and Sara L. Rynes. 2003. Corporate social and financial performance: A meta-analysis. *Organization Studies*, 24:403 – 441.

Katharina Pilgrim and Sabine Bohnet-Joschko. December 2022. Corporate social responsibility on Twitter: A review of topics and digital communication strategies' success factors. *Sustainability*, 14(24).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *arXiv preprint arXiv:1908.10084*.

Svetlana Rybalko and Trent Seltzer. 2010. Dialogic communication in 140 characters or less: How fortune 500 companies engage stakeholders using Twitter. *Public Relations Review*, 36(4):336–341.

Camilla Salvatore, Silvia Biffignandi, and Annamaria Bianchi. 2022. Corporate social responsibility activities through Twitter: From topic model analysis to indexes measuring communication characteristics. *Social Indicators Research*, 164(3):1217–1248.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. 2013. Exploiting topic based Twitter sentiment for stock prediction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Sinno Jialin Pan, Qing Li, and Huayi Li. 2014. Exploiting social relations and sentiment for stock prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1139–1145.

Lucie Kvasniková Stanislavská, Ladislav Pilař, Xhesilda Vogli, Tomá Hlavsa, Katerina Kuralová, Abby Feenstra, Lucie Pilařová, Richard Hartman, and Joanna Rosak-Szyrocka. 2023. Global analysis of Twitter communication in corporate social responsibility area: sustainability, climate change, and waste management. *PeerJ Computer Science*, 9.

US SIF Foundation. 2020. Report on US sustainable and impact investing trends 2020. Technical report, US SIF Foundation.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 1445–1456, New York, NY, USA. Association for Computing Machinery.

A. Full Topic List

The table in the following pages summarizes the topics by high-probability words and prominent companies. We include both our find-grained labels and our broader categorization of these topics: textbfIndustry-Specific Speech (I), **Advertising** (A), **Corporate News and Reporting** (R), **Community and People** (C), **Sustainability** (S), and **Other** (O).

#	Label	Top Words	Top Companies
0 (I)	Healthcare Sector	help care health new support provide program people access make	@ShopSimon @Take2Interactiv @tuicruises @RealtyIncome @EXPD_Official
1 (I)	Biotech	clinical new research development patients help cell drug discuss learn	@CatalentPharma @Incyte @CRiverLabs @CorningLifeSci @moderna_tx
2 (O)	Social Media	new latest episode shares future video look trends social blog	@nielsen @expediamedia @CBOE @GoldmanSachs @TrimbleCorpNews
3 (C)	Community Support	proud support employees communities students local community commitment efforts help	@VentasREIT @comcast @DevonEnergy @HeyCisco @PPLCorp
4 (R)	Financial Reporting	market · global prices supply bond demand high rose economic	@SPGlobal @Prologis @ICE_Markets @TRowePrice @MoodysInvSvc
5 (C)	Community Support	help work employees make better technology improve business new people	@Paycom @ServiceNow @Ceridian @kroger @Paychex
6 (C)	Community Support	health people help mental impact support care safety work important	@ElevanceHealth @Cigna @ViatrisInc @Humana @Centene
7 (R)	News Reporting	global markets economic market impact new credit pandemic growth insurance	@TRowePrice @MoodysInvSvc @BlackRock @FTI_US @GoldmanSachs
8 (S)	Sustainability	energy sustainable climate global future sustainability emissions carbon commitment #sustainability	@Trane_Tech @mhkgreenworks @sempra @nexteraenergy @Edison_Energy
9 (C)	Recognition	proud named year recognized honored list celebrate 2021 years 100	@KeurigPepper @Omnicom @VentasREIT @nexteraenergy @DowNewsroom
10 (I)	Healthcare Sector	health help access care support healthcare improve provide resources program	@Centene @cvshealth @ElevanceHealth @UnitedHealthGrp @ViatrisInc
11 (O)	Other	✓ help need new right business information know make online	@KelloggsUS @AskAmex @InsidePMI @VERISIGN @AltriaNews
12 (S)	Products and Packaging	new make help packaging products like food - work team	@WestRock @packagingcorp @BallCorpHQ @Sealed_Air @IntlPaperCo
13 (R)	Technology Reporting	new industry learn latest technology digital trends help supply experts	@McKesson @Gartner_Inc @PTC @healthcare_abc @Applied4Tech
14 (A)	Positive Advertising	new holiday favorite season time like make just best perfect	@Ross_Stores @LambWeston @RealReddiWip @bathbodyworks @smuckers
15 (I)	Information Technology	new business digital help data learn technology latest customer financial	@FISglobal @Fiserv @Broadridge @QuickBooks @StateStreet
16 (C)	People	people help work world make women like support helping we're	@Meta @tuicruises @3M @Intuit @AbbottNews
17 (O)	Other	make food like help time people know way water new	@VlasicStork @KeurigPepper @OpenTable @smuckers @pizzahut
18 (I)	Information Technology	data business security digital help learn discuss key organizations leaders	@Fortinet @Gartner_Inc @Protiviti @DXCTechnology @Equinix
19 (O)	Other (Short Hashtags)	- | / + – new love favorite great like	@skyworksinc @DukesMeats @ChipotleTweets @newell_brands @InvitationHomes
20 (R)	Financial Reporting	financial growth results quarter earnings market 2021 year strong new	@WECEnergyGroup @RealtyIncome @FactSet @MarathonOil @mhkgreenworks
21 (C)	People	new look team looking forward learn experience great time -	@iTeroScanner @poolcorp @IFF @amphenol @AmericanAir
22 (I)	Information Technology	new solutions technology data learn design help software technologies digital	@NXP @L3HarrisTech @ANSYS @MicrochipTech @Qualcomm
23 (I)	Financial Sector	volume near options • contracts trading futures million day term	@MarketAxess @CBOE @CMEGroup @ICE_Markets @FactSet
24 (C)	Diversity and Inclusion	culture work inclusive diversity diverse employees inclusion commitment women create	@Intuit @VentasREIT @ADP @KeurigPepper @Accenture_US

25 (I)	Biotech	learn new help using - design webinar cell process booth	@mettlertoledo @BioRadLifeSci @CorningLifeSci @BioRadFlowAbs @WatersCorp
26 (C)	Recognition	support team employees honor help work service thank members military	@HCAhealthcare @sbasite @ONEOK @UHS_inc @genuinepartsco
27 (O)	Other (Quantities)	- new million customers products years 2 support provide –	@PACCARFinancial @MarriottBonvoy @CharterNewsroom @Prologis @skyworksinc
28 (I)	Surgery, Medicine	patients help treatment disease heart care people risk patient cancer	@zimmerbiomet @DaVita @IntuitiveSurg @Abiomed @Hologic
29 (A)	Advertising	new look latest series collection – features — featuring iconic	@RalphLauren @Delta @EsteeLauder @Silversea @CarnivalPLC
30 (R)	Financial Reporting	risk help global companies risks impact financial challenges health need	@mercer @MarshGlobal @MarshMcLennan @GuyCarpenter @BRINKNewsNow
31 (C)	Power Service	power customers help stay weather safety crews safe outages report	@DominionEnergy @PSEGdelivers @EversourceMA @DTE_Energy @DukeEnergy
32 (O)	Other (Informal)	time tips know make help just sure you're need home	@Invisalign @OurTimeDating @hinge @KelloggsUS @Discover
33 (S)	Sustainable Energy	new team power energy help - future solar water make	@Enphase @nscorp @CSX @SolarEdgePV @CrownCastle
34 (A)	Events	learn today live event discuss - virtual join booth miss	@AristaNetworks @FactSet @ONEOK @LiveNation @IntuitiveSurg
35 (I)	Technology	power help new solutions make energy electric technology safety learn	@LKQCorp @autozone @IRProducts @ParkerHannifin @monolithicpower
36 (S)	Energy Responsibility	energy new help emissions reduce gas power carbon save electric	@Enphase @SolarEdgePV @Humana @EversourceMA @DTE_Energy
37 (A)	Advertising	new help customers look support business meet team make world	@McDonalds @RaytheonTech @WeAreHII @LockheedMartin @DukeEnergy
38 (A)	Advertising	win - chance time day booth sure ready new just	@exocad @UPS @AmericanAir @SlimJim @MonsterEnergy
39 (I)	Technology	power data new energy help performance solution customers network solutions	@monolithicpower @TXInstruments @MicrochipTech @Equinix @SEAGATE
40 (S)	Energy and Waste	water energy help gas reduce natural use waste air clean	@Pentair @RepublicService @amwater @Xylem @AOSmithHotWater
41 (R)	Energy Reporting	new energy gas million years largest - announced facility production	@Lindeplc @conocophillips @KeurigPepper @Kinder_Morgan @northropgrumman
42 (I)	Home Renovation	home like new space tips make kitchen room perfect living	@Lennar @PulteHomes @DRHorton @HomeDepot @LarsonDoors
43 (C)	Recognition, Announcements	proud excited team new announce support work students share sponsor	@TruistNews @tuicruises @genuinepartsco @Allstate @CaesarsEnt
44 (C)	Workplace	help work employees new career people business make need talent	@mercer @roberthalf @Paycom @Paychex @CamdenLiving
45 (I)	Healthcare	new data learn drug help using testing use development clinical	@thermofisher @WestPharma @WatersCorp @CatalentPharma @PerkinElmer
46 (R)	News, Announcements	new - people shares know latest like help learn —	@HeyCisco @travelocity @kroger @Orbitz @HLCruises
47 (C)	Community Support	help food support employees local team families communities million donated	@molinahealth @ConagraBrands @KelloggCompany @IDEXCorp @IntlPaperCo
48 (R)	Financial Reporting	latest new report year - 2021 credit 10 – impact	@VERISIGN @turbotax @creditkarma @TheHartford @CFIndustries
49 (A)	Time, Dating	. time years day team love @ summer #dating	@united @SherwinWilliams @kroger @OurTimeDating @Match

Table 2: The top 10 words and top 5 accounts for each topic. Each topic is hand-labeled with an approximate subject for the topic. Top words that include non-visible ASCII characters have been omitted, and the first 10 words with visible characters are included.

LLaMA-2-Econ: Enhancing Title Generation, Abstract Classification, and Academic Q&A in Economic Research

Onur Keleş, Ömer Turan Bayraklı
Boğaziçi University, Istanbul University
Department of Linguistics, Department of Econometrics
onur.keles1@bogazici.edu.tr, omerturanbayrakli@ogr.iu.edu.tr

Abstract

Using Quantized Low Rank Adaptation and Parameter Efficient Fine Tuning, we fine-tuned Meta AI's LLaMA-2-7B large language model as a research assistant in the field of economics for three different types of tasks: title generation, abstract classification, and question and answer. The model was fine-tuned on economics paper abstracts and syntheticically created question-answer dialogues based on the abstracts. For the title generation, the results of the experiment demonstrated that LLaMA-2-Econ (the fine-tuned model) surpassed the base model (7B and 13B) with few shot learning, and comparable models of similar size like Mistral-7B and Bloom-7B in the BLEU and ROUGE metrics. For abstract categorization, LLaMA-2-Econ outperformed different machine and deep learning algorithms in addition to state-of-the-art models like GPT 3.5 and GPT 4 with both single and representative few shot learning. We tested the fine-tuned Q&A model by comparing its output with the base LLaMA-2-7B-chat with a Retrieval Augmented Generation (RAG) pipeline with semantic search and dense vector indexing, and found that LLaMA-2 performed on a par with the base model with RAG.

Keywords: LLaMA-2, economics, SFT, QLoRA, PEFT

1. Introduction

The evolution of neural networks like RNN (Rumelhart et al., 1986), and LSTM (Hochreiter and Schmidhuber, 1997) architectures and later the invention of the transformer architecture (Vaswani et al., 2017) paved the way for the development of the state-of-the-art Large Language Models (LLMs) such as GPT 3.5 (Ouyang et al., 2022), ChatGPT-4[1] by OpenAI, Gemini (Team et al., 2023) by Google or popular open-source LLMs such as LLaMA-2 (Touvron et al., 2023) by Meta AI, Bloom (Scao et al., 2022), Mistral (Jiang et al., 2023), OPT (Zhang et al.), GPT Neo (Black et al., 2021) and Bart (Lewis et al., 2019) especially for text generation tasks with causal language modeling. Earlier models were trained largely on general corpora (e.g., Wikipedia and books) but now there are a myriad of attempts at injecting open-source LLMs with domain-specific knowledge, including transformers pre-trained on medical and biomedical (Lee et al., 2020), financial (Peng et al., 2021), and scientific text (Beltagy et al., 2019).

Fine-tuning is a process where a pre-trained language model, like BERT (Devlin et al., 2018) or GPT, is specialized for a specific task by further training it on a related dataset, enhancing its performance in the target domain. Furthermore, with the advent of newly emerging methodologies such as Retrieval Augmented Generation (RAG) (Lewis et al., 2020), language models can easily retrieve information and use external data sources. Also, the introduction of Low Rank Adaptation (LoRA) (Hu et al., 2021) and more recently Quantized Low Rank Adaptation (QLoRA) (Dettmers et al., 2023) significantly reduced parameters with less memory needed, enabling training on smaller hardware and faster training times and helped with scalability. Likewise, Parameter Efficient Fine Tuning (PEFT) techniques helped optimize LLMs in terms of applicability across various domains/tasks by fine-tuning only a subset of base model parameters. Although there are a few number of attempts made for large-scale domain or task adaptation purposes (Gema et al., 2023), we acknowledge that there is a scarcity of work dedicated to fine-tuning an open-source LLM assistant in economics for specific research tasks like title generation, abstract classification, and open-ended question & answer (Q&A). To address this gap, this work introduces LLaMA-2-Econ, a PEFT adapted version of the open-source LLaMA-2-7B model by Meta AI, fine-tuned on economics paper abstracts and synthetically created question-answer data for research tasks, specifically for title generation, abstract classification and academic open-ended Q&A.

1.1. Related Work

There are a number of applications of LLMs on related tasks like news headline generation (Gavrilov et al., 2019) or summary generation (Xiao and Chen, 2023). Previous classification methods using transformers include few-shot financial text classification with LLMs like ChatGPT (Loukas et al., 2023b,a). In particular, Loukas et al. demonstrated

[1] http://chat.openai.com

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 212–218

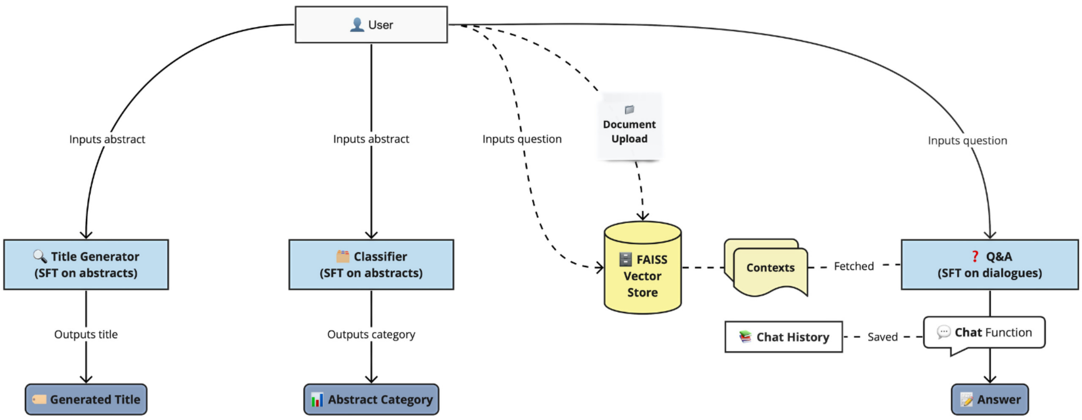


Figure 1: Proposed Workflow for LLaMA-2-Econ. A user inputs an abstract to generate a title and categorize the content using our specialized supervised fine-tuned models. Questions asked by the user are answered through an interactive QA chatbot system that optionally retrieves information from a document vector store, with interactions saved in the chat history.

autoregressive models like GPT 4 can surpass MLM models in text classification. However, their own fine-tuned MP-Net model achieved comparable results in such domain specific tasks. Despite the paucity of research in adapting LLMs in economics, one important contribution is the FinBERT models (Araci, 2019; Yang et al., 2020), BERT-based models trained for financial NLP tasks to tackle financial sentiment analysis and classification problems, outperforming previous state-of-the-art models.

There has been further exploration into financial sentiment analysis by analyzing sentiments in cryptocurrency-related social media posts (Kulakowski and Frasincar, 2023). The authors introduced CryptoBERT, a model fine-tuned on the cryptocurrency domain from BERTweet, and LUKE, a language-universal cryptocurrency emoji sentiment lexicon, to address the challenges in sentiment analysis across languages in social media, and providing tools for enhancing quantitative trading models with sentiment analysis of social media.

As for Q&A, a significant domain adaptation work is PaperPersiChat, which is an open chat-bot designed for discussing scientific papers for computer science (Chernyavskiy et al., 2023). The authors incorporated summarization and Q&A within a single end-to-end online chat-bot pipeline. They trained a dialogue system with scientific grounding. Finally and more relevantly, a recent work employed a PEFT/LoRA based approach for LLaMA-2 fine-tuning in a multitask financial news analysis, and the experimental results showed that the fine-tuned model performs various tasks like main point highlighting, text summarization, and named-entity extraction with sentiments (Pavlyshenko, 2023). Overall, It is clear that LLMs can prove to be helpful agents in (economic) research, performing tasks ranging from paper summaries, generating headlines and text classification to synthesizing information and editing (Korinek, 2023; Dowling and Lucey, 2023; Horton, 2023). However, most available applications for such tasks are not open-source, and there is a lack of research integrating especially decoder-only and open-source LLMs and economics.

2. Methodology

To this end, this paper will attempt at the following: (i) fine-tune LLaMA-2-7B, an open-source and decoder-only model, for the tasks of paper title generation and abstract classification (econometrics, general economics, and theoretical economics) and LLaMA-2-7B Chat for open-ended academic Q&A with QLoRA and PEFT; (ii) perform experiments on metrics to test fine tuned model against the baseline and other language models for these tasks; (iii) propose a Web application acting as a research assistant in economics, utilizing the fine-tuned models with these tasks and an end-to-end chatbot with RAG integration (Figure 1).

2.1. Data

We obtained the data with the arXiv API[2] and searched for economics papers in the following categories/classes: ec.EM (econometrics), ec.GN (general economics), and ec.TH (theoretical economics). In addition to the category, title, abstract, and other metadata were added to our dataset. We preprocessed the data and filtered out low-quality samples following a manual inspection. In the end, we obtained 6362 samples for the train dataset and 707 for the test dataset (Figure 2).

[2] https://arxiv.org/help/api/index

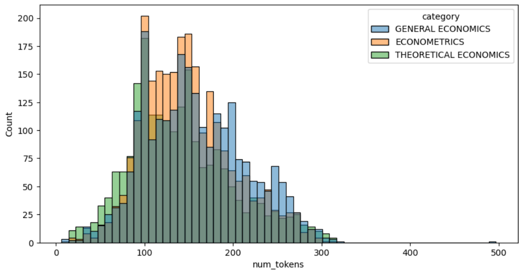


Figure 2: Number of tokens in the training data and their categorical distribution

2.1.1. Creation of Q&A Synthetic Data

Following research that creates synthetic data with state-of-the-art closed source models like ChatGPT (Askari et al., 2023), we have fine-tuned LLaMA-2-7B Chat with a syntheticically created question-and-answer dialogue dataset from academic paper abstracts (7079 in total), employing an approach that utilizes GPT 3.5 Turbo model (costs $0.002 per 1K tokens) from OpenAI that to train the question and answer component of the research assistant. We generated contextual dialogues, where the model both acts as an assistant and user, posing questions and providing answers relevant to a given abstract. Per abstract, we generated 2 to 3 questions and answers. We then filtered out low quality samples, short and incorrectly parsed dialogues, and ended up with 3340 pairs. A sample Q&A pair is given below:

> **What distinguishes revenue management systems? A:** Key characteristics include *fixed capacities*, *homogeneous products*, and sensitivity to *customer pricing decisions*.

> **What's problematic with current policy-making indicators? A:** They often subjectively combine a limited number of indicators, overlooking crucial *inter-indicator relationships*.

2.2. Fine Tuning

We fine-tuned Meta AI's LLaMA-2-7B model[3] for the title generation and abstract classification tasks, and LLaMA-2-7B Chat[4] (reinforcement learning with human feedback) using the `transformers` library (Wolf et al., 2020) on a NVIDIA A100 GPU. For the fine-tuning, we used Quantized Low Rank Adaption (QLoRA) with a `lora_r` of 64 and a `lora_dropout` of 0.1. To enhance computational efficiency, we utilized 4-bit precision with a computation dtype of `float16` and the quantization type was set to `nf4`, with nested quantization enabled. The models were scheduled to train for 8 epochs with an early stopping patience of 2 epochs with `bf16` training. Gradient checkpointing and a maximum gradient norm of 0.3 was used. The learning rate was initialized at $2e-4$, using a cosine learning rate scheduler and a warmup ratio of 0.03. Sequences were grouped by length for efficiency and we employed paged AdamW (Loshchilov and Hutter, 2017) with 32-bit precision as the optimizer. The batch size, maximum input and target length were selectively optimized for each task and model.

The PEFT technique we use here integrates fine-tuned components, specifically LoRA weights, into a baseline model, conserving computational resources while keeping the model's task-specific performance. After reloading the model in FP16 for better efficiency and setting up the tokenizer with precision, we then merge these enhancements with the baseline model. This crucial step ensures that our fine-tuning efforts are fully integrated, enhancing the model's overall efficiency and effectiveness. Combined with QLoRA, PEFT allowes for optimized fine-tuning performance and scalability. Our fine-tuned models and dataset are openly available on Huggingface[5].

3. Results

In this section, we report LLaMA-2-Econ's performance on BLEU and ROUGE metrics for the title generation task and compare the results with the baseline LLaMA-2-7B as well as LLaMA-2-13B, Mistal-7B, Bloom-7B and smaller open-source models like GPT Neo and OPT with few shot (5 for this task) learning. As for the classification, we computed the performance metrics and compared the results with those of GPT 3.5 and GPT 4 with one shot and representative few shot (one for each class) learning. We also trained and evaluated different machine learning (ML) and neural network (NN) classifiers. Finally, to evaluate our Q&A model, we measure similarity between LLaMA-2-Econ's generated answers and reference answers obtained through RAG with human verification.

3.1. Experiment 1: Title Generation

As can be seen from the results in Table 1, the fine-tuned model surpasses the baseline and other open-source LLMs of different sizes that use few shot learning. LLaMA-2-13B performs second best in these metrics, followed by other smaller size models.

[3] https://huggingface.co/meta-llama/Llama-2-7b
[4] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[5] https://huggingface.co/onurkeles/llama-2-7b-econ-abstract-classifier
https://huggingface.co/onurkeles/llama-2-7b-econ-title-generator
https://huggingface.co/onurkeles/llama-2-7b-econ-chat-qa

Model	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
LLaMA-2-Econ (ours)	**0.16**	**0.45**	**0.24**	**0.41**
LLaMA-2-7B (few shot)	0.10	0.41	0.18	0.36
LLaMA-2-13B (few shot)	0.12	0.40	0.19	0.36
Mistral-7B (few shot)	0.11	0.37	0.18	0.33
Bloom-7B (few shot)	0.10	0.37	0.16	0.33
GPT Neo-2.7B (few shot)	0.03	0.19	0.05	0.17
OPT-2.7B (few shot)	0.06	0.25	0.10	0.22

Table 1: Comparison of LLaMA-2-Econ with Other Models in Title Generation

Model	Accuracy	Precision	Recall	F1 Score
LLaMA-2-Econ (ours)	**0.88**	**0.88**	**0.88**	**0.88**
GPT 3.5 (one shot)	0.43	0.63	0.43	0.40
GPT 3.5 (few shot)	0.59	0.72	0.59	0.53
GPT 4 (one shot)	0.70	0.73	0.70	0.64
GPT 4 (few shot)	0.84	0.85	0.84	0.83
Decision Tree Classifier	0.77	0.72	0.77	0.71
K-Nearest Neighbors Classifier	0.79	0.79	0.79	0.79
Logistic Regression	0.85	0.86	0.85	0.85
Random Forest Classifier	0.85	0.86	0.85	0.85
SVC	0.86	0.86	0.86	0.86
XGB Classifier	0.83	0.83	0.83	0.83
RNN	0.81	0.81	0.80	0.81
LSTM	0.81	0.82	0.81	0.82

Table 2: Comparison of LLaMA-2-Econ with Other Models in Abstract Classification

3.2. Experiment 2: Classification

Table 2 shows that LLaMA-2-Econ outperformed other classifiers, having an F1 score of 0.88. Logistic Regression, Random Forest Classifier and SVC achieve comparable scores to our fine-tuned model, followed by GPT 4 with representative few shot (one for each class) and one shot learning, and other ML and neural models. GPT 3.5 both one and few shot (one for each class) performs worst in this abstract classification task.

3.3. Experiment 3: Q&A

As for the neural evaluation for our Q&A Model, we obtained reference open-ended answers to a subset of our synthetically created questions from the base chat model with RAG integration. Following human verification of the answers and inspection, we compared them with LLaMA-2-Econ's generated answers without RAG. We use BERT-Score (Zhang et al., 2019) as our evaluation metric, which calculates the cosine similarity between the embeddings of tokens in our generated answers and those in the reference answers. The formulas to calculate the precision (P), recall (R), and F1-score (F1) where S_{ij} is the similarity score between token i from the candidate answers and token j from the reference answers, $|C|$ is the total number of tokens in the candidate answers, and $|R|$ is the total number of tokens in the reference answers are:

$$P = \frac{1}{|C|} \sum_{i \in C} \max_{j \in R} S_{ij} \quad (1)$$

$$R = \frac{1}{|R|} \sum_{j \in R} \max_{i \in C} S_{ij} \quad (2)$$

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (3)$$

The generated answers by our LLaMA-2-Econ model without RAG (to the questions in the test dataset) received an average precision value of 0.90, recall value of 0.89, and F1 value of 0.90. This means that it achieved commendable similarity with human verified reference responses provided by a RAG implemented base chat model to academic open-ended questions in the domain of economics.

4. Proposed Workflow

Finally, we propose an open application (Figure 1) that can act as an online research assistant which will be openly available to researchers in economics by using open-source fine-tuned models with QLoRA and PEFT. For the chat module of the system, RAG and Facebook AI Similarity Search are employed as well as Langchain[6]'s loader libraries to allow users to load their own economics paper of their own choice or choose one from the provided database.

[6] http://langchain.com

5. Conclusion

In conclusion, we introduced the LLaMA-2-Econ model a QLoRA and PEFT-based model fine-tuned for specific research tasks in the domain economics. Our fine-tuned model performed well in executing different research related tasks, as supported by the metrics achieved against baseline and other state-of-the-art model across various metrics. Our model was also successful in generating reference-like answers to academic questions related to economics research. Overall, we conclude that smaller adapted models with PEFT can be trained on small set of domain specific papers to perform personalized research tasks and obtain comparable results to larger or more advanced models. The integration of QLoRA and PEFT in this study has also shown that scaling large models to new tasks can be more accessible, as it can reduce the need for extensive computational resources. This, of course, further democratizes the use of LLMs in the social sciences, allowing more entities to fine-tune and deploy state-of-the-art models for their specific research needs.

6. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. 2023. A test collection of synthetic documents for training rankers: Chatgpt vs. human experts. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5311–5315.

Salvador Balkus and Donghui Yan. Improving short text classification with augmented data using GPT-3. pages 1–30.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Alexander Chernyavskiy, Max Bregeda, and Maria Nikiforova. 2023. PaperPersiChat: Scientific paper discussion chatbot using transformers and discourse flow management. In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 584–587.

Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. On the use of ArXiv as a dataset.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient fine-tuning of quantized LLMs.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Michael Dowling and Brian Lucey. 2023. Chatgpt for (finance) research: The bananarama conjecture. *Finance Research Letters*, 53:103662.

Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2023. Qlarify: Bridging scholarly abstracts and papers with recursively expandable summaries.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. 2020. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. In Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra, editors, *Advances in Information Retrieval*, volume 11438, pages 87–93. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Aryo Gema, Luke Daines, Pasquale Minervini, and Beatrice Alex. 2023. Parameter-efficient fine-tuning of llama for the clinical domain. *arXiv preprint arXiv:2307.03042*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Zhiqiang Hu, Yihuai Lan, Lei Wang, Wanyu Xu, Ee-Peng Lim, Roy Ka-Wei Lee, Lidong Bing, and Soujanya Poria. 2023. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models. *arXiv preprint arXiv:2304.01933*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models.

Anton Korinek. Language models and cognitive automation for economic research.

Anton Korinek. 2023. Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Mikolaj Kulakowski and Flavius Frasincar. 2023. Sentiment classification of cryptocurrency-related social media posts. *IEEE Intelligent Systems*, 38(4):5–9.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. Label supervised LLaMA finetuning.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lefteris Loukas, Ilias Stogiannidis, Odysseas Diamantopoulos, Prodromos Malakasiotis, and Stavros Vassos. 2023a. Making llms worth every penny: Resource-limited text classification in banking. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 392–400.

Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023b. Breaking the bank with ChatGPT: Few-shot text classification for finance.

Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13.

Bohdan M. Pavlyshenko. Financial news analytics using fine-tuned llama 2 GPT model.

Bohdan M Pavlyshenko. 2023. Financial news analytics using fine-tuned llama 2 gpt model. *arXiv preprint arXiv:2308.13032*.

Chandrashekhar S. Pawar and Ashwin Makwana. Comparison of BERT-base and GPT-3 for marathi text classification. In Pradeep Kumar Singh, Sławomir T. Wierzchoń, Jitender Kumar Chhabra, and Sudeep Tanwar, editors, *Futuristic Trends in Networks and Computing Technologies*, volume 936, pages 563–574. Springer Nature Singapore. Series Title: Lecture Notes in Electrical Engineering.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Is domain adaptation worth your investment? comparing bert and finbert on financial tasks. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 37–44.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Issey Sukeda, Masahiro Suzuki, Hiroki Sakaji, and Satoshi Kodera. JMedLoRA:medical domain adaptation on japanese large language models using instruction-tuning.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, {\textbackslash}Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

Sharon Whitfield and Melissa A. Hofmann. Elicit: AI literature review research assistant. 19(3):201–207.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-LLaMA: Towards building open-source language models for medicine.

Le Xiao and Xiaolin Chen. 2023. Enhancing llm with evolutionary fine tuning for news summary generation. *arXiv preprint arXiv:2307.02839*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models, 2022. *URL https://arxiv.org/abs/2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Multi-Lingual ESG Impact Duration Inference

Chung-Chi Chen,[1] **Yu-Min Tseng,**[2] **Juyeon Kang,**[3] **Anaïs Lhuissier,**[3]
Yohei Seki,[4] **Hanwool Lee**,[5] **Min-Yuh Day**,[6] **Teng-Tsai Tu**,[7] **Hsin-Hsi Chen**[8]
[1]AIST, Japan
[2] Data Science Degree Program, National Taiwan University and Academia Sinica, Taiwan
[3]3DS Outscale, France, [4]University of Tsukuba, Japan, [5]NCSOFT, South Korea
[6]Graduate Institute of Information Management, National Taipei University, Taiwan
[7]Graduate Institute of International Business, National Taipei University, Taiwan
[8]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

Abstract

To accurately assess the dynamic impact of a company's activities on its Environmental, Social, and Governance (ESG) scores, we have initiated a series of shared tasks, named ML-ESG. These tasks adhere to the MSCI guidelines for annotating news articles across various languages. This paper details the third iteration of our series, ML-ESG-3, with a focus on impact duration inference—a task that poses significant challenges in estimating the enduring influence of events, even for human analysts. In ML-ESG-3, we provide datasets in five languages (Chinese, English, French, Korean, and Japanese) and share insights from our experience in compiling such subjective datasets. Additionally, this paper reviews the methodologies proposed by ML-ESG-3 participants and offers a comparative analysis of the models' performances. Concluding the paper, we introduce the concept for the forthcoming series of shared tasks, namely multi-lingual ESG promise verification, and discuss its potential contributions to the field.

Keywords: argument relation, argument mining, cross-lingual

1. Introduction

In recent years, the Environmental, Social, and Governance (ESG) criteria have emerged as vital measures for evaluating a company's impact on the world. These criteria not only inform investors about the sustainability and ethical implications of investing in a company but also help consumers and employees align with organizations that share their values. However, accurately assessing a company's performance in these areas remains a complex challenge, exacerbated by the dynamic and multifaceted nature of ESG-related information. To address this challenge, our research community has initiated the ML-ESG series of shared tasks.

Given the increasing importance of ESG for accounting departments and investors, many rating companies have emerged, such as DJSI, CDP, FTSE, MSCI, and Sustainalytics. In the ML-ESG shared tasks series, we selected MSCI's rating standard for annotations on ESG-related news articles. In ML-ESG-1 (Chen et al., 2023a), we explored the ESG Issue Identification task. In ML-ESG-2 (Chen et al., 2023b), we focused on ESG Impact Type Identification. After understanding the issue (up to 44 aspects) and the type (opportunity or risk), ML-ESG-3 goes a step further to infer the impact duration. This task aims to estimate how long the effects of certain events or actions taken by a company will last, impacting its ESG scores. It involves not only interpreting the immediate effects of an event but also predicting its long-term consequences—something that even experienced human analysts find challenging.

This paper presents an overview of ML-ESG-3, including the datasets developed, and the insights gained from compiling these datasets. ML-ESG-3 includes news articles in five different languages, acknowledging the global nature of ESG issues and the importance of diverse linguistic representation in ESG analysis. Moreover, we summarize the methodologies proposed by participants in ML-ESG-3, offering an analysis of their models' performance. Finally, we conclude with a discussion on the next series of shared tasks, focusing on multi-lingual ESG promise verification. This forthcoming task is designed to further the field's understanding of how companies' promises regarding ESG performance align with their actual actions and impacts. By exploring the verification of these promises across different languages, we aim to enhance the transparency and accountability of companies on a global scale.

2. Dataset

2.1. Guidelines

The MSCI guidelines delineate the timeline for impact duration as follows: short-term is under 2 years, long-term is 5+ years, and medium-term encompasses the period in between. Given that all actions carry long-term consequences, the following advice is provided to avoid indiscriminately assigning the label "long" to each time frame:

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 219–227
20 May, 2024.

	Train			Test		
	within 2 years	2 to 5 years	longer than 5 years	within 2 years	2 to 5 years	longer than 5 years
Chinese	97	69	226	11	8	25
English	82	198	265	6	47	83
French	122	222	293	31	32	83
Korean	446	212	142	96	40	64
Japanese	15	7	5	291	167	715

Table 1: Statistics of impact duration dataset.

	Train			Test		
	High	Medium	Low	High	Medium	Low
English	196	243	106	60	59	17
French	198	317	122	45	53	48

Table 2: Statistics of impact-level.

- Pay attention to any time indications within the text, as these can serve as reliable indicators of the intended duration, such as references to political agendas or statements from scientists.
- Consider the subject matter of the sentence: if the focus is on contract negotiations or diplomacy rather than the issue itself, it may be appropriate to classify the paragraph as short-term, despite potential long-term benefits or harms.
- Recognize that some topics inherently imply a specific impact duration based on common sense. For issues that cannot be predicted with absolute certainty, opting for a safe, neutral mean or the most likely impact duration is advisable.
- In the absence of explicit date references or common-sense driven topics, focus on keywords that indicate the type of issue being discussed or the nature of the debate, rather than the overarching topic.

In addition to the impact duration, English and French datasets provide additional impact-level annotations. Since evaluating the impact of an event can be utterly subjective, to minimize this, here are some pieces of advice to remain objective and indications as to what could be considered low, medium, and high impact.

- Take into consideration the broader issue at stake and not only the discussed matter, to get a better picture of the potential impact.
- Reference similar previous events as a benchmark.
- National or international events do not always signify high impact. Decision-makers can take small steps towards their goals, and these should be assessed as such for the sake of our shared task.

	Train	Test
Opputunity	462	105
Risk	229	66
Cannot distinguish	109	29

Table 3: Statistics of Korean impact type annotations.

- The impact level may be adjusted according to a balance of positive and negative impacts. For example, a highly impactful/problematic event may be partially resolved.

Korean is the new language of ML-ESG, and impact-type labels are also provided at this time. Please refer to our previous paper (Chen et al., 2023b; Tseng et al., 2023) for more details.

2.2. Statistics

The Cohen's Kappa coefficient (Cohen, 1960) for datasets in Chinese, Korean, and Japanese yielded values of 0.21, 0.26, and 0.31, respectively. This variation underscores the challenges inherent in inferring the duration of the impact. To ensure the quality of the training and testing data, we exclusively utilized instances from the Chinese dataset that received uniform labels from the annotators. Table 1 details the statistics of the annotation results. The distribution of impact levels and types are presented in Tables 2 and 3, respectively. Table 1 and 2 demonstrate that low impact duration and length data are less abundant for the English and French languages.

2.3. Challenges

This edition faced a double challenge due to the previously mentioned nature of ESG news: unbalanced label distribution and annotation disagreements. For the first issue, the detailed guidelines guaranteeing a certain objectivity cannot ignore the fact that annotators having different backgrounds can still interpret the guidelines with a biased view, adjusting the impact level and duration accordingly during the annotation process. Thus, we accentuated our efforts on both cross- and group reviews to reach a high level of objectivity and coherence. For the latter, as most of ESG-related actions carry relatively long-term consequences with a medium to

	Best-Performing Method	Paper
Chinese	Longformer (Beltagy et al., 2020)	Tseng et al. (2023)
English	DeBERTa-v3 (He et al., 2023)	Dakle et al. (2024)
French	BERT (Devlin et al., 2019) & FinBERT (Araci, 2019)	Banerjee et al. (2024)
	GPT4 (OpenAI et al., 2024)	Tian and Chenn (2024)
Korean	XLM-RoBERTa (Conneau et al., 2020)	Dakle et al. (2024)
	KF-DeBERTa (jeo, 2023)	Kim et al. (2024)
Japanese	DeBERTaV3 (He et al., 2023)	Dakle et al. (2024)
	XLM-RoBERTa (Conneau et al., 2020)	Abburi et al. (2024)

Table 4: Best-performing methods.

English	Micro-F1	Macro-F1
Jetsons_3	64.71%	52.47%
Team Tredence_3	66.18%	50.12%
LIPI_1	58.82%	49.62%
fin-turbo_2	69.12%	46.89%
Jetsons_1	61.03%	46.70%
Team Tredence_2	58.09%	45.45%
upaya_3	60.29%	44.23%
Drocks_2	59.56%	44.14%
CriticalMinds_3	65.44%	43.86%
Drocks_1	58.82%	43.37%
Drocks_3	57.35%	43.01%
CriticalMinds_1	64.71%	42.81%
upaya_1	57.35%	42.75%
DICE_2	55.88%	42.53%
Jetsons_2	56.62%	42.28%
CompLx_1	56.62%	42.07%
SamNLP_2	57.35%	41.94%
MLG-TRDDCPune_1	52.21%	41.75%
MLG-TRDDCPune_3	52.21%	41.75%
MLG-TRDDCPune_2	52.21%	41.75%
CriticalMinds_2	59.56%	41.53%
LIPI_3	52.21%	40.73%
fin-turbo_3	61.76%	40.35%
fin-turbo_1	58.82%	39.83%
SamNLP_1	61.76%	39.59%
FinTwin_1	62.50%	38.90%
Team Tredence_1	61.03%	38.74%
upaya_2	51.47%	38.55%
DICE_3	55.15%	37.84%
DICE_1	44.85%	37.07%
kaka_1	52.94%	36.36%
LIPI_2	50.00%	32.70%

French	Micro-F1	Macro-F1
LIPI_1	56.16%	49.19%
upaya_2	52.05%	48.73%
French-_kaka_1	46.58%	47.42%
fin-turbo_1	56.16%	46.22%
Drocks_3	50.00%	45.77%
Team Tredence_2	53.42%	45.66%
fin-turbo_3	48.63%	44.88%
fin-turbo_2	57.53%	43.35%
Drocks_1	50.00%	43.31%
upaya_1	46.58%	42.86%
Drocks_2	49.32%	42.52%
Jetsons_3	54.11%	42.23%
Team Tredence_3	49.32%	40.54%
Team Tredence_1	41.78%	39.70%
Jetsons_1	47.95%	37.06%
SamNLP_2	43.84%	36.84%
LIPI_3	37.67%	36.41%
Jetsons_2	46.58%	34.62%
DICE_1	34.93%	34.45%
SamNLP_1	46.58%	33.70%
CriticalMinds_3	54.11%	32.88%
CriticalMinds_2	46.58%	32.19%
upaya_3	41.10%	32.09%
CriticalMinds_1	54.79%	30.33%
LIPI_2	41.10%	30.02%

Korean	Micro-F1	Macro-F1
Jetsons_1	70.00%	66.24%
3idiots_1	67.50%	61.98%
3idiots_3	67.50%	61.54%
3idiots_2	66.50%	61.02%
Team Tredence_2	64.00%	58.18%
Jetsons_3	64.00%	57.39%
Drocks_3	62.50%	55.17%
kaka_1	56.00%	52.94%
Team Tredence_3	57.50%	52.36%
Team Tredence_1	59.50%	51.58%
fin-turbo_3	65.50%	51.26%
Drocks_2	57.50%	48.39%
Drocks_1	60.50%	48.02%
fin-turbo_1	65.00%	47.32%
fin-turbo_2	64.50%	47.30%
FIT_2	61.50%	43.98%
FIT_1	52.50%	43.82%
Jetsons_2	42.00%	38.11%
FinNLP_1	49.00%	36.87%
FinNLP_2	49.50%	36.75%
FinNLP_3	49.00%	36.47%
LIPI_1	3.50%	4.38%

Japanese	Micro-F1	Macro-F1
Jetsons_2	36.50%	25.60%
Drocks_1	36.30%	25.60%
Jetsons_3	36.50%	25.50%
kaka_1	34.90%	25.50%
Team Tredence_1	43.10%	24.80%
Drocks_2	34.40%	24.50%
Albatross_1	31.90%	23.70%
Drocks_3	32.70%	23.40%
Team Tredence_2	39.90%	21.80%
Jetsons_1	30.90%	21.50%
LIPI_3	27.90%	19.20%
LIPI_1	29.90%	18.60%
Team Tredence_3	29.90%	18.00%
LIPI_2	24.30%	16.10%
ABC_1	18.90%	11.80%
IMNTPU_2	11.90%	7.10%
IMNTPU_1	11.10%	5.00%

Table 5: Performance — Impact Duration.

high impact on society and industries, a substantial analytical work was conducted to reveal which topics and impact type could entail a low impact level and/or low impact duration in order to obtain quality datasets. Overcoming these challenges evidences the necessity to assist human analysts.

3. Methods

A total of 12 teams share their methods in ML-ESG-3. We show the best-performing method in Table 4, and provide an overview of participants' methods in this section.

3.1. Impact Duration

In the Korean subtask of the ML-ESG-3 challenge, two teams, Jetsons (Dakle et al., 2024) and 3idiots (Kim et al., 2024), showcased strategies for improving ESG impact duration prediction accuracy amidst challenges like class imbalance and data scarcity. The Jetsons team led the field by implementing a data augmentation strategy that utilized self-training with supplementary English and French ESG articles to generate pseudo labels, thus enriching their training dataset. This approach, coupled with the fine-tuning of an XLM-RoBERTa model (Jetsons_1) (Conneau et al., 2020) , showcased the effectiveness of integrating sophisticated language models with data augmentation to improve multilingual ESG impact duration predictions. The 3idiots team distinguished themselves with a semi-supervised learning (SSL) approach, utilizing a finance-specialized pre-trained language model, KF-DeBERTa (jeo, 2023), along with advanced data augmentation techniques (Wei and Zou, 2019). By enriching their dataset with unlabeled ESG-related news articles, they achieved significant results, illustrating the potential of SSL and domain-specific models in enhancing NLP tasks with limited labeled data.

In the Japanese impact duration subtask, both Jetsons_2 (Dakle et al., 2024) and Drocks_1 (Abburi et al., 2024) achieved first place with the highest Macro F1 score. Dakle et al. (2024) implemented three strategies in the Japanese subtask: the English translation approach (Jetsons_2), the ensemble approach (Jetsons_3), and the fine-tuned multilingual model approach (Jetsons_1). For the English translation approach, Japanese texts were translated into English using the Google API, followed by a fine-tuning of the DeBERTa-v3-small model (He et al., 2023) on the class labels using the translated text. In the ensemble approach, they combined three models: XLM-RoBERTa, Longformer, and DeBERTa. The comparative results indicated that both the English translation and ensemble approaches outperformed the fine-tuned multilingual model approach, which was based on XLM-RoBERTa. Abburi et al. (2024) employed a data augmentation approach based on English text translated using the DeepL service, augmented with PEGASUS and GPT-mix, and then translated back into Japanese. They also trained an ensemble model that combined transformers (XLM-RoBERTa), CNN, and Voyage AI embeddings. It is noteworthy that a common characteristic of both teams was their reliance on English translation.

3.2. Impact Level

In the English impact duration and level subtasks, Jetsons_3 and Jetsons_1 (Dakle et al., 2024), respectively, proposed the best performing model with the highest Macro F1 score, while LIPI_1 (Banerjee et al., 2024) acheived the best score for the French impact duration task and kaka_1 (Tian and Chenn, 2024) for the French impact level task.

To handle multilingual datasets with relatively low volume and issues of label imbalance, most participants translated all datasets into English using tools like DeepL and Google Translate and explored data augmentation techniques using recent LLMs (e.g. GPT, Gemini, T5) to generate more samples. Those efforts on the dataset show improvements in some cases (Banerjee et al., 2024) (Dakle et al., 2024) but not in others (Atanassova et al., 2024). This observation indicates that processing ESG-related information seems to be language-dependent, so that it requires a strategy determining the relevance of data to each specific language (Dakle et al., 2024).

Most participants largely explored pre-trained transformer-based models, particularly, BERT, RoBERTa, DeBERTa and Longformer, by fine-tuning them on the ESG dataset. We observe that training various transformer models separately and subsequently combining them through an ensembling process has proven to yield the best results in impact duration and level classification (Yang and Rong, 2024) (Kao et al., 2024) (Bougiatiotis et al., 2024) (Dakle et al., 2024). An alternative approach involves fine-tuning Mistral-7B on a dataset generated by GPT-4, which contains articles along with information on the impact level, length, and rationale behind the classification (Rajpoot et al., 2024).

Another approach relies on classical machine learning classification algorithms such as Random Forest, XGBoost and KNN, which have shown less optimal performance in theses tasks due to challenges related to data imbalance (Shetty, 2024) (Atanassova et al., 2024).

3.3. Impact Type

Building upon their successful semi-supervised learning (SSL) approach for predicting ESG impact duration, the 3idiots team (Kim et al., 2024) applied a similar methodology to classify the impact type of ESG-related events on companies. Employing the same finance-specialized pre-trained language model, KF-DeBERTa (jeo, 2023), the team enriched their dataset with additional unlabeled ESG news articles, paralleling their strategy in the impact duration challenge. Through the use of advanced data augmentation techniques, including both weak (Wei and Zou, 2019) and strong augmentations, they effectively leveraged the model's capabilities to capture domain-specific nuances.

4. Performances

4.1. Impact Duration

Table 5 shows the performance of the official evaluation of participants' models.

In Korean Impact Duration, the application of advanced NLP models, notably KF-DeBERTa (jeo, 2023) and XLM-RoBERTa (Conneau et al., 2020), showcased exemplary performance among encoder models such as FinBERT (Araci, 2019), BERT (Devlin et al., 2019), and so on. Particularly, the integration of semi-supervised learning (SSL) (Tarvainen and Valpola, 2018) and diverse augmentation strategies (Wei and Zou, 2019; Lee et al., 2023) played a crucial role, enhancing model robustness and comprehension of ESG-related news articles, thereby leading to superior outcomes in classification tasks. Moreover, a noteworthy innovation was observed from a team (Yun Hyojeong and Son, 2024) employing GPT-4 (OpenAI et al., 2024), which diverged from traditional methodologies by leveraging prompting and dynamic in-context learning without direct model fine-tuning on the provided datasets. This approach highlighted how advanced generative language models can understand and tackle specialized area.

Banerjee et al. (2024)[1] proposed an English translation approach using Google Translate for the Japanese subtask and augmented the translated dataset with a T5-based model. They utilized the

[1]Their team ID is "LIPI."

	English		French		
	Micro-F1	Macro-F1		Micro-F1	Macro-F1
Jetsons_1	65.44%	60.90%	kaka_1	63.70%	63.29%
Team Tredence_3	58.09%	57.69%	upaya_1	58.22%	56.78%
LIPI_1	60.29%	56.57%	upaya_2	58.22%	56.69%
Jetsons_2	60.29%	56.51%	Team Tredence_3	54.79%	53.80%
Team Tredence_2	59.56%	56.16%	Team Tredence_2	50.00%	51.06%
DICE_2	55.88%	55.27%	Drocks_1	48.63%	48.81%
DICE_3	58.82%	55.08%	Drocks_2	48.63%	48.70%
Drocks_1	57.35%	55.03%	LIPI_2	48.63%	48.30%
IMNTPU_2	58.82%	55.03%	Team Tredence_1	47.95%	47.56%
DICE_1	55.15%	53.11%	IMNTPU_1	47.26%	47.16%
CompLx_1	60.29%	51.88%	DICE_1	49.32%	44.80%
LIPI_2	58.09%	51.48%	Drocks_3	43.15%	42.90%
LIPI_3	56.62%	51.42%	LIPI_3	41.78%	40.45%
kaka_1	51.47%	51.07%	SamNLP_2	43.15%	38.00%
Drocks_2	53.68%	48.65%	CriticalMinds_2	39.04%	37.96%
upaya_1	54.41%	48.40%	upaya_3	42.47%	37.64%
Team Tredence_1	50.00%	48.10%	SamNLP_1	42.47%	37.63%
MLG-TRDDCPune_3	52.21%	47.78%	IMNTPU_2	37.67%	34.46%
Drocks_3	52.21%	46.41%	LIPI_1	41.10%	26.89%
SamNLP_2	50.74%	46.30%	CriticalMinds_3	36.30%	26.21%
upaya_3	51.47%	46.09%	CriticalMinds_1	36.30%	22.48%
upaya_2	53.68%	45.93%			
SamNLP_1	52.21%	45.24%			
MLG-TRDDCPune_1	49.26%	44.74%			
MLG-TRDDCPune_2	50.00%	43.95%			
FinTwin_1	50.00%	43.55%			
CriticalMinds_1	47.06%	43.16%			
CriticalMinds_3	45.59%	40.64%			
CriticalMinds_2	42.65%	39.59%			
IMNTPU_3	19.12%	17.22%			
IMNTPU_1	18.38%	15.54%			

Table 6: Performance — Impact Level.

pretrained BERT-base multilingual uncased model for content concatenated with the impact type feature and classified it using a linear layer. Kao et al. (2024)[2] also employed the BERT-base multilingual-cased model for the Japanese subtask and augmented the dataset using GPT-3.5-turbo. Shetty (2024) explored the efficacy of various classifiers using the scikit-learn library and demonstrated that the decision tree approach was effective for the Japanese subtask. One reason for the comparative deficiency in performance against the top teams appeared to be their lack of use of state-of-the-art pretrained models such as DeBERTa-v3-xsmall or XLM-RoBERTa.

4.2. Impact Level

Table 6 shows the performance of participants' methods on the impact level task.

4.3. Impact Type

Table 7 shows the results of the impact type task in the Korean dataset.

[2]Their team ID is "IMNTPU."

	Micro-F1	Macro-F1
3idiots_3	84.00%	79.85%
FIT_2	81.50%	76.13%
Team Tredence_2	82.50%	75.95%
3idiots_2	81.50%	73.98%
3idiots_1	80.50%	73.43%
Team Tredence_1	80.00%	73.17%
Team Tredence_3	80.00%	71.76%
FIT_1	78.50%	64.46%
FinNLP_2	79.50%	62.46%
FinNLP_3	79.50%	62.46%
FinNLP_1	79.50%	62.46%
kaka_1	63.00%	55.53%
LIPI_1	64.00%	45.53%

Table 7: Performance — Impact Type

In the MLESG-3 shared task, the approaches to the Korean Impact Type mirrored those of the Korean Impact Duration, leveraging advanced NLP models such as KF-DeBERTa (jeo, 2023) with consistent effectiveness. This parallel strategy was reinforced by the adoption of semi-supervised learning (SSL) (Tarvainen and Valpola, 2018) or data augmentation, enhancing both tasks. Furthermore, the use of GPT-4 (OpenAI et al., 2024) by a team (Yun Hyojeong and Son, 2024) showcased in-context learning and prompting techniques, proving that specialized tasks like Impact Type classification can achieve significant outcomes without conventional fine-tuning.

5. Verifying Virtue — Promise Verification

In the ML-ESG shared tasks series, we focus on analyzing news articles from various countries to understand ESG-related events, thereby dynamically scoring a company's ESG performance based on third-party news. To advance our research, the upcoming shared tasks series will concentrate on the ESG-related promises made by companies. This series will encompass tasks such as (1) identifying ESG-related promises, (2) linking evidence to these promises, (3) determining the type of promise-evidence relationship, and (4) inferring the timing for verifying these promises. Our goal is to continue enhancing our multilingual and cross-country datasets.

For the forthcoming series, participants are encouraged to utilize ML-ESG datasets to improve their task performances. For instance, the dataset from ML-ESG-1 can aid in understanding the types of promises, which is crucial for the promise-evidence type task. Similarly, the ML-ESG-3 dataset can be instrumental in inferring the duration of events, a key factor in the task of verifying timing inference.

6. Conclusion

In the ML-ESG series of shared tasks, we have explored three tasks for dynamically scoring a company's ESG score based on news articles. ML-ESG-3, in particular, introduced the challenge of inferring the duration of impacts. Unlike ESG issue identification (ML-ESG-1) and impact type (ML-ESG-2), the impact duration (ML-ESG-3) is much more subjective, evidenced by low agreements in the annotation results across different languages. The performance in ML-ESG-1 and ML-ESG-2 is significantly better than in ML-ESG-3. Based on participants' findings, we observe that pre-trained LMs and LLMs perform well in well-defined tasks but still face challenges with this kind of subjective task. Thus, one of our suggestions is for ESG scoring companies to share more details about the assessment results of experts' discussions and experiences. This would help make the process more transparent and increase the possibility of models performing the task automatically.

Furthermore, we reveal our plan for the next shared task series, which focuses on multi-lingual ESG promise verification. This future direction promises to further refine our understanding of corporate ESG performance, enhancing transparency and accountability across languages and borders. We hope the ML-ESG task series will contribute to promoting sustainability and equity in the financial sector.

7. Acknowledgments

This work was supported by National Science and Technology Council, Taiwan, under grants MOST 110-2221-E-002-128-MY3, NSTC 112-2634-F-002-005 -, and Ministry of Education (MOE) in Taiwan, under grants NTU-112L900901. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). The work of Yohei Seki was partially supported by the Japanese Society for the Promotion of Science Grant-in-Aid for Scientific Research (B) (#23H03686), and Grant-in-Aid for Challenging Exploratory Research (#22K19822).

8. Bibliographical References

2023. *KF-DeBERTa: Financial Domain-specific Pre-trained Language Model*. Korean Institute of Information Scientists and Engineers.

Harika Abburi, Ajay Kumar, Edward Bowen, and Balaji Veeramani. 2024. Multilingual esg news impact identification using ensemble approach. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models.

Iana Atanassova, Marine Potier, Maya Mathie, Marc Bertin, and Panggih Kusuma Ningrum. 2024. Criticalminds: Enhancing ml models for esg impact analysis categorisation using linguistic resources and aspect-based sentiment analysis. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Neelabha Banerjee, Anubhav Sarkar, Swagata Chakraborty, Sohom Ghosh, and Sudip Naskar.

2024. Fine-tuning language models for predicting the impact of events associated to financial news articles. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Konstantinos Bougiatiotis, Andreas Sideras, Elias Zavitsanos, and Georgios Paliouras. 2024. Dice @ ml-esg-3: Esg impact level and duration inference using llms for augmentation and contrastive learning. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115, Macao. -.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023b. Multi-lingual esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Parag Pravin Dakle, Alolika Gon, Sihan Zha, Liang Wang, Sai Krishna Rallabandi, and Preethi Raghavan. 2024. Jetsons at finnlp 2024: Towards understanding the esg impact of a news article using transformer-based models. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Yu Han Kao, Vidhya Nataraj, Ting-Chi Wang, Yu-Jyun Zheng, Hsiao-Chuan Liu, Wen-Hsuan Liao, Chia-Tung Tsai, and Min-Yuh Day. 2024. Imntpu at ml-esg-3: Transformer language models for multi-lingual esg impact type and duration classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Jungdae Kim, Eunkwang Jeon, and Jeon Sang Hyun. 2024. Leveraging semi-supervised learning on a financial-specialized pre-trained language model for multilingual esg impact duration and type classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. Easyguide : Esg issue identification framework leveraging abilities of generative large language models.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Pawan Rajpoot, Ashvini Jindal, and Ankur Parikh. 2024. Adapting llm to multi-lingual esg impact and length prediction using in-context learning and fine-tuning with rationale. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Poorvi Shetty. 2024. Esg impact inference in english, french, korean and japanese. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Antti Tarvainen and Harri Valpola. 2018. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.

Ke Tian and Hua Chenn. 2024. Esg-gpt: Gpt4-based few-shot prompt learning for multi-lingual esg news text classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of The 32nd ACM International Conference on Information and Knowledge Management (CIKM'23)*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks.

Weijie Yang and Xinyun Rong. 2024. Duration dynamics: Fin-turbo's rapid route to esg impact insight. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

Moonjeong Hahm Kyuri Kim Yun Hyojeong, Chan-Yeong Kim and Guijin Son. 2024. Esg classification by implicit rule learning via gpt-4. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and The 4th Workshop on Economics and Natural Language Processing*.

IMNTPU at ML-ESG-3: Transformer Language Models for Multi-Lingual ESG Impact Type and Duration Classification

Yu-Han Kao[a], Vidhya Nataraj[b], Ting-Chi Wang[c], Yu-Chun Cheng[c], Hsiao-Chuan Liu[a], Wen-Hsuan Liao[a], Chia-Tung Tsai[a], Min-Yuh Day[a]*
Graduate Institute of Information Management, National Taipei University[a],
Smart Healthcare Management, National Taipei University[b],
Department of Accountancy, National Taipei University[c]
Taiwan
{s711236116, s411177056, s411177047, s711136108, s711136109, s711136106, myday}@gm.ntpu.edu.tw,
vidhyanataraj99@gmail.com

Abstract
Team IMNTPU participated in the multi-lingual Environmental, Social, and Governance (ESG) classification task, focusing on datasets in three languages: English, French, and Japanese. This study leverages Pre-trained Language Models (PLMs), with a particular emphasis on the Bidirectional Encoder Representations from Transformers (BERT) framework, to analyze sentence and document structures across these varied linguistic datasets. The team's experimentation with diverse PLM-based network designs facilitated a nuanced comparative analysis within this multi-lingual context. For each language-specific dataset, different BERT-based transformer models were trained and evaluated. Notably, in the experimental results, the RoBERTa-Base model emerged as the most effective in official evaluation, particularly in the English dataset, achieving a micro-F1 score of 58.82 %, thereby demonstrating superior performance in classifying ESG impact levels. The major contribution of this paper highlights the adaptability and effectiveness of PLMs in tackling the complexities of multi-lingual ESG classification tasks. The practitioner implications of this paper provide ESG analysts with more reliable tools for assessing the impact duration and level of sustainability initiatives.

Keywords: Multi-Lingual ESG, Data Augmentation, ESG impact analysis, Classification, Pre-trained Language Models (PLMs)

1. Introduction

In recent years, the global investment and corporate governance community has increasingly recognized the pivotal role of ESG factors as essential perspectives for driving a company's long-term growth and informing investment decision-making. Evaluating the sustainability and ethical impact of investment opportunities, ESG considerations have underscored the necessity for robust tools and methodologies to address related issues. Meanwhile, the escalating risk associated with non-financial factors highlights ESG elements as a primary threat to the stability of financial systems (Ziolo et al., 2019). As a response to this imperative, initiatives have emerged to tackle the challenges of automatically identifying and categorizing ESG-related themes in textual data.

Responding to the imperative of incorporating ESG considerations, initiatives leveraging Natural Language Processing (NLP) technologies have been developed to automate the identification and categorization of ESG-related themes in textual data and revolutionizing the approach within the financial services sector. NLP serves as a potent instrument for extracting profound semantic insights from vast pools of unstructured data, ranging from financial reports to chat transcripts and news articles. Through such analysis, NLP has the potential to bolster scenario recognition and risk assessment across various financial contexts. Given the prevalence of individual opinions on financial matters, conveyed through diverse channels such as news outlets and social media platforms, strategic analysis of these sentiments offers invaluable insights, shaping decision-making processes and influencing both user and organizational perspectives within the financial domain.

In the progression from ML-ESG-2 to ML-ESG-3, the domain of ESG analysis has seen the introduction of sophisticated tasks aimed at enhancing the precision of ESG rating systems. In ML-ESG-2, a novel challenge was introduced, focused on ESG impact type identification, requiring models to discern whether a piece of news represents an opportunity or risk from an ESG perspective. Advancing further, ML-ESG-3 expanded the scope to include the classification of news articles based on impact duration and impact level, utilizing a multilingual dataset to reflect the global nature of ESG considerations. In this context, our team, IMNTPU, has employed in ML-ESG-3 utilizing the PLMs to adeptly classify sentences that describe a company's ESG efforts, assigning them to distinct labels for both impact duration and impact level, thereby showcasing the evolving complexity and understanding required in contemporary ESG analysis. Building upon the foundation laid by the tasks of ML-ESG-3, and the employment of advanced pre-trained language models by team IMNTPU for precise classification, this methodology facilitates the extraction of textual evidence for ESG impact duration and impact level from the often-noisy environment of news article reports. Consequently, this approach supports more informed investment decisions by leveraging the refined insights gained from the automated analysis

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 228–233
20 May, 2024.

of ESG-related textual data.The contributions of this work can be summarized as follows:

- **Implementing Data Augmentation**: To combat class imbalance within the datasets, enhancing model robustness and ensuring a balanced representation for more accurate ESG impact duration and level classification.
- **Training with BERT-based Transformer Models**: Leveraging the sophisticated capabilities of BERT-based models across multilingual datasets to significantly improve the precision and comprehensiveness of ESG impact duration and level classifications.

Our research revealed a good correspondence in classifying the ESG impact duration and level in textual evidence. This finding will be helpful in future work on automatic estimation of ESG scores from textual resources.

The remaining part of the paper proceeds as follows: The second chapter introduces the related work related to the ML-ESG-3 shared task. Chapter three presents our approaches for each of the datasets. Chapter four provides a comprehensive account of the official experiment results and includes a detailed analysis. Finally, chapter five outlines the conclusions obtained from this study.

2. Related Work

In light of the heightened attention toward ESG issues, machine learning (ML) and NLP techniques have increasingly been leveraged in recent years to conduct sophisticated analyses of ESG ratings and predict impacts. By harnessing the predictive capabilities of Artificial Intelligence (AI) models have been created not only to assess current ESG ratings, classify them into various categories, but also to forecast future trajectories pertaining to both financial and societal impacts.(Tseng et al., 2023; Wang et al., 2023)

2.1 ESG in NLP

Lee et al. (2022)highlight the growing trend of companies disclosing their sustainability practices through various forms of unstructured text, such as reports and transcripts. They point out that NLP plays a crucial role in automating the classification and measurement of ESG-related news articles, enabling the parsing of extensive datasets to identify pertinent information efficiently. Furthermore, Zhuang et al. (2020) underscore the significance of transfer learning techniques within NLP, utilizing large language models to facilitate the transfer of knowledge across different sustainability domains and languages. These advancements underscore the pivotal role of NLP in enhancing the accessibility and analysis of sustainability information, contributing significantly to the field of ESG research.

2.2 Previous approach in multi-lingual ESG issues classification

In the realm of multi-lingual classification, the identification ESG issues across varied disclosure mediums presents a complex challenge. Recent efforts have explored numerous solutions, predominantly harnessing advanced NLP techniques to navigate this multifaceted landscape. A significant milestone in this ongoing journey was the 5th Workshop on Financial Technology and NLP (Kannan & Seki, 2023)which organized a shared task dedicated to ESG issue detection, attracting participation from 26 teams. Within this competitive context, a diversity of innovative approaches emerged, targeting a dataset encompassing 44 distinct ESG issues.

Armbhurst, Schäfer, and Klinger (2020) analyzed the impact of a company's environmental performance, derived from MD&A sections in financial filings, on its financial outcomes. They concluded that, while the MD&A text does not predict financial performance, environmental performance can be effectively identified using NLP techniques.

Wang et al. (2023) introduced the application of the MacBERT model(Cui et al., 2020), enhancing its capabilities with additional pre-training and contrastive learning strategies for the meticulous examination of ESG issues within the Chinese language track. In a similar vein, Pontes et al. (2023) employed a combination of models, including a Support Vector Machine (SVM) model (Platt, 1999) integrated with Sentence BERT (SBERT) embeddings (Reimers & Gurevych, 2019) and RoBERTa-based models (Liu et al., 2019), to classify multi-lingual ESG issues. Glenn et al. (2023) and Devlin et al. (2018) leveraged the potential of open-source large language models(LLM), notably gpt, for data augmentation purposes, thereby enhancing the performance of model. Mehra et al. (2022) made a notable contribution by developing ESGBERT, a tool specifically fine-tuned on a BERT model for sequence classification and conducting a Masked Language Model (MLM) task on an ESG-focused corpus, showcasing ESGBERT's efficacy in capturing the nuanced context of ESG for specialized text classification tasks.

Drawing inspiration from these pioneering contributions, our study leverages PLMs to classify the impact duration and level of ESG issues within news articles. By integrating the insights and methodologies from these notable works, our approach seeks to further refine the accuracy and applicability of NLP technologies in dissecting and understanding the complex domain of ESG disclosures, illustrating the interconnected progress within the field.

3. Proposed Methods

3.1 Dataset

Figure 1 shows the architecture used in this study.In the multi-lingual ESG-3 shared task, the organizers provided datasets in five languages, which were divided into different subtasks as outlined in Appendix 1. The training datasets included English,

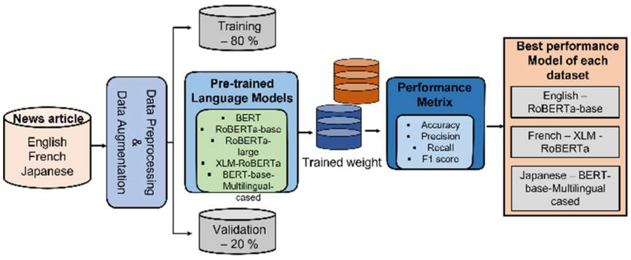


Figure 1: System architecture of Multi-Lingual ESG Impact type and Impact duration classification

French, Japanese, Korean, and Chinese. English, Korean and French datasets were associated with two subtasks, impact level and impact length, while Japanese, and Chinese had only one subtask, impact length. Team IMNTPU participated in three languages: English, French, and Japanese which consists of 545 news articles in the English dataset, 664 in the French dataset, and 50 news articles in the Japanese dataset, related to ESG issues.The English and French datasets include the following columns: "URL," "news_title," "news_content," "impact_level," and "impact_length." In contrast, the Japanese dataset contains "ID," "Text," "Relevancy," "ESG_type," "impact_type," and "impact_duration."

3.2 Data Augmentation

We observe that the dataset presents two primary challenges: a constrained overall size and uneven label distribution across different languages. To tackle these issues, we employed gpt-3.5-turbo in view of cost effective, an open-source large language model, for data augmentation purposes. This strategy not only expanded our dataset but also aimed at rectifying the imbalance in label distribution. Data augmentation, in this context, is crucial for enhancing the diversity and representativeness of our dataset, thereby improving model training outcomes and ensuring a more robust and accurate classification performance across the multilingual ESG classification task. Appendix 2 illustrates the prompt used to generate additional text, showcasing our methodology for augmenting the dataset effectively.

Figure 2 presents the dataset before and after augmentation. The English dataset has expanded from 545 to 11,556 news articles, the French dataset from 661 to 10,104 articles, and the Japanese dataset from 50 to 1,430 articles. Additionally, the label distribution for impact level and impact length is more balanced compared to the original dataset.

3.3 Pretrained language model

The surge in leveraging PLMs such as BERT and its transformer-based counterparts has marked a significant stride in the field of NLP, extending its impact to domain-specific applications. This growing fascination with large-scale language models, underscored by their remarkable efficacy across diverse NLP applications, is well-documented in recent scholarly discourse (Liu et al., 2023). In the ambit of our current endeavor, we strategically deployed an array of PLMs — namely, BERT, RoBERTa (both Base and Large variants), XLM-RoBERTa, and BERT-base-multilingual-cased.

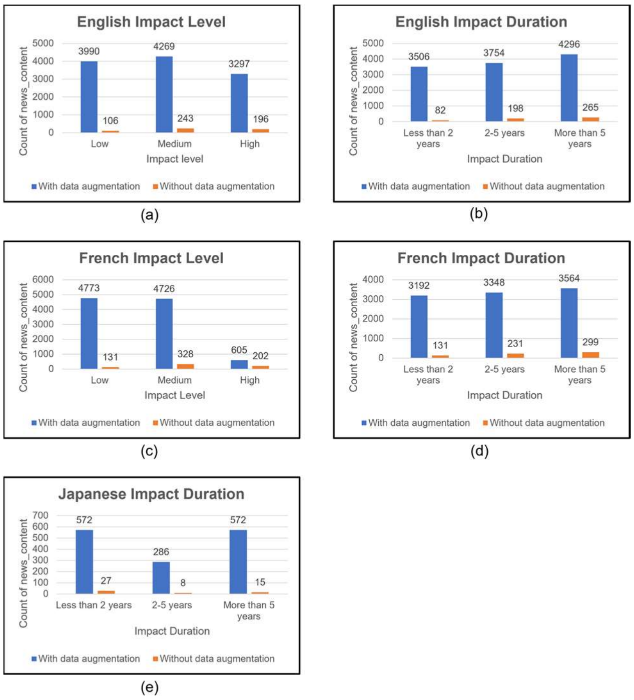


Figure 2: Comparative Analysis of Multilingual dataset before and after data augmentation (a) English Impact Level, (b) English Impact Duration, (c) French Impact Level, (d) French Impact Duration and (e) Japanese Impact Duration.

While these models share the foundational BERT architecture, they diverge in their pre-training approaches and the scale of parameters, which are pivotal in learning comprehensive language representations.

Our project aims to harness these models for the nuanced task of classifying the impact level and duration from the textual content of news articles. To this end, we utilize the transformative capabilities of Hugging Face's transformer models, meticulously chosen for their proficiency in comprehending and analyzing text. Given the multilingual nature of our dataset, we allocated the BERT and RoBERTa (Base and Large) for the English dataset, and the XLM-RoBERTa and BERT-base-multilingual-cased models for French and Japanese datasets, aligning with their inherent language processing strengths. Our methodology concentrates on parsing the news content, excluding titles, to derive predictions for designated labels.

To evaluate the models' performance, we first divided each original training dataset into an 80% slice for training purposes and a 20% segment for validation and we compared the models performance with data augmentation dataset. This structured approach not only amplifies the precision of our classification task but also underscores the adaptability of these PLMs in dissecting and understanding multilingual news narratives, setting a precedent for future research in domain-specific NLP applications.The hyperparameters of each model is mentioned in Appendix 3.

4. Experimental Results

4.1 Submitted runs

In our comprehensive experimental setup, we evaluated the efficacy of five distinct models across English, French, and Japanese datasets, with our findings meticulously documented in Table 1. This table encapsulates the culmination of our official submissions and their corresponding performance metrics. Within the English dataset evaluations spanning three submission rounds, the RoBERTa model stood out, securing the premier position with an impressive Micro-F1 score of 58.82% and a Macro-F1 score of 55.03%. This achievement underscores RoBERTa's nuanced understanding and processing capabilities of the English language.

Transitioning to the French dataset, our exploration across two submission rounds revealed the XLM-RoBERTa model as the frontrunner, achieving a notable Micro-F1 score of 47.26% and a Macro-F1 score closely aligned at 47.16%. This result highlights XLM-RoBERTa's adeptness at navigating the linguistic intricacies of the French language, cementing its status as a potent tool for multilingual analysis.

Further delving into the Japanese dataset, again over two rounds of submissions, the Bert-base-multilingual-cased model emerged as the victor, albeit with a Micro-F1 score of 11.90% and a Macro-F1 score of 7.10%. Despite the lower scores relative to the other languages, this outcome signals the model's capacity to grapple with the Japanese language, albeit indicating potential areas for improvement and refinement.

Table 1 not only serves as a testament to the comparative strengths and areas for enhancement across the models but also illuminates the path forward for optimizing multilingual ESG classification tasks. The distinguished performance of the RoBERTa model in English, in particular, delineates a benchmark for excellence, suggesting a fertile ground for future investigations to build upon and extend its application across diverse linguistic landscapes. Appendix 4 and Appendix 5 shows the comparison report of performance metrics report obtained before and after data augmentation in the development dataset .

5. Conclusion

In conclusion, team IMNTPU engaged in the multi-lingual ESG classification task, with the aim of discerning impact levels and durations from ESG-related news articles across English, French, and Japanese datasets. Leveraging transformer models, notably the RoBERTa-base model, we focused on optimizing our approach to accurately classify the given information. The RoBERTa-base model, in particular, demonstrated superior performance in the English dataset, achieving a commendable Micro-F1 score of 58.82%, which stands as our best result. This was followed by the French dataset with a score of 47.26%, and the Japanese dataset at 11.90%, highlighting a significant opportunity for improvement in handling Japanese language data, potentially through parameter adjustments such as learning rate and epochs.

Dataset	Subtask	Runs	Team ID	Model	Micro-F1	Macro-F1
English	Impact Level	Run 1	English_IMNTPU_1	BERT	18.38%	15.54%
		Run 2	English_IMNTPU_2	RoBERTa-base	**58.82%**	**55.03%**
		Run 3	English_IMNTPU_3	RoBERTa-large	19.12%	17.22%
French	Impact Level	Run 1	French_IMNTPU_1	XLM-RoBERTa	**47.26%**	**47.16%**
		Run2	French_IMNTPU_2	BERT-base-multilingual cased	37.67%	34.46%
Japanese	Impact Length	Run1	Japanese_IMNTPU_1	XLM-RoBERTa	11.10%	5.00%
		Run 2	Japanese_IMNTPU_2	BERT-base-multilingual cased	**11.90%**	**7.10%**

Table 1: Official evaluation results submitted to ML-ESG 3

Additionally, our application of data augmentation techniques played a critical role in enhancing our model's performance, particularly by addressing issues of data scarcity and label imbalance. However the data augmentation has not improved the performance in all models but it improved the performance in RoBERTa-base model which outperformed in official run. The results underscore the effectiveness of the RoBERTa-base model and data augmentation in advancing our understanding and classification capabilities within the multi-lingual ESG domain.

5.1 Research Contributions

This study advances the field of ESG impact assessment by training with BERT-based Transformer Models across multilingual datasets to significantly improve the precision and comprehensiveness of ESG impact duration and level level. Furthermore, implementing the Data Augmentation to balance the class within the datasets, enhanced the model robustness and ensuring a balanced representation for more accurate ESG impact duration and level classification.

5.2 Managerial Implications

These advancements offer substantial benefits. First the enhanced models provide ESG analysts with more reliable tools for assessing the impact duration and level of sustainability initiatives, thus supporting more informed and strategic decision-making. And organizations can better align their operations with sustainable practices, accurately track their ESG performance, catering to a globally diverse audience.

6. Appendices

Dataset	Subtasks	Labels
English	Impact Level	low, medium, high
	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
French	Impact Level	low, medium, high
	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
Japanese	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
Korean	Impact Level	Opportunity,risk,cannot distinguish
	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years
Chinese	Impact Length	Less than 2 years, 2 to 5 years, More than 5 years

Appendix 1: Different classification subtasks for each language.

Prompt: ** "Using the provided examples as a reference, generate additional examples of 'news_title' and 'news_content' focused on impactful Human Capital, Environmental, or Governance initiatives".
** "Each example should represent realistic and relevant ESG activities that align with MSCI ESG standards. Highlight innovative efforts and solutions addressing pressing ESG issues, ensuring the content is insightful and adheres to current industry trends and guidelines".
** "Maintain the distinctiveness of each new entry while covering a variety of ESG themes to reflect the depth and breadth of content expected".
'''Example 1'''
" 'news_title': 'Sustainability Advisory ERM Acquires Consulting Firm Sustainalize',"
" 'news_content': 'Keryn James, CEO, ERM, said: "The Covid-19 pandemic and the global movement on racial justice have accelerated and enhanced the focus on ESG risks and opportunities and the need for businesses to be more resilient and sustainable. Boards and executive teams need to become more knowledgeable, proactive and effective on ESG matters, from improving diversity within their companies, to linking executive pay to ESG metrics. We are delighted to announce the acquisition of Sustainalize which further strengthens our capabilities and capacity in being able to support clients in Europe and beyond as they navigate these increasingly complex issues."',"
'''Example 2'''
" 'news_title': 'Guest Post: ESG isn't About Altruism – it's About Survival',"
" 'news_content': 'Companies have always been accountable to their stakeholders. Shareholder value has been at the center of accountability for decades now. But the long term success of every company has always also been dependent on the ability to recruit and retain talent, to build brand identification, to maintain the social license to operate and to build resilient relationships with customers and communities. Investors, talent, customers and communities today want to work for, buy from, invest in and associate with companies that align with their personal values. Today, particularly among people in their 30s and younger, the quality of a product or of a company includes environmental sustainability and economic justice. An effective approach to Environment, Social and Governance (ESG) issues is a core part of the value chain.', ".....

Other 4 samples GPT to know

Appendix 2: Prompt generated using GPT 3.5 Turbo

Model	Batch size	Epoch	Optimizer	Learning rate
BERT	16	10	Adam	5e-5
RoBERTa-base	16	10	Adam	5e-5
RoBERTa-Large	16	10	Adam	5e-5
XLM-RoBERTa	16	3	Adam	5e-5
BERT-base-multilingual-cased	16	3	Adam	5e-5

Appendix 3: The main hyperparameters used in this study

Dataset	Models	Accuracy	F1 score	Precision	Recall
English	BERT	0.82	0.82	0.82	0.82
	RoBERTa-base	0.79	0.79	0.79	0.79
	RoBERTa-large	0.34	0.17	0.12	0.34
French	XLM-RoBERTa	0.87	0.86	0.86	0.87
	BERT base-multilingual-cased	0.88	0.87	0.88	0.88
Japanese	XLM-RoBERTa	0.40	0.23	0.16	0.40
	BERT-base-multilingual-cased	0.36	0.19	0.13	0.36

Appendix 4: Performance matrix of development dataset after data augmentation

Dataset	Models	Accuracy	F1 score	Precision	Recall
English	BERT	0.64	0.61	0.61	0.64
	RoBERTa-base	0.53	0.50	0.50	0.53
	RoBERTa-large	0.49	0.22	0.16	0.33
French	XLM-RoBERTa	0.42	0.25	0.17	0.42
	BERT base-multilingual-cased	0.45	0.31	0.44	0.42
Japanese	XLM-RoBERTa	0.60	0.45	0.36	0.60
	BERT-base-multilingual-cased	0.40	0.34	0.30	0.40

Appendix 5: Performance matrix of development dataset without data augmentation

7. Acknowledgments

This research was supported in part by the National Science and Technology Council (NSTC), Taiwan, under grants NSTC 112-2425-H-305-002-, and NSTC 112-2627-M-038-001-, and National Taipei University (NTPU), Taiwan under grants 113-NTPU_ORDA-F-003, 113- NTPU-ORDA-F-004, USTP-NTPU-TMU-113-03, and NTPU-112A413E01.

8. Bibliographical References

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2020). Revisiting pre-trained models for Chinese natural language processing. arXiv preprint arXiv:2004.13922.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Glenn, P., Gon, A., Kohli, N., Zha, S., Dakle, P. P., & Raghavan, P. (2023). Jetsons at the finnlp-2023: Using synthetic data and transfer learning for multilingual esg issue classification. Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting,

Kannan, N., & Seki, Y. (2023). Textual Evidence Extraction for ESG Scores. Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting,

Lee, O., Joo, H., Choi, H., & Cheon, M. (2022). Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms. Sustainability, 14(14), 8745.

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., & Liu, Z. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. Meta-Radiology, 100017.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Mehra, S., Louka, R., & Zhang, Y. (2022). ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. arXiv preprint arXiv:2203.16788.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers, 10(3), 61-74.

Pontes, E. L., Benjannet, M., & Ming, L. K. (2023). Leveraging bert language models for multi-lingual esg issue identification. arXiv preprint arXiv:2309.02189.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084.

Tseng, Y.-M., Chen, C.-C., Huang, H.-H., & Chen, H.-H. (2023). DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles. Proceedings of the 32nd ACM International Conference on Information and Knowledge Management,

Wang, W., Wei, W., Song, Q., & Wang, Y. (2023). Leveraging contrastive learning with bert for esg issue identification. Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting,

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. Proceedings of the IEEE, 109(1), 43-76.

DICE @ ML-ESG-3: ESG Impact Level and Duration Inference Using LLMs for Augmentation and Contrastive Learning

Konstantinos Bougiatiotis, Andreas Sideras, Elias Zavitsanos, Georgios Paliouras
Institute of Informatics and Telecommunications, NCSR "Demokritos"
Patriarhou Gregoriou and Neapoleos St., Aghia Paraskevi
{bogas.ko, andreasideras, izavits, paliourg}@iit.demokritos.gr

Abstract

We present the submission of team DICE for ML-ESG-3, the 3rd Shared Task on Multilingual ESG impact duration inference in the context of the joint FinNLP-KDF workshop series. The task provides news articles and seeks to determine the impact and duration of an event in the news article may have on a company. We experiment with various baselines and discuss the results of our best-performing submissions based on contrastive pre-training and a stacked model based on the bag-of-words assumption and sentence embeddings. We also explore the label correlations among events from the same news article and the correlations between impact level and impact length. Our analysis shows that even simple classifiers trained in this task can achieve comparable performance with more complex models under certain conditions.

Keywords: ESG, NLP, machine learning, impact, sustainability, duration, fintech

1. Introduction

Environment, Social, and Governance (ESG) in the financial industry includes environmental, social, and governance issues within a company that may impact its performance. Their effect may be mild, moderate, or severe, and their duration may vary. Each of the three aspects of ESG involves various indicators that contribute to the ESG profile of a company. The environmental element focuses primarily on climate considerations, waste management, and resource preservation. The social direction concerns human rights, employee health and safety, training, and consumer rights protection. The governance dimension is related to board issues, business ethics, and issues related to the company's strategic decisions.

ESG has recently become particularly important, forcing organizations to incorporate ESG criteria into their processes and operations. Assembling a company's ESG profile is critical because of the need to evaluate companies' activities and investments, as well as the adoption of regulations and the transparency of communication about their sustainability. Therefore, it is apparent from the business perspective that ESG issues may impact the company and its investors when there is doubt about its decision-making strategies and sustainability. Given the above, companies must periodically release ESG reports, as they represent an essential guide for potential new investors.

In this context, automating the analysis of ESG reports, indicators, or related news has gained much attention in the academic literature. Recently, an ESG shared task was proposed (Kang and El Maarouf, 2022) in the context of the FinNLP workshop series, including two subtasks that focused on ESG taxonomy enrichment and sustainable sentence prediction. The following year, the task was extended to a multilingual ESG issue identification (Chen et al., 2023) that aimed at integrating the ESG paradigm into financial natural language processing (NLP) systems. The objective of the task was to classify news articles into 35 key ESG issues and identify the affected company and the corresponding industry.

This third task on multilingual ESG inference (ML-ESG-3) aims to determine the impact and duration an event in the news article may have on a target company. This challenging task comprises two subtasks: impact level identification and duration identification, including news articles in five languages. In this work, we present the submission of the team *DICE* for ML-ESG-3, along with the baseline models we experimented with. Our primary focus was on the English language. In this setting, our best system ranked in the 6th position out of 32 submissions in the subtask of impact level identification, while our best-performing system in the subtask of impact duration ranked in the 14th position.

The rest of the paper is organized as follows. Section 2 provides an overview of the related work in the ESG domain. Section 3 presents the datasets given by the organizers and the task design. In sections 4 and 5, we discuss our methods and empirical results, while section 6 concludes the paper and highlights future directions.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 234–243
20 May, 2024.

2. Related Work

The ESG paradigm has gained increasing attention, especially since 2020. The idea of analyzing ESG data and factors has matured over time, and nowadays, the academic community supports the automated analysis of such data using machine learning (ML) and deep learning (DL) methods that target various aspects and use cases.

A body of work focuses on predicting ESG scores and the related variables and factors that affect these scores. The work in (Gupta et al., 2021) is based on statistical analysis and traditional ML to measure the importance of ESG parameters in financial performance and how they affect investment decisions. Similarly, in (D'Amato et al., 2021) and (D'Amato et al., 2022), the authors aim to identify the variables that affect the ESG score by leveraging random forests, and they conclude that balance sheet items, i.e., numerical indices, constitute significant predictors of the ESG score.

In addition, some work focuses on the impact of ESG data on investments and stock returns. The work in (Utkarsh Sharma and Gupta, 2024) investigates whether ESG data can lead to profitable investments. According to this, the higher the ESG scores, the better the financial performance, especially when ESG data are combined with other financial variables. In another study (Yu et al., 2022), the authors tried to discover the relationship between ESG scores and stock returns using credit rating agency data. Finally, the work in (Margot et al., 2021) uses ML to identify patterns between ESG profiles and the financial performance of companies by mapping ESG data to excess returns.

A common characteristic of the above efforts is that they rely on structured data analysis. However, ESG data are available at several levels and modalities. This variety raises interesting questions from an ESG perspective regarding the implications of differences in ESG data from different providers. For this reason, much work focuses on becoming independent of data providers by using other data sources, such as Corporate Social Responsibility (CSR) reports, company communications, and the news. For example, the work in (Wang et al., 2020) uses the news to classify the relevance and sentiment of the articles to the economy by using DL and traditional ML methods. In (Nugent et al., 2021), the authors analyzed news articles and classified them into twenty ESG categories using domain adaptation and data augmentation techniques to improve classification performance. Using transformer-based language models, the work in (Guo et al., 2020) used news data to examine the impact of ESG issues in financial news and to analyze the predictive power of ESG news on stock volatility.

In the previous multilingual ESG shared task (ML-ESG-2) (Chen et al., 2023) for news classification into ESG issues, most submitted methods focused on large language models. The authors in (Pontes et al., 2023) used RoBERTa and SBERT and found that the best results in both monolingual and multilingual data are achieved with RoBERTa, while the work in (Glenn et al., 2023) relies on fine-tuning multilingual BERT with augmented data produced by GPT-3.5. Similarly, the authors in (Lee et al., 2023) use generative models, zero-shot techniques, and translation to augment the training data and experiment with BERT-based models, such as RoBERTa and FinBERT. The work in (Mashkin and Chersoni, 2023) experiments with transformer representations that were used in traditional ML methods, such as Logistic Regression (LR), Random Forests (RF), and Support Vector Machines (SVM) for classification. Finally, the authors in (Billert and Conrad, 2023) and (Wang et al., 2023) also rely on BERT models. The former exploits a strategy for efficient transfer learning, introduced in (Houlsby et al., 2019), to fine-tune a multilingual BERT, while the latter leverages MacBERT in a contrastive learning framework utilizing pseudo-labeled data.

In this ML-ESG-3 shared task, we experiment with several baselines and focus on our submitted systems based on contrastively pre-trained and stacked models.

3. Datasets and Task Design

The organizers released the datasets in two phases. First, the annotated training data, including five languages, were released, and then, the blind test sets for the corresponding five languages. A training sample from the English dataset with the corresponding fields and values is shown below.


```
{
    "URL": "https://www.esgtoday.com/arabesque-ai-appoints-carolina-minio
          -paluello-as-new-ceo/",

    "news_title":"Arabesque AI Appoints Carolina Minio Paluello as New
          CEO",

    "news_content": "ESG-focused financial technology company Arabesque
          AI announced today the appointment of Dr. Carolina Minio
          Paluello as the company\u2019s new Chief Executive Officer.",

    "impact_level": "low",

    "impact_length": "2 to 5 years"
},
```


As depicted, apart from the news content, we also have the corresponding news title and URL from which the text was extracted. In this work, we focused on the English data, and we submitted systems for the English and French datasets where each text sample is annotated with the following labels:

- **Impact Length**, was selected among "Less than 2 years" ($x < 2$), "2 to 5 years" ($2 < x > 5$), and "More than 5 years" ($x > 5$).

- **Impact Level**, qualifies the opportunity or risk as being “low”, “medium” or “high”.

The English dataset consists of 545/136 train/test samples, while the French dataset is split into 661/146 respectively. The number of samples in each class for the English data is, as shown in Fig. 1, in paired format. The class distribution is not balanced. For the *Impact Length*, 48.62% of the data are annotated as “More than 5 years”, 36.33% of the data are annotated as “2 to 5 years”, and 15.05% of the data concern “Less than 2 years”. On the other hand, the impact level annotations are distributed as follows: 44.59% of the samples belong to the “medium” category, a percentage of 35.96% belongs to “high”, and the remaining 19.45% belongs to the “low” category. An important observation is that the “high” impact level category seems strongly correlated with a duration of “More than 5 years”.

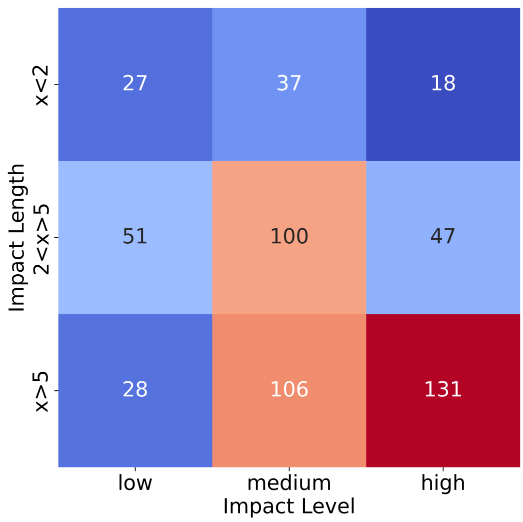


Figure 1: Number of samples in each class for both tasks in the English dataset.

4. Methods and System Selection

The current task entails several intricacies. As previously emphasized, there is a discrete correlation not only between the classes of impact length and level but also between the text snippets originating from the same article. Such instances occur in both the training and test data. Also, we operate within a low-resource environment with limited data. Thus, we experiment with methods that encapsulate the above observations. All our experiments were performed five times, using different splitting seeds on the full English training set, splitting the data in $70\%/10\%/20\%$ train/val/test stratified (concerning class label) splits in each run. The evaluation is performed in terms of macro-averaged F1, also reporting the standard deviations.

4.1. Features and Task Engineering

First, we experimented independently for the length and level identification tasks with ML methods, such as Logistic Regression and input representations like TF-IDF, to establish baseline performance and gain insights regarding the feature importance and problem difficulty. This analysis indicated that the model highly correlates specific people and company names with its prediction. By exploring the dataset, we validated that there are companies (e.g., Microsoft) that are almost always classified into the same classes for both prediction tasks. Also, given that multiple texts belong to the same article, we noticed that their labels match rather frequently. Consequently, we experimented with several pre- and post-processing techniques, as well as different ways to split the data for model selection.

Using a simple TF-IDF vectorization process, we noticed that specific words highly correlate with specific classes. Table 1 provides such examples and shows the number of occurrences of each word, alongside its distribution over the classes. The first set of words, namely “2035”, “2050”, and “trillion”, correspond to simple cases where it is straightforward to deduce the label of the texts containing them, solely using these context words. For instance, it is easy to understand that when talking about things that have a horizon up to 2050, the time context is probably “More than 5 years” ($x > 5$), or when talking about matters in the context of trillions of dollars, the impact level is probably “high”.

Table 1: Example of specific word occurrences and their distribution among both task labels.

Word (Occur.)	**Distribution**
2035 (5)	{x>5:4}, {high:3}
2050 (7)	{x>5:7}, {high:7}
trillion (9)	{x>5:8}, {high:9}
water (38)	{x>5:32}, {high or medium:34}
appoint (30)	{2<x>5:29}, {low:29}
hydrogen (11)	{x>5:7}, {high:10}
microsoft (6)	{x>5:5}, {high:6}
verizon (6)	{x>5:4}, {high:6}
hsbc (6)	{x<2 or 2<x>5:6}

The second group of words, which contains words like “water” and “appoint”, captures ESG-related issues. As expected, the “water-themed” news is mainly of “high” or “medium” impact and always corresponds to $x > 5$ years in terms of impact length, showcasing the long-term gravitas of water management. On the other hand, the word "appoint" refers to changes in personnel, mainly on the board of directors, and corresponds to "low" impact levels in terms of ESG.

The final group of words focuses on specific companies, for which all related news usually corresponds to "high" and long-term (i.e., $x > 5$ years) impact. One intuitive explanation for these "company-related" news exhibiting the same class could be the size of the companies, as any news related to companies of large capitalization may have severe implications in terms of ESG risks and opportunities. However, another explanation could be that many text samples that refer to specific companies originate from the same URL, hosting a specific news item, and the impact level/length class label is common among the samples in the same news article.

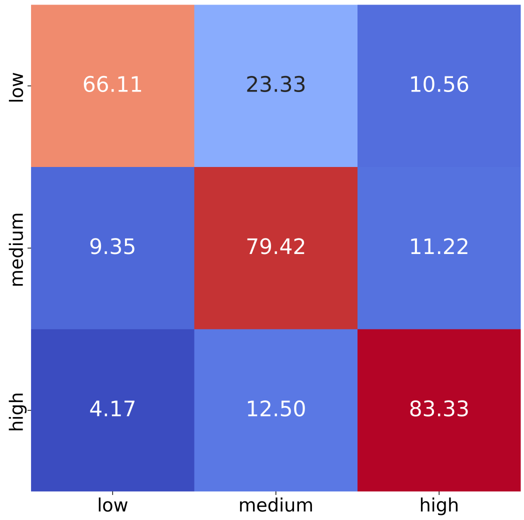


Figure 2: Correlation of class labels among same-URL instances regarding impact level.

Motivated by the above, we measured the correlation between class labels among same-URL samples, as shown in Fig. 2. Each cell $C[i, j]$ indicates the probability of encountering a sample with class label j, related to a specific URL, given that we have already seen a sample with class label i from the same URL (intra-URL class correlation). For example, for a news article containing two distinct text samples and given that one of them has a "high" impact level label, the prior probabilities of the labels for the second text are indicated by the last row of the table, i.e., $\{\texttt{low} : 4.17, \texttt{medium} : 12.50, \texttt{high} : 83.33\}$. This means that with a very high probability, the second text sample will have a "high" impact in most cases, regardless of its content. Finally, the diagonal of the correlation matrix has the highest values, validating our intuition that, in most cases, the news items found in a specific article exhibit the same impact level label.

To empirically validate our intuition, we devised a small-scale experiment, starting with a baseline classification model with a Bag-of-Words (BoW), TF-IDF weighted feature representation for each text, and a Logistic Regression (LR) model on top. We create two variants of this model. The first one uses a Named Entity Recognition (NER) component (we use spaCy (Honnibal and Montani, 2017)) and masks each named-entity identified in the text with a corresponding label string for the entity (e.g. "*Jeff Bezos*" is mapped to "*PERSON*", "*Microsoft*" to "*ORG*" etc.) to anonymize the text and mitigate any information that is bound to specific entities. The second deploys a simple post-processing strategy (dubbed *PostProcess*) using the prior-probability table of Fig. 1. Specifically, at inference time, if the sample for which we predict the labels originates from a URL that was already seen in training, we weigh the predicted class probabilities of the LR model with the corresponding prior probability for this specific URL based on the class labels of the other same-URL texts seen in training. This is a simple way to "steer" the predictions of the classifier toward the "expected" distribution of same-URL texts.

Table 2: Results using a baseline model and its variants on impact level prediction under different stratification splits.

Model	**Impact Level**	
	Class	*Class + URL*
BoW-LR	52.51 ± 3.05	47.68 ± 3.55
+*NER*	52.71 ± 3.91	**48.43** ± **2.30**
+*PostProcess*	**56.75** ± 4.06	47.68 ± 3.55

The performance of these models is reported in Table 2, in terms of macro-F1 averaged across five different runs. We also test their performance under two different stratification methods. The first one, corresponding to the second column in the Table, denotes the vanilla stratification setup based on the class labels. The second one, corresponding to the third column in the Table, is a stratified group split where the samples also follow a group split based on their URLs. In this setup, samples belonging to the same URL are always found in the same split, either train, validation, or test, so there is no intra-URL "leak" among the splits.

Focusing on the vanilla setup first, we observe that adding the NER pre-processing step does not improve the generalization capabilities of the model much. On the other hand, the post-process strategy improved the performance of this simple model significantly, which empirically validates the usefulness of knowing other same-URL labels. For the group-based split, we observe that the performance of the models drops for all variants, indicat-

Table 3: Results of different baseline models on both tasks, on either the full test dataset or focusing only on those test samples belonging to a URL already seen in training. The reported score is macro-F1, averaged over five runs alongside the standard deviation.

Model	**Impact Length**		**Impact Level**	
	Full	*SameURL*	*Full*	*SameURL*
BoW + LR	50.36 $\pm$ 4.17	48.32 $\pm$ 2.30	52.51 $\pm$ 3.05	45.82 $\pm$ 8.47
Emb + kNN(k=5)	47.01 $\pm$ 4.70	43.73 $\pm$ 3.56	50.14 $\pm$ 1.90	50.12 $\pm$ 5.20
SameURL-Labels	-	46.34 $\pm$ 9.23	-	56.59 $\pm$ 4.96
SameURL-BoW + LR	-	45.61 $\pm$ 8.62	-	55.67 $\pm$ 5.24
SameURL-Emb + kNN(k=5)	-	47.53 $\pm$ 4.28	-	56.51 $\pm$ 2.57
Stacked Model (*DICE 1*)	**51.52 $\pm$ 3.87**	**49.54 $\pm$ 5.89**	**59.68 $\pm$ 3.26**	**60.78 $\pm$ 4.66**

ing a much harder setup for the BoW-based model. This can be of interest to the organizers of similar future challenges if they want to restrict the models from taking advantage of the whole news article and making predictions based solely on the given text. Moreover, the post-process variant performs the same as the original baseline. This is expected since there are no cases where the test samples' URLs are in any of the training samples. Finally, the NER variant is the best-performing one (while also decreasing the standard deviation in performance), indicating that over-fitting on specific words that correspond to entities is not good for generalization. Thus, adding a NER pre-processing step could be helpful if the test set was created following this regime. For our submissions, we did not add the NER pre-processing, as the splits given by the organizers did not conform to this setup.

4.2. Baseline Approach

Following the observations mentioned above, we aimed to create a system that could do the following:

1. Capture specific words that are highly correlated with labels. To this end, we use a *BoW+LR* model as before (with no pre-/post-processing techniques).

2. Generalize to cases where the (highly) label-correlated vocabulary from (1) is not useful. To this end, we use a sentence embedding model (Reimers and Gurevych, 2019), specifically *all-mpnet-base-v2*, first to embed the news content of each item and then use a k-NearestNeighbor (kNN) classifier on-top. We denote this model as *Emb + kNN*.

3. Encapsulate information from same-URL training samples when possible to do so. To do this, we create three simple models that activate only in cases where a sample originates from a URL already seen in training.

   (a) *SameURL-Labels*: Calculates the probability of each label based on the frequency of the labels of all same-URL training samples.

   (b) *SameURL-BoW + LR*: Retrieves the BoW representations of all same-URL training samples and aggregates them by summation, using an LR classifier on the resulting feature vectors.

   (c) *SameURL-BoW + LR*: Retrieves the sentence embedding representations of all same-URL training samples and aggregates them by summation, using a kNN classifier on the resulting embeddings.

Having these five base models in place, our first submission is a stacked model that considers the probabilities for each class according to these models as input (i.e., a feature vector of length $3(labels) \times 5(models) = 15$) and uses an LR model for the final classification. The final LR classifier is trained using the predictions of the base models on the validation split. No hyper-parameter tuning is performed here.

The results of these models for both tasks on the English dataset are shown in Table 3. We report the performance both on the vanilla setup of the full (5-fold created) test sets (denoted with *Full*) and focusing only on the test samples that we've already seen in training (denoted with *SameURL*). The *SameURL-* models can only generate predictions for the *SameURL* subset of the test samples, so their performance is omitted (denoted with $-$). Essentially, that means that for the cases where a test sample originates from a URL not seen during training, the stacked model only utilizes the predicted probabilities of *Bow+LR* and *Emb+kNN*.

Regarding the performance of the models, predicting impact length seems much more difficult across all settings than impact level. If we focus on the difference under the *Full* setting between the two tasks, we see that the ensemble of *Bow+LR* and *Emb+kNN* is much more effective in the impact level task, denoting that these models make complementary predictions, while the slight increase in the performance of the ensemble indicates that they probably make the same mistakes when predicting impact length.

Regarding the *SameURL* setting and models, in the impact length task, the information from the *SameURL* models is not as helpful as in the level task. Interestingly, when we focus only on the *SameURL* test samples, the *SameURL-X* models, which use aggregates of information between the intra-URL data, perform better than the *Bow+LR* and *Emb+kNN* that use the actual test sample. This provides evidence that we should exploit the information from the *SameURL* samples.

4.3. Deep Learning Approaches

Having created the stacked baseline model, we now focus on improving performance, mainly on the impact length task with DL approaches. We experimented with models that utilize contextualized embeddings and incorporate prior knowledge from their pre-training process, whether domain-specific or general. Table 4 presents the performance of all such models.

We began with the generic BERT model (Devlin et al., 2018) in a frozen state, using it as an embedding model for the news content by averaging over the token embeddings of the last layers. Subsequently, we appended two additional layers and trained the model independently on impact level and length tasks. The results were much worse than the previously established baseline. Thus, we moved on to experimenting only with fine-tuned models. The performance of the fine-tuned BERT model, with the same classification heads as above, is shown in the second line in Table 4.

Since ESG-related narrative is too specific and domain-oriented and the amount of available data is limited, there is strong evidence that generic pretrained models may not capture the linguistic semantics of this particular task. Thus, we experimented with RoBERTa (Liu et al., 2019) and FinBERT (Araci, 2019), which are trained on larger and domain-specific data, respectively. However, they both failed to surpass generic BERT's performance. We therefore focused on learning representations for our data that uncover the actual ESG semantics. SetFit (Tunstall et al., 2022) is an efficient framework for few-shot tuning in low-resource scenarios, where a pre-training representation learning step is evolved. SetFit finetunes a sentence encoder while optimizing a triplet loss. Each triple tuple consists of three samples: two that share the same label (positive pair) and one sample of a different label. Then, it builds a classifier on top. SetFit achieved an improved performance at the expense of being too slow to train. However, it inspired us to implement a Contrastive Learning pre-training step.

Contrastive representation learning (Le-Khac et al., 2020) tries to distinguish between similar and dissimilar samples by comparing them. This unsupervised technique can be used as a pre-training step where the model tries to learn meaningful features to address a downstream task. What we contrast upon is called the "pretext task" and has to be aligned with the downstream task. In other words, when the model addresses this pretext task, it should learn highly informative features for the downstream task.

The pretext task we define is to distinguish between sentences that refer to the same ESG issue. Such sentences would be rephrases of a single news text. Thus, the task involves taking a news text, providing a rephrased version of it, and several other unrelated news texts, with the objective of learning a metric space that brings the original and rephrased sentences closer while distancing the irrelevant ones. We assume that this pre-training step will uncover the underlying semantics of ESG news and that the ensuing classifier will capitalize on this information.

In the contrastive learning setting, we need to define a similarity distribution to sample a positive or a negative sample pair (according to the pretext task). A common approach is to use augmentation techniques to get a positive pair for each sample and treat all the rest as negative pairs. We want an augmentation technique that keeps the ESG-related information intact. We used OpenAI's *gpt-3.5-turbo* model and generated three augmentations per sample with the following prompt: "*Rephrase the following in 3 ways. Use synonyms and keep the length close to the original*". An example of the original text and the corresponding generated augmentations can be seen below:

```
Original text: ESG-focused financial technology company Arabesque AI
        announced today the appointment of Dr. Carolina Minio Paluello as
        the company's new Chief Executive Officer.

Augmentation 1: Arabesque AI, a fintech firm with an emphasis on ESG,
        today declared the induction of Dr. Carolina Minio Paluello as
        their new CEO.

Augmentation 2: Today, Arabesque AI, a finance technology corporation
        focused on ESG, introduced Dr. Carolina Minio Paluello as its
        latest Chief Executive Officer.

Augmentation 3: Dr. Carolina Minio Paluello was announced today as the
        new CEO of ESG-dedicated fintech company Arabesque AI.
```

Having multiple ways to express the same ESG news, we consider a pair consisting of the original text and its augmented version as positive and two randomly selected original texts as negative pairs. In the generated pairs, we always include one original sample. It is possible for the augmentations to include texts with vocabulary that may not necessarily align with the narrative of our original data, along with ambiguities or even meaningless passages. This is why we demanded three of them and also added the sampling technique to address such cases and add variability to the vocabulary.

The contrastive loss used to learn this metric space is the following NTXent loss (Sohn, 2016).

Table 4: Final results on both tasks for the English language. The reported score is macro-F1, averaged over five runs alongside the standard deviation

Models	Impact Length	Impact Level
Stacked Model (*DICE 1*)	**51.52** $\pm$ **3.87**	**59.68** $\pm$ **3.26**
BERT	48.67 $\pm$ 5.20	55.17 $\pm$ 6.00
RoBERTa	45.50 $\pm$ 4.07	52.65 $\pm$ 4.37
FinBERT	45.89 $\pm$ 6.11	53.45 $\pm$ 3.31
SetFit	50.00 $\pm$ 4.00	57.22 $\pm$ 6.20
CL Variant 1 (*DICE 2*)	**51.77** $\pm$ **6.06**	53.66 $\pm$ 3.36
CL Variant 2 (*DICE 3*)	50.57 $\pm$ 7.48	52.23 $\pm$ 2.38
CL Variant 3	47.25 $\pm$ 6.00	50.17 $\pm$ 3.43

$$L_q = -\log \frac{\exp\left(q \cdot k_+/\tau\right)}{\sum_{i=0}^{K} \exp\left(q \cdot k_i/\tau\right)} \tag{1}$$

where q is an original sample in our batch of size K, τ is the temperature parameter set to 0.07. k_i is any original sample within the current batch, and k_+ is a positive augmentation of the current q original sample, uniformly sampled from the three available augmentations. We computed the loss per sample and optimized its mean aggregation over the batches. This loss maximizes the dot product (similarity) of the positive samples on the enumerator while pushing away the negative pairs (denominator). We pre-trained the generic BERT with two extra linear layers of 768 dimensions and a ReLU applied between them on the available data for 70 epochs, with a learning rate of $5e-6$, a batch size of 32 and an early stopping criterion of 7 epochs. Keeping only the BERT backbone, we froze its parameters and appended three linear layers, applying ReLU to the first two and dropout to the first one for the downstream classification task. We trained for 30 epochs, with a patient of 6, a batch size of 32, and a learning rate of $1e-3$.

We report three variants of this setting that involve the same pre-trained model. The first one targets impact length and level independently. The second adds the post-processing of the predictions as described earlier, and the third one uses two classification heads and tries to solve both tasks simultaneously. Table 4 summarizes all results plus the Stacked Model for comparison reasons.

The unexpected dominance of BERT over RoBERTa and FinBERT has already been noted. However, we should stress that we did not conduct thorough hyperparameter tuning for these models. SetFit was very promising but too slow and did not allow further experimentation. Additionally, it exhibited considerable variation among the five runs, especially in the impact level task, where it achieved the best macro F1 score. Regarding the Contrastive Learning setting, it is interesting that the post-processing step (Variant 2) resulted in a performance drop, unlike the baseline models, where we observed the opposite effect. That is also the case with the third variant, where we tried to leverage the tasks' correlation depicted in Fig. 1.

5. Official Results

Table 5: Final results on the official test sets, macro-F1 reported.

Models	Length (Rank)	Level (Rank)
DICE 1 - Eng.	37.07 (30)	53.11 (10)
DICE 2 - Eng.	**42.53** (14)	**55.27** (6)
DICE 3 - Eng.	37.84 (29)	55.08 (7)
DICE 1 - Fr.	34.45 (19)	44.80 (11)

Table 5 presents the results for our submissions in the blind test set. There is a noticeable deviation between our anticipated performance and the official evaluation, particularly concerning the impact length task. However, for the impact level task, we are much more aligned with our expectations and rank relatively high on the leaderboard. *DICE 1*, although it was our best-performing model in our evaluation setting, performed poorly. We also noticed a significant decrease in performance when applying our post-processing step to the contrastive pre-trained model to the impact length task. However, the impact level appears to remain unaffected.

Since the workshop organizers released the test set ground truths, we also performed an error analysis. Following our intuitions regarding the information shared between same-URL samples, we analyzed the performance of the models separately on two subsets. The first subset contains all the test samples with URLs that exist in our training set (denoted as *SameURL*). The second contains those that originate from unseen URLs (denoted as *!SameURL*). Tables 6 and 7 display the corresponding test results.

Overall, there is a massive increase in the scores concerning not previously seen URL articles except

Table 6: Performance of submitted models on the test set for impact length, when grouping samples on whether we've encountered a same-URL sample in training (*SameURL*) or not *!SameURL*.

Model	Impact Length		
	Full	*SameURL*	*!SameURL*
DICE 1	37.07	29.15 ↓	48.09 ↑
DICE 2	42.53	35.92 ↓	**52.86** ↑
DICE 3	37.84	28.80 ↓	**52.86** ↑

Table 7: Performance of submitted models on the test set for impact level, when grouping samples on whether we've encountered a same-URL sample in training (*SameURL*) or not *!SameURL*.

Model	Impact Level		
	Full	*SameURL*	*!SameURL*
DICE 1	53.11	**59.52** ↑	44.99 ↓
DICE 2	55.27	48.52 ↓	**63.22** ↑
DICE 3	55.08	44.46 ↓	**63.22** ↑

for the *DICE 1* model on the level task. All the models seem to have overfitted entities found in the training data, with the contrastive models being the ones that generalize better in both cases. Moreover, it is essential to note the effectiveness of the *DICE 1* in utilizing information on the *SameURL* group for the impact level task, as shown in Table 7. This is the only case that performs better on the *SameURL* group than the entire test set. This is in line with the findings of our analysis, as also shown in Table 3, where the models that utilize information from other *SameURL* articles perform very well when predicting the impact level of the sample at hand. This effect is not observed, though, for impact length in both cases as expected (i.e., both in Tables 3, 6), which is due to the much lower intra-URL label correlation.

Concerning the contrastive learning models, we observe a drop in performance for the *SameURL* setting. This drop is probably related to the way we conducted the contrastive pretraining. Due to the pretext task we defined, the embeddings of *SameURL* samples are forced apart because they constitute negative pairs in this context. This, when combined with high intra-URL label correlation (e.g., impact level), has a negative effect on the final downstream task. It would be interesting to incorporate the above observations in the contrastive learning setting, which we leave as a future work.

6. Conclusion

The complex nature of the ML-ESG-3 shared task provides an excellent opportunity to experiment with various methods in the domain of ESG under challenging conditions. In this work, we focused on identifying the impact level and length duration of ESG issues found in news articles, based on the English dataset that the organizers distributed. In this setting, we demonstrated how the correlation between texts originating from the same articles impacts the overall performance of different models. Our explanatory analysis revealed that the class labels, at least in the English data, were closely linked to specific tokens, such as the names of companies, nouns, and verbs related to specific ESG issues. To mitigate this bias in the data, we experimented with various baseline systems, pre-/post-processing techniques, and contrastive pre-training. In both subtasks of the ML-ESG-3, our best-performing system was the one based on contrastive pre-training.

Regarding future directions and following our findings regarding news events originating from the same news article, as well as correlations between impact length and impact level, we plan to focus on methodologies that consider multiple sources of information. For example, information stemming from the latest SEC filing regarding any ESG disclosure or other news sources at the time of the news event under examination, alongside other historical information regarding ESG-related activities of the company.

7. Acknowledgements

The authors would like to acknowledge the financial support of Qualco SA for this project. The opinions of the authors expressed herein do not necessarily state or reflect those of Qualco SA.

8. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *CoRR*, abs/1908.10063.

Fabian Billert and Stefan Conrad. 2023. Team HHU at the FinNLP-2023 ML-ESG task: A multi-model approach to ESG-key-issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 146–150, Macao. -.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal*

AI For Financial Forecasting, pages 111–115, Macao.

Valeria D'Amato, Rita D'Ecclesia, and Susanna Levantesi. 2021. Fundamental ratios as predictors of esg scores: a machine learning approach. *Decisions in Economics and Finance*, 44(2):1087–1110.

Valeria D'Amato, Rita D'Ecclesia, and Susanna Levantesi. 2022. Esg score prediction through random forest algorithm. *Computational Management Science*, 19(2):347–373.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

Tian Guo, Nicolas Jamet, Valentin Betrix, Louis-Alexandre Piquet, and Emmanuel Hauptmann. 2020. Esg2risk: A deep learning framework from esg news to stock volatility prediction. *arXiv preprint arXiv:2005.02527*.

Akshat Gupta, Utkarsh Sharma, and Sandeep Kumar Gupta. 2021. The role of esg in sustainable development: An analysis through the lens of machine learning. In *2021 IEEE International Humanitarian Technology Conference (IHTC)*, pages 1–5.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp.

Juyeon Kang and Ismail El Maarouf. 2022. FinSim4-ESG shared task: Learning semantic similarities for the financial domain. extended edition to ESG insights. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 211–217, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907–193934.

Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. EaSyGuide : ESG issue identification framework leveraging abilities of generative large language models. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 127–132, Macao. -.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Vincent Margot, Christophe Geissler, Carmine De Franco, Bruno Monnier, France Advestis, and France Ossiam. 2021. Esg investments: filtering versus machine learning approaches. *Applied Economics and Finance*, 8(2):1–16.

Ivan Mashkin and Emmanuele Chersoni. 2023. HKESG at the ML-ESG task: Exploring transformer representations for multilingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 140–145, Macao. -.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging BERT language models for multi-lingual ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 121–126, Macao.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

Akshat Gupta Utkarsh Sharma and Sandeep Kumar Gupta. 2024. The pertinence of incorporating esg ratings to make investment decisions: a quantitative analysis using machine learning. *Journal of Sustainable Finance & Investment*, 14(1):184–198.

Jingli Wang, Ashok Bhowmick, Mucahit Cevik, and Ayse Basar. 2020. Deep learning approaches to classify the relevance and sentiment of news articles to the economy. In *Proceedings of the 30th Annual International Conference on Computer Science and Software Engineering*, pages 207–216.

Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging contrastive learning with BERT for ESG issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 116–120, Macao. -.

Guangliang Yu, Yukun Liu, William Cheng, and Chun-Te Lee. 2022. Data analysis of esg stocks in the chinese stock market based on machine learning. In *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 486–493.

Fine-tuning Language Models for Predicting the Impact of Events Associated to Financial News Articles

Neelabha Banerjee[1], Anubhav Sarkar[1], Swagata Chakraborty[1], Sohom Ghosh[2], Sudip Kumar Naskar[2]
[1]Christ (Deemed to be University), [2]Jadavpur University
India
{neelabha.12.banerjee, sarkaranubhav2001, swagatac652}@gmail.com
{sohom1ghosh, sudip.naskar}@gmail.com

Abstract

Investors and other stakeholders like consumers and employees, increasingly consider ESG factors when making decisions about investments or engaging with companies. Taking into account the importance of ESG today, FinNLP-KDF introduced the *ML-ESG-3* shared task, which seeks to determine the duration of the impact of financial news articles in four languages - English, French, Korean, and Japanese. This paper describes our team, LIPI's approach towards solving the above-mentioned task. Our final systems consist of translation, paraphrasing and fine-tuning language models like BERT, Fin-BERT and RoBERTa for classification. We ranked first in the impact duration prediction subtask for French language.

Keywords: ESG, financial natural language processing, impact prediction, language models, ESG impact prediction

1. Introduction

The Multi-Lingual ESG Impact Duration Inference (ML ESG-3) task being organised in conjunction with the FinNLP-KDF@LREC-COLING-2024 deals with predicting the impact of events on companies. Determining the duration of an impact, an event might have on a company in the context of Environmental Social and Governance (ESG) factors could be crucial for understanding and managing the risks or opportunities associated with that event. Predicting the duration of an impact might involve fine-grained analysis of historical data, sentiment analysis, and other relevant information from news articles. In this paper, we talk about our team LIPI's approach of solving the subtasks of ML ESG- 3. This can be the first step towards achieving the long-term goal of developing multilingual systems that can assess the potential short-term and long-term effects of specific events on a company's performance, reputation, or other ESG-related aspects. We present this in Figure 1.

Our contributions

Our contributions include developing a framework that finetunes pre-trained language models for classifying the impact and duration of an event associated with multi-lingual news articles. We open-sourced the code[1] so that the research community can utilize them as baselines.

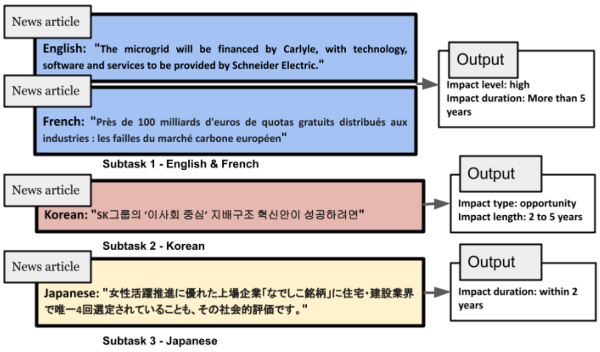


Figure 1: Overview of the ML-ESG3 task

2. Problem Statement

The multilingual data set of the shared task ML-ESG-3[2] consists of financial news articles in different languages such as English, French, Japanese, and Korean (Chen et al., 2024) (Kannan and Seki, 2023). The design of the task varies slightly across different languages. It is described as follows:

- **English and French**: Given a financial news article in English or French, the objective is to determine its *impact level* and predict its *impact length*. The impact length can be "low",

This paper got accepted at FinNLP-KDF-ECONLP workshop of LREC-COLING 2024.

[1]https://github.com/Neel-132/ML-ESG3_LIPI

[2]https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-kdf-2024/shared-task-ml-esg-3 (accessed on 3rd Feb 2024)

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 244–247
20 May, 2024.

"medium" or "high". The impact length can be "Less than 2 years", "2 to 5 years", or "More than 5 years".

- **Japanese**: Given a financial news article in Japanese, the objective is to predict its *impact duration*. The impact duration can be "Less than 2 years", "2 to 5 years", or "More than 5 years".
- **Korean**: Given a financial news article in Korean, the goal is to determine its *impact type* and predict its *impact length*. The impact type can be between "opportunity", "risk", or "cannot distinguish" and the impact length can be "less than 2 years", "2 to 5 years", or "more than 5 years".

3. System Descriptions

The pipeline for handling the tasks mentioned above comprises the following steps:

- Step 1: **Translation** - Although there are several powerful multilingual encoder models present, our experiments revealed that they were not very efficient in learning the intricate patterns in the dataset and thereby correctly predicting the impact type and duration of news articles. Thus, we primarily translated the non-English datasets into English before proceeding with modelling.
- Step 2: **Paraphrase** - We found that as the given data set was small, the classification models were overfitting. To solve this, we paraphrased the translated data set returned by the translation module as mentioned in Step 1 using a T5-based model (Vladimir Vorobev, 2023).
- Step 3: **Classification** - After paraphrasing comes the final module of the pipeline. This is the classification module. Since the target variable differed slightly across different datasets, we designed two different classification modules for the three tasks given as follows:
  - Module 1 (for English, French & Korean): The English, French, and Korean dataset has two target variables. For English and French, they are *impact level* and *impact length*. For Korean, they are *impact type* and *impact length*. We used pre-trained encoder models like BERT (Devlin et al., 2018), DistilBERT (Sanh et al., 2019), etc. to learn the embeddings of the content as given by the paraphrase module, followed by a linear layer to predict the target which can be impact length, impact type, or impact level. The number of classes in each of these target variables is used as a hyperparameter to specify the output of the linear layer.
  - Module 2 (For Japanese): The Japanese dataset has only one target variable, *impact duration*. The *impact type* was given for this dataset. So, we developed the second module to learn the pre-trained text embeddings using the same encoder models, but for two features which are news content and impact type, followed by a concatenation operation. Finally, we added a linear layer to predict the output.

We present this in Figure 2.

Dataset	Model	macro F-1	micro F-1
English	XG-Boost	**0.35**	**0.31**
	SVM	0.29	0.26
	DNN	0.32	0.27
French	XG-Boost	0.23	0.22
	SVM	0.21	0.21
	DNN	**0.33**	**0.34**
Japanese	XG-Boost	**0.12**	**0.09**
	SVM	0.08	0.05
	DNN	0.11	0.10
Korean	XG-Boost	0.34	0.34
	SVM	0.27	0.22
	DNN	**0.42**	**0.34**

Table 1: Result of the Baselines

4. Experiments and Results

In this section, we describe the experiments we performed, and the corresponding results.

4.1. Baseline

For the baseline, we chose BERT-base uncased (for English) and BERT-base multilingual (Devlin et al., 2018) uncased (for other languages) to learn the pre-trained embeddings of news content and used them to train classical machine learning algorithms like XG-Boost (Chen and Guestrin, 2016), Support Vector Machine (Cortes and Vapnik, 1995), and deep learning based algorithms like Multi-layered Perceptron with just one hidden layer.
The results corresponding to it are presented in Table 1.

4.2. Experiment 1

The first experiment towards improving on the baseline had three stages, depending on the language of the dataset. For the non-English datasets like

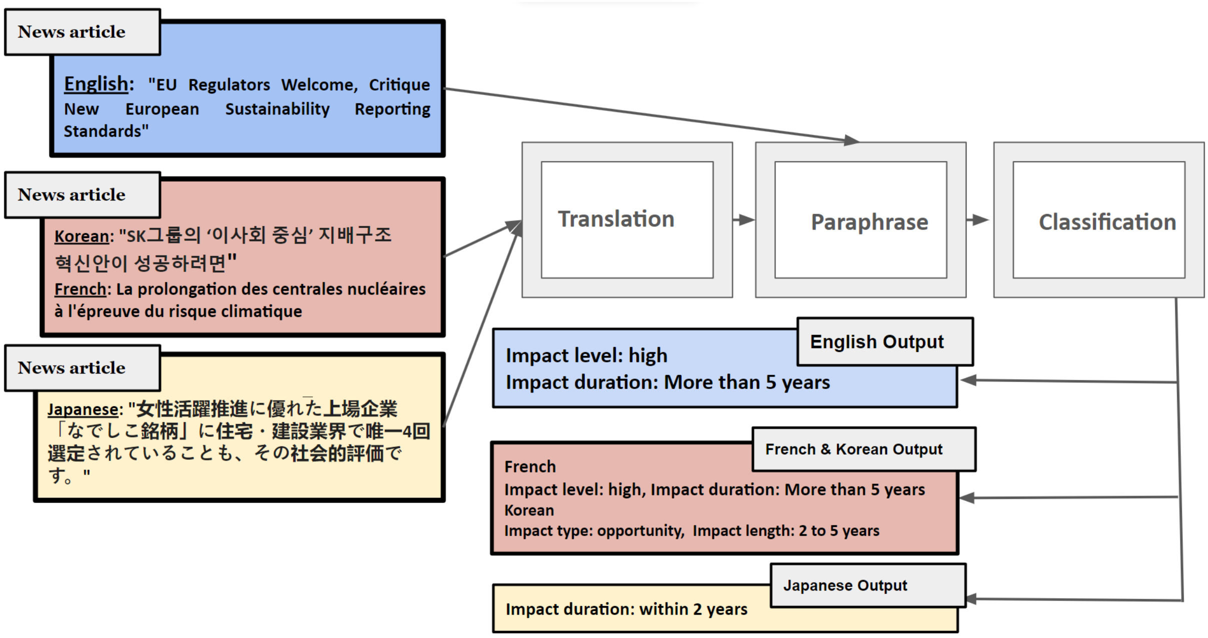


Figure 2: Our framework

French, Korean, we firstly translated the news content into English using Google Translate. In the next step, we paraphrased each data point using a T5 based paraphraser (Vladimir Vorobev, 2023) with a beam size of 5, temperature of 0.7, and repetition penalty of 10.
In the final step, we fine-tuned pre-trained encoder models like BERT (Devlin et al., 2018), DistillBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2020), Fin-BERT (Araci, 2019), etc. for the task of classifying the news articles to their respective impact type/level. We used a learning rate of e^{-5}, and a weight decay of 0.005 and fine-tuned the models for 30 epochs. Our best-performing models were BERT-base-uncased (Devlin et al., 2018) for English, RoBERTa (Liu et al., 2020) for Korean, and FinBERT (Araci, 2019) for French .
The results are presented in Table 2.

Dataset	Model	macro F-1	micro F-1
English	BERT-base-uncased	**0.99**	**0.99**
	FinBERT	0.97	0.97
	DistillBERT-multiling	0.70	0.68
	DistillBERT-base	0.68	0.69
	NLI-Distilroberta-base	0.81	0.80
	Distilroberta Financial	0.75	0.78
	XLI Roberta base	0.83	0.81
	RoBERTa-base	0.98	0.97
Korean	BERT-base-uncased	**0.95**	**0.94**
	FinBERT	0.94	0.93
	DistillBERT-multiling	0.78	0.71
	DistillBERT-base	0.76	0.69
	NLI-Distilroberta-base	0.82	0.81
	Distilroberta Financial	0.67	0.64
	XLI Roberta base	0.75	0.71
	RoBERTa-base	0.96	0.93
French	BERT-base-uncased	0.93	0.93
	FinBERT	**0.94**	**0.93**
	DistillBERT-multiling	0.57	0.49
	DistillBERT-base	0.91	0.90
	NLI-Distilroberta-base	0.51	0.45
	Distilroberta Financial	0.47	0.46
	XLI Roberta base	0.63	0.67
	RoBERTa-base	0.91	0.92

Table 2: Results of Experiment-1

4.3. Experiment 2

Since the Japanese dataset had only one objective, i.e. to predict the *impact duration*, we used the *impact type* as another feature along with the news content. Like the first experiment mentioned above, we translated the data into English, followed by paraphrasing with the same model, and configurations as mentioned in Experiment 1. Finally, we fine-tuned pre-trained models mentioned in Experiment 1 for assessing impact duration of news articles in Japanese.
Furthermore, we concatenated the embeddings of news content and impact type followed by a linear layer before the final output layer. We used a learning rate of e^{-4} and a weight decay of 0.006 and trained the models for 30 epochs. Our top performing models were BERT-base-uncased (Devlin et al., 2018), RoBERTA-base (Liu et al., 2020) and FinBERT (Araci, 2019).
The results are presented in Table 3.

4.4. Experiment 3

Since the English and French datasets had the same objective of predicting the impact level and impact length, we experimented with fine-tuning the pre-trained models (mentioned in both of the

Dataset	Model	macro F-1	micro F-1
Japanese	BERT-base-uncased	**0.67**	**0.69**
	FinBERT	0.55	0.52
	DistillBERT-multiling	0.52	0.48
	DistillBERT-base	0.36	0.32
	NLI-Distilroberta-base	0.51	0.48
	Distilroberta Financial	0.43	0.48
	XLI Roberta base	0.49	0.51
	RoBERTa-base	0.68	0.67

Table 3: Results of Experiment-2

Dataset	Model	macro F-1	micro F-1
English and French	BERT-base-uncased	**0.79**	**0.79**
	FinBERT	0.67	0.62
	DistillBERT-multiling	0.34	0.41
	DistillBERT-base	0.51	0.55
	NLI-Distilroberta-base	0.57	0.61
	Distilroberta Financial	0.47	0.45
	XLI Roberta base	0.51	0.50
	RoBERTa-base	0.76	0.76

Table 4: Results of Experiment-3

previous experiments) on the English dataset and testing them on the French dataset. The hyperparameters were the same as those of Experiment 2 and the results corresponding to it are mentioned in Table 4.

5. Conclusion

In this paper, we share our team, LIPI's approach for determining the duration of an event's impact on the company. We translated the non-English datasets into English and further paraphrased them before fine-tuning the encoder-based pre-trained language models on them. Our observations revealed the best performing models were BERT(Devlin et al., 2018) for English and Japanese; RoBERTa (Liu et al., 2020) for Korean, and FinBERT(Araci, 2019) for French. We achieved a significant increase in performance with translation and paraphrasing. Finally, we proposed a unified framework for all the languages.
Our team ranked 3^{rd} in both of the sub-tasks of the English dataset, 1^{st} in the first sub-task(impact-length) and 8^{th} in the second sub-task(impact-level) of the French dataset, 20^{th} in the first sub-task(impact-length) and 13^{th} in the second sub-task(impact-type) of the Korean dataset, and 11^{th} in the Japanese dataset.
However, we did not consider the semantic loss while paraphrasing and also had to translate the dataset into English to seek improvement. The future scope of this paper involves, designing better language models for low-resourced languages, improving the computational aspect of the algorithms, and extending the solution to cater to bigger and more important needs.

6. Bibliographical References

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for ESG scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao. -.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

CriticalMinds: Enhancing ML Models for ESG Impact Analysis Categorisation Using Linguistic Resources and Aspect-Based Sentiment Analysis

Iana Atanassova[*,†], **Marine Potier**[*], **Maya Mathie**[*], **Marc Bertin**[‡], **Panggih Kusuma Ningrum**[*]
[*]Université de Franche-Comté, CRIT
F-25000 Besançon, France
{iana.atanassova, panggih_kusuma.ningrum}@univ-fcomte.fr,
{marine.potier02, maya.mathie}@edu.univ-fcomte.fr

[†]Institut Universitaire de France (IUF)

[‡]ELICO, Université Claude-Bernard Lyon 1
43 Bd. du 11 novembre 1918, 69622 Villeurbanne cedex, France
marc.bertin@univ-lyon1.fr

Abstract

This paper presents our method and findings for the ML-ESG-3 shared task for categorising Environmental, Social, and Governance (ESG) impact level and duration. We introduce a comprehensive machine learning framework incorporating linguistic and semantic features to predict ESG impact levels and durations in English and French. Our methodology uses features that are derived from FastText embeddings, TF-IDF vectors, manually crafted linguistic resources, the ESG taxonomy, and aspect-based sentiment analysis (ABSA). We detail our approach, feature engineering process, model selection via grid search, and results. The best performance for this task was achieved by the Random Forest and XGBoost classifiers, with micro-F1 scores of 47.06 % and 65.44 % for English Impact level and Impact length, and 39.04 % and 54.79 % for French Impact level and Impact length respectively.

Keywords: ABSA, ESG, Impact level, Impact length, ESG taxonomy, linguistic resources

1. Introduction

After the establishment of Environmental, Social, and Governance (ESG) criteria in 2004 (United Nations, 2004), the incorporation of ESG principles within corporations has become a topic of extensive discussion (Berg et al., 2022). The advent of FinNLP challenges explore the opportunity to employ Natural Language Processing methodologies in this domain (Aue et al., 2022; Del Vitto et al., 2023; Schimanski et al., 2024).

The ML-ESG 2024 shared task focuses on multilingual ESG impact type and duration inference, particularly in languages including English and French. The tasks for English and French involve annotations for "Impact Level" (low, medium, high) and "Impact Length" (less than 2 years, 2 to 5 years, more than 5 years) based on the MSCI ESG rating guidelines (Chen et al., 2024).

Our objective in participating in this task, as CriticalMinds team, is to propose a competitive Machine Learning (ML, low resource) approach and evaluate the contribution of several types of features: manually crafted linguistic resources exploiting the ESG taxonomy, and features derived from aspect-based sentiment analysis (ABSA).

2. Method

In this section, we first introduce the datasets employed in the analysis. We then detail the feature types implemented in our experiments with ML models, along with specifications regarding the feature sets' dimensions. Finally, we describe the procedure for model selection and present the corresponding results.

2.1. Data Description

The datasets used in this experiment cover two languages, English and French. For both languages, the training and test sets were provided in json format, with the following variables for each news article: URL, news_title, news_content, impact_level, impact_length. The latter two variables contain the annotated categories in the training set.

We identified a total of 48 duplicate entries within the French training dataset. These duplicates were excluded from subsequent analyses due to inconsistencies between the 'impact_level' and 'impact_length' labels, which rendered the determination of the correct labels ambiguous. Following this data cleaning processes, Table 1 presents the distributions of annotations for 'Impact Length' and 'Impact Level' for the training datasets.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 248–253
20 May, 2024.

Table 1: Distribution of annotations in the training sets in English and in French

Category		En	Fr
Impact length	Less than 2 years	82	110
	Between 2 and 5 y.	198	218
	More than 5 years	265	285
Impact level	low	106	117
	medium	243	305
	high	196	191
Total		545	613

2.2. Features extraction and selection

In our experiment, we tested combinations of different types of features that we describe below. We designed five types of features:

1. FastText embeddings (Bojanowski et al., 2017; Grave et al., 2018) word vectors;
2. TF-IDF vectors;
3. Features derived from the ESG taxonomy;
4. Linguistic resources to capture expressions of uncertainty and temporal data;
5. Aspects extracted by ABSA.

To calculate the first two types of features, FastText embeddings and TF-IDF, we used the text from the news_title and news_content fields. These were concatenated, then tokenized and lemmatized using `nltk WordNetLemmatizer`. Stop words were also removed. To reduce the dimension of TF-IDF vectors, we used only the 25 terms having the highest discriminatory power. This value was adjusted experimentally.

For the rest of the features, the original values of news_title and news_content fields were used. We describe these features in more detail in the following subsections.

2.2.1. Features derived from ESG taxonomy

As the task of classifying EGS impact durations and levels is essentially related to the semantics of the ESG taxonomy[1], we used the terms denoting ESG issues, sectors and subsectors in the following way. We defined as features the number of occurrences of the issues, sectors and subsectors in the ESG taxonomy. Moreover, for each issue, sector and subsector, we consider lists of synonym expressions that can be present in the news articles and that were curated manually and represented as regular expressions. The figure 1 shows an example of regular expressions in English related to the 'energy' subsectors.

```
'energy': [
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Dd]rilling|[Oo]il\s+[Dd]rilling|[Gg]as\s+[Dd]rilling\b',
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Ee]quipments?\s+(?:and|&)\s+[Ss]ervices?|[Oo]il\s+(?:and
r'\b[Ii]ntegrated\s+[Oo]il\s+(?:and|&)\s+[Gg]as|[Ii]ntegrated\s+[Oo]il|[Ii]ntegrated\s+[Gg]
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Ee]xplorations?\s+(?:and|&)\s+[Pp]roductions?|[Oo]il\s+(
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Rr]efining\s+(?:and|&)\s+[Mm]arkets?(?:ing)?|[Oo]il\s+(?
r'\b[Oo]il\s+(?:and|&)\s+[Gg]as\s+[Ss]torage\s+(?:and|&)\s+[Tt]ransports?(?:ation)?|[Oo]il\
r'\b[Cc]oal\s+(?:and|&)\s+[Cc]onsumable\s+[Ff]uels?|[Cc]oal\s+[Ff]uels?|[Cc]onsumable\s+[Ff
```

Figure 1: Excerpt from the lists of regular expressions related to the 'energy' subsectors

2.2.2. Linguistic resources

The prediction of Impact level is related to the notion of uncertainty. For this reason, we used as features the number of occurrences of lists of uncertainty and hedging cues in news_title and news_content. In particular, we used the lists defined in (Atanassova et al., 2018).

For the prediction of Impact length, we created lists of temporal expressions that denote various time spans such as "over the next 2 years", "by 2026", etc. They were implemented as regular expressions and their numbers of occurrences were used as features.

Experimentally, we found that these linguistic resources features improve the micro-F1 scores of our models of about 1 % to 2 %.

2.2.3. Aspects extraction

In our study, we leveraged Aspect-Based Sentiment Analysis (ABSA) to dissect and extract significant aspects from textual content, marking it as an advanced segment of sentiment analysis that precisely pinpoints text components and evaluates the sentiments tied to them (Hua et al., 2023). By integrating a combination of linguistic, statistical, and machine learning techniques, and utilizing resources like annotated datasets, lexicons, and ontologies, ABSA achieves a high level of analytical precision (Fan et al., 2020).

ABSA provides a way to examine the textual aspects, which is particularly useful when working with complex datasets such as ESG news articles. These articles often contain discussions on multiple aspects of ESG criteria within the same paragraph or article. By employing a transfer learning approach with a fine-tuned ABSA model[2], we could effectively parse and understand the nuanced sentiments associated with specific ESG aspects. This selected model, optimized within the SetFit ABSA framework and utilizing Sentence Transformer embeddings (Tunstall et al., 2022), is

[1] https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map

[2] *joshuasundance/setfit-absa-all-MiniLM-L6-v2-laptops-aspect* from Hugging Face

particularly suited for natural language understanding tasks, enabling precise analysis at the sentence level in ESG news dataset.

Upon reviewing the ESG news dataset, we noted a predominance of neutral sentiments (82.4 %), reflecting the objective presentation style typical of news articles. However, this neutrality does not diminish the utility of ABSA; on the contrary, it allows us to mine the texts for the specific aspects they discuss, shedding light on crucial ESG themes relevant to corporate conduct. This aspect-oriented analysis method, as supported by Hua et al. (2023), provides a deeper dive into key detail information in texts, reaching beyond the surface level of sentiment polarization.

These extracted aspects were then incorporated as features in our ML model, grouping them by their impact_level and impact_length. We calculated the frequency of these aspect occurrences in the news_title and news_content, where the numbers of occurrences were calculated with respect to several cut-off values of the lists for French and for English. The choice of the cut-off values was optimized through grid search.

Figure 2 shows the aspects detected from the English training set grouped by category.

Table 2 shows the cut-off values that were used for English and French, leading to 17 and 11 derived features, respectively.

Table 2: Aspect lists cut-off values N

En	[10, 25, 50, 75, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900, 1100]
Fr	[25, 50, 100, 150, 200, 300, 400, 500, 750, 1000, 1500]

2.3. Feature set dimensions

We employed Principal Component Analysis (PCA) (Jolliffe, 2002) to reduce the dimensions of some of the sets of features, namely the number of dimensions for the FastText embeddings and for the features derived from the ESG taxonomy. This was necessary for two reasons. Firstly, high-dimensional data can complicate model training and possibly lead to overfitting. Secondly, the features that are based on the linguistic resources and the aspects have a fixed dimension, and therefore we need to find the correct balance between the number of dimensions for these features and the ones derived from the embeddings and the ESG taxonomy.

During the grid search phase of our model optimization, we tested various combinations for the numbers of these dimensions, ranging from 5 to 80 dimensions, to find the best configuration for the prediction of each category. Table 3 presents the dimensions of the different types of features that were used with the best model configurations.

2.4. Model Selection

In order to identify the optimal Machine Learning (ML) models, hyperparameters, and to adjust the number of dimensions that were used for the FastText embeddings and TF-IDF features, we performed grid search on the training set. 20 % of the dataset was used for performance evaluation and the rest was used for training with 4-fold cross validation. We used grid-search by maximizing the micro F1 score to test models, including Support Vector Machines (SVM), Random Forest, Gradient Boosting, Logistic Regression, K-Nearest Neighbors (KNN), Extreme Gradient Boosting (XGBoost), LightGBM, and CatBoost. Key hyperparameters tested included kernel types and regularization parameters for SVM, number of estimators and depth for tree-based models, to distance metrics and weights for KNN. For the implementation of the models we used the python `sklearn`, `xgboost`, `catboost` and `lightgbm` libraries.

Table 4 presents the two best models with their hyperparameters, dimensions of features after PCA and results on the training set.

3. Results

Table 5 shows the results obtained by the CriticalMinds team on the test set. To obtain these results, we executed both the Random Forest (RF) and Extended Gradient Boosting (XGB) models five times each, and then selected the most consistently observed predictions across these iterations.

To show the contribution of the different types of features, table 6 presents the results of both models and compares the scores obtained using: the features derived from embeddings (Emb), for TF-IDF and linguistic resources (LR), with adding the features derived from the ESG taxonomy (F-ESG), and those from ABSA. These results show that the features derived from the ESG taxonomy and ABSA improve the performance in most cases. In particular, adding ABSA derived features improves the micro-F1 scores in 4 cases with 2.85 % on average, while it reduces the performance in three cases but with only 1.87 % on average.

4. Discussion

The use of Aspect-Based Sentiment Analysis (ABSA) as strategy in feature engineering is an original approach that aims to improve the semantic representation of textual data. The results in table 6 show the variable impact of ABSA across

Table 3: Number of dimensions for the different models and types of features

Category	Embeddings	TF-IDF	ESG taxonomy	Linguistic	ABSA-derived	Total
Random Forest						
En Impact Level	19	25	12	4	17	77
En Impact Length	75	25	10	3	17	130
Fr Impact Level	12	25	36	4	11	88
Fr Impact Length	70	25	28	3	11	137
XGBoost						
En Impact Level	20	25	15	4	17	81
En Impact Length	75	25	20	3	17	140
Fr Impact Level	18	25	40	4	11	98
Fr Impact Length	75	25	36	3	11	150

Table 4: Best models and results on the training set

Category	Hyperparameters	Micro-F1
Random Forest		
En Impact Level	'criterion': 'gini', 'n_estimators': 400, 'max_depth': None	86.24 %
En Impact Length	'criterion': 'log_loss', 'n_estimators': 400, 'max_depth': None'	79.82 %
Fr Impact Level	'criterion': 'log_loss', 'n_estimators': 500, 'max_depth': None'	71.54 %
Fr Impact Length	'criterion': 'log_loss', 'n_estimators': 200, 'max_depth': None	66.67 %
XGBoost		
En Impact Level	'learning_rate': 0.1, 'n_estimators': 200, 'max_depth': 9	84.40 %
En Impact Length	'learning_rate': 0.1, 'n_estimators': 400, 'max_depth': 9	77.06 %
Fr Impact Level	'learning_rate': 0.1, 'n_estimators': 300, 'max_depth': 7	65.04 %
Fr Impact Length	'learning_rate': 0.1, 'n_estimators': 400, 'max_depth': 5	68.29 %

Table 5: Micro-F1 and Macro-F1 Scores for Impact Length and Impact Level on the test set

		English		French	
	Model	Impact Length	Impact Level	Impact Length	Impact Level
micro F1	CriticalMinds_1 (RF)	64.71 %	47.06 %	54.79 %	36.30 %
	CriticalMinds_2 (XGB)	59.56 %	42.65 %	46.58 %	39.04 %
	CriticalMinds_3 (RF + XGB)	65.44 %	45.59 %	54.11 %	36.30 %
macro F1	CriticalMinds_1 (RF)	42.81 %	43.16 %	30.33 %	22.48 %
	CriticalMinds_2 (XGB)	41.53 %	39.59 %	32.19 %	37.96 %
	CriticalMinds_3 (RF + XGB)	43.86 %	40.64 %	32.88 %	26.21 %

Table 6: Micro-F1 scores on the training set with different subsets of features. Emb = Embeddings, LR = Linguistic resources, F-ESG = ESG taxonomy features. The last column presents the final results (as in table 5) using Emb+TF-IDF+LR+F-ESG and also Aspect-based Sentiment Analysis features.

Category	Features		
	Emb+TF-IDF+LR	Emb+TF-IDF+LR+F-ESG	All
Random Forest			
En Impact Level	44.85 %	45.59 %	47.06 %
En Impact Length	61.76 %	62.50 %	64.71 %
Fr Impact Level	36.30 %	37.67 %	36.30 %
Fr Impact Length	54.11 %	54.79 %	54.79 %
XGBoost			
En Impact Level	42.65 %	45.59 %	42.65 %
En Impact Length	61.76 %	57.35 %	59.56 %
Fr Impact Level	38.36 %	33.56 %	39.04 %
Fr Impact Length	45.89 %	47.95 %	46.58 %

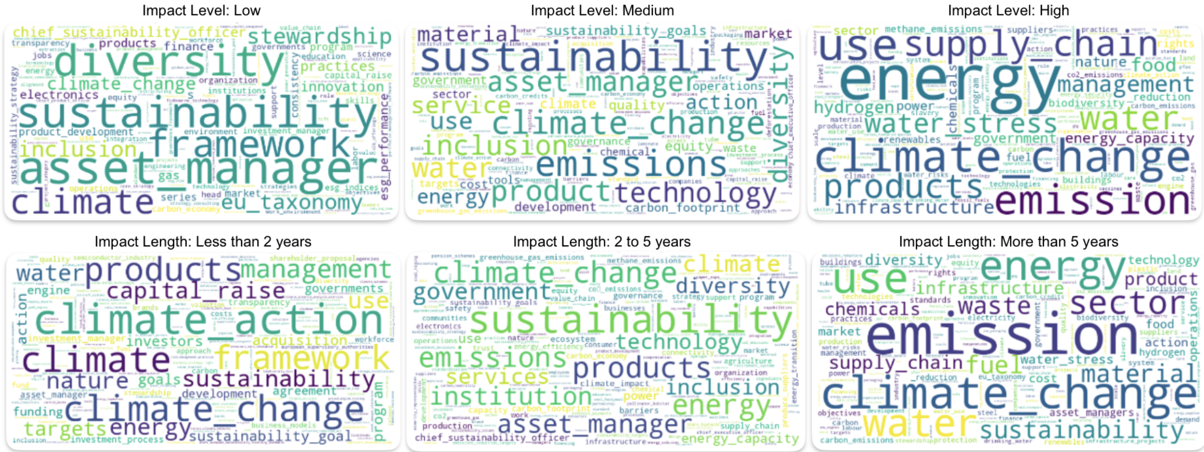


Figure 2: Wordclouds representing aspects detected from the English training set grouped by impact_level and impact_length

different models and languages. Specifically, the inclusion of features from ABSA appears to enhance the predictions in English, underscoring the value of capturing sentiment at a granular level. However, the results also indicate a complex interplay between aspects and other features, suggesting that the contribution of ABSA depends on the model and the linguistic characteristics of the dataset.

The results of our study should be interpreted in the light of several limitations. Firstly, the dependence on linguistic resources makes this approach difficult to deploy for multilingual processing. We specifically curated the lists of regular expressions for English and for French. This task is often time-consuming. We will publish all resources in order to ensure the reproducibility of this experiment.

We choose to use FastText embeddings because of the relatively small size of the models and the ease of use on low-resource machines. FastText embeddings capture subword information and allow representing out-of-vocabulary words. This makes them particularly relevant for processing news articles that may contain numerous new terms and named entities. However, other types of embeddings should be tested as they might improve the results.

The quality of the training data is critical for the success of ML models. During our investigation, we encountered several cases of duplicated annotations, particularly within the French dataset, which were inconsistent and required meticulous cleaning before proceeding with data processing.

Furthermore, in our experimentation, we explored whether augmenting the training set with translated datasets can improve the performance of the models. Specifically, we augmented the training datasets by translating the English dataset into French and vice versa, using ChatGPT-4. Contrary to our expectations, we observed a systematic decline in the performance of all models when the training sets were augmented in this manner. This suggests that the expression of ESG-related information is highly language-specific. This finding underscores the importance of developing language-specific models and training sets for such tasks.

5. Acknowledgements

This work was supported by French ANR grant number ANR-21-CE38-0003-01 (Project InSciM).

6. Bibliographical References

Iana Atanassova, François-C. Rey, and Marc Bertin. 2018. Studying Uncertainty in Science: a distributional analysis through the IMRaD structure. In *7th International Workshop on Mining Scientific Publications (WOSP 2018) at 11th edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.

Tanja Aue, Adam Jatowt, and Michael Färber. 2022. Predicting companies' esg ratings from news articles using multivariate timeseries analysis.

Florian Berg, Julian F Kölbel, and Roberto Rigobon. 2022. Aggregate Confusion: The Divergence of ESG Ratings*. *Review of Finance*, 26(6):1315–1344.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-Lingual ESG Impact Duration Inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Alessandro Del Vitto, Daniele Marazzina, and Davide Stocco. 2023. ESG ratings explainability through machine learning techniques. *Annals of Operations Research*, pages 1–30.

Shouxiang Fan, Junping Yao, Yangyang Sun, and Ying Zhan. 2020. A summary of aspect-based sentiment analysis. *Journal of Physics: Conference Series*, 1624(2):022051.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yan Cathy Hua, Paul Denny, Katerina Taskova, and Jörg Wicker. 2023. A systematic review of aspect-based sentiment analysis (absa): Domains, methods, and trends.

Ian T. Jolliffe. 2002. *Principal Component Analysis*, 2nd edition. Springer Series in Statistics. Springer New York, NY.

Tobias Schimanski, Andrin Reding, Nico Reding, Julia Bingler, Mathias Kraus, and Markus Leippold. 2024. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61:104979.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts.

United Nations. 2004. Who cares wins: Connecting financial markets to a changing world. Technical report, United Nations Global Compact.

Jetsons at FinNLP 2024: Towards Understanding the ESG Impact of a News Article using Transformer-based Models

Parag Pravin Dakle, Alolika Gon, Sihan Zha, Liang Wang, SaiKrishna Rallabandi, and Preethi Raghavan
Fidelity Investments, AI Center of Excellence
{firstname.lastname}@fmr.com

Abstract

In this paper, we describe the different approaches explored by the Jetsons team for the Multi-Lingual ESG Impact Duration Inference (ML-ESG-3) shared task. The shared task focuses on predicting the duration and type of the ESG impact of a news article. The shared task dataset consists of 2,059 news titles and articles in English, French, Korean, and Japanese languages. For the impact duration classification task, we fine-tuned XLM-RoBERTa with a custom fine-tuning strategy and using self-training and DeBERTa-v3 using only English translations. These models individually ranked first on the leaderboard for Korean and Japanese and in an ensemble for the English language, respectively. For the impact type classification task, our XLM-RoBERTa model fine-tuned using a custom fine-tuning strategy ranked first for the English language.

Keywords: ESG, Langauge Models, self-training, multi-lingual

1. Introduction

ESG (environment, social, and governance) related news can impact the performance and reputation of companies, investors, and regulators. One of the key challenges in ESG impact assessment is to estimate the duration of the ESG impact of a news article (Tseng et al., 2023). Different news articles may have different levels of salience, credibility, and relevance for different stakeholders and thus may have different effects on their behavior and outcomes. The LREC-COLING shared task (Chen et al., 2024) presents a multi-lingual impact duration and level classification task based on news articles.

We approach the shared task using the following strategies - (1) Traditional NLP techniques like TF-IDF with logistic regression, SVM (Cortes and Vapnik, 1995), and Random Forest classifiers, (2) De-noising the data to evaluate the impact of removing noisy or less informative samples, (3) Fine-tuning multilingual BERT-style models on individual language and entire dataset, (4) Complementing direct fine-tuning for impact duration with self-training using additional English and French ESG articles, (5) Translating all articles to English to simplify the impact duration task, and (6) Creating an ensemble of the best models for the impact duration task.

2. Related Work

There has been an increased focus on evaluating the nonfinancial activities of a company, which is typically encapsulated under the title of ESG. Park et al. (2022) show that the various topics included in ESG have gradually evolved. Mandas et al. (2023) perform a similar analysis across 11 sectors and show that the best ESG-performing financial institutions are actively committed to the code of best practices in governance. Language Modeling and NLP techniques have been the de facto approaches toward automating the estimation of ESG ratings.

Embeddings for ESG Classification: Raman et al. (2020) investigated employing embeddings from pre-trained language models for classifying sentences relevant to the ESG domain. Mehra et al. (2022) pre-trained a BERT model on ESG-related text, demonstrating improvements in classification tasks related to ESG factors.

Fine-tuning: Nugent et al. (2021) fine-tuned an English BERT-style model specifically for ESG document classification. They explored data generation as an augmentation strategy, enhancing model performance. (Jørgensen et al., 2021, 2023) extended the concept of pre-training language models from financial text to multilingual text and evaluated sentence classification and financial topic classification.

3. Data

The training dataset consists of 2,059 news articles in four languages: 545 English(en), 661 French(fr), 800 Korean(kr), and 53 Japanese(jp) articles. Each article has an associated title and the main content. News articles in all four languages are annotated with impact duration labels: less than 2 years, 2 to 5 years, and more than 5 years[1]. The distribution across the 3 impact duration classes is highly skewed, as shown in Figure 1a. The French and English articles are also annotated with 'low', 'medium',

[1]The label names are different for some languages

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 254–260
20 May, 2024.

or 'high' impact level classes (Figure 1c). The Korean dataset also contains impact type annotations with the following classes - opportunity, risk, cannot distinguish. This paper does not focus on the Korean impact type classification task.

Additionally, 31 and 24 duplicates were encountered while pre-processing the data for the Korean and French training data, respectively. We ignore all duplicates with the same class labels, but for 17 of the 24 French duplicates, we randomly select one of the duplicates to be part of the training dataset. This data is split into 10 parts for 10-fold cross-validation with consistent data distribution across all folds in the training and validation sets. The training and validation set lengths were about 1800 and 200, respectively. Lastly, we also found that the test set for Korean contains 1 duplicate, and for Japanese, it contains 19 duplicates and 327 samples with no class label.

4. Impact Duration Task

4.1. Traditional NLP Methods

4.1.1. Baseline Model

The small size of the dataset and high frequency of ESG keywords motivated us to evaluate naive TF-IDF classifier models as a traditional NLP baseline.

We consider logistic regression, random forest, and SVM as our baseline models and adopt 10-fold cross-validation for model training and evaluation. To enable the hyperparameters tuning for those baseline models, we further divide the 10-fold training set into train/val with a ratio of 80/20. Wangperawong (2022) show that using a single vocabulary for all languages and subword tokenization greatly improves the classification results. We use SentencePiece[2] for multilingual tokenization. We convert the obtained tokens to lowercase and compute TF-IDF statistics with filters of maximum frequency(0.7). We tune the penalty parameters C for SVM and logistic regression, and *number of trees, maximum depth parameter, and minimum sample of internal nodes* parameters for the Random Forest (RF) model. The averaged statistics in percentage from the 10-fold testing set are reported in Table 1. The RF model does a good job predicting the impact duration with large variation for the Japanese due to the smaller dataset.

4.1.2. Learning with De-noised Labels

Although the impact duration of the ESG news has been cross-validated with agreement statistics across different annotators, it is sometimes challenging to classify an ESG event into less than 2 years, 2 to 5 years, and more than 5 years window. For example, *The new agreements bring Verizon's projected renewable energy capacity to more than 3GW, enough to power more than 707,000 homes for a year and position the company to meet its goal to source or generate renewable energy equivalent to 50% of its total annual electricity consumption by 2025.* This article was annotated to be '2 to 5 years' probably due to the knowledge of the time difference between 2025 and the year of the annotation. However, the text clearly indicates a time window of one year, which could or should be annotated as "less than 2 years". The ground truth label of this event can hence be ambiguous. Brodley and Friedl (1999) demonstrated that direct training based on the "mislabeled" data generates less desirable models than training with less but de-noised data. Following a similar idea in Wang et al. (2023), we explored a data quality model to score each text-label pair.

Using the RF baseline model fine-tuned on TF-IDF tokens, we evaluate on each of the 10-fold testing sets to obtain the confidence of the prediction $\mathbb{P}$ and the label of the prediction $\hat{\mathbb{Y}}$. Then comparing against the annotation from the ground truth labels $\mathbb{Y}$, we compute a quality score $\mathbb{Q}$: $\mathbb{Y}\times\hat{\mathbb{Y}}\times\mathbb{P}\rightarrow[-1,1]$ using $\mathbb{Q}(\mathbb{P},\mathbb{Y},\hat{\mathbb{Y}}) = -\mathbb{P}$ if $\mathbb{Y}\neq\hat{\mathbb{Y}}$ and $\mathbb{Q}(\mathbb{P},\mathbb{Y},\hat{\mathbb{Y}}) = +\mathbb{P}$ if $\mathbb{Y} = \hat{\mathbb{Y}}$. Hence, a high-quality score $\mathbb{Q}$ would indicate agreement and high confidence between the predicted labels and the actual labels, whereas a low-quality score $\mathbb{Q}$ indicates agreement with high confidence. Computing on each of the 10-fold testing sets, we obtained the quality score $\mathbb{Q}$ for the entire 2,059 observations, based upon which we delete x% of the data that are potentially of low quality/agreement. Through our evaluation, we have found that deleting 10% of the original data provides a decent improvement with the weighted F1 score shown in Table 1. This indicates a certain level of noisiness within the duration labels.

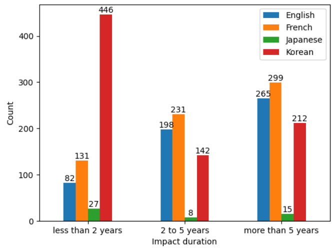
4.2. Modern NLP Methods

All models described in this subsection have been fine-tuned using 10-fold cross-validation, and the metric used for comparison is the average of weighted F1 scores across the folds. For the winning models, the fold model with the highest evaluation F1 score was further fine-tuned on the dataset for 2 additional epochs.

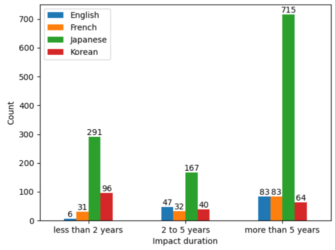
4.2.1. Fine-tuning Language Models (LMs)

We first fine-tune the XLM-RoBERTa (large) model (Liu et al., 2019) using both the title and the main content of the news articles in each language. We also consider the Longformer (large-4096) (Beltagy et al., 2020) model since some articles surpass

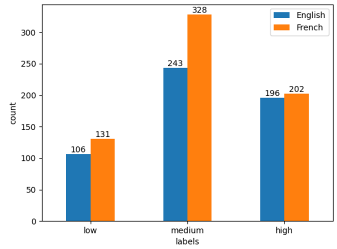
[2]https://github.com/google/sentencepiece



(a) Impact Duration training data



(b) Impact duration test data



(c) Impact level training data

Figure 1: Data distribution across classes

Model	**Setting**	**overall**	**kr**	**jp**	**en**	**fr**
Logistic	Normal	53.85	63.96	64.78	53.47	50.44
SVM	Normal	53.61	64.02	72.59	51.34	49.27
Random Forest (RF)	Normal	58.23	71.13	61.45	61.45	58.67
Logistic	De-noised	54.61	66.82	72.49	52.36	52.41
SVM	De-noised	55.96	68.42	73.13	54.98	52.79
Random Forest (RF)	De-noised	58.60	71.03	70.00	62.98	60.35

Table 1: Impact Duration: Weighted F1 averaged across 10-folds for baseline models

the maximum token length of conventional BERT-style models(Kannan and Seki, 2023). Given the small size of the datasets per language, specifically Japanese, we also fine-tune multilingual models by combining the news articles in all four languages. Table 2 shows the results of the fine-tuning experiments. We did not fine-tune a monolingual model for Japanese due to the small training data size.

Model-Lang	**PSL**	**en**	**fr**	**kr**	**jp**
XLM-en	_	57.9	_	_	_
XLM-fr	_	_	62.6	_	_
XLM-kr	_	_	_	**66.5**	_
LF-en	_	57.7	_	_	_
LF-fr	_	_	**72.9**	_	_
XLM-all	_	57.4	71	66.1	56.7
LF-all	_	50.2	57	39.8	39
XLM-all	direct	**60.2**	71.3	63.4	**59.9**
LF-all	direct	46.6	55.2	44	44.1
LF-all	avg. conf.	50.8	70.4	42.8	58.2

Table 2: Weighted F1 score for impact duration classification averaged across 10-folds for a) fine-tuned LMs (rows 1-7), and b) semi-supervised learning (rows 8-10). Note: **XLM**: XLM-RoBERTa, **LF**: Longformer, **PSL**: Pseudo-label generation methods, **XLM-kr**: Korean_Jetsons_1 submission

4.2.2. Semi-supervised Learning

The training dataset is small and skewed across the impact duration labels. For English and French, 45-48% of the articles belong to the 'more than 5 years' class. For Korean, 55% of the data belongs to the 'less than 2 years' class. To overcome this class imbalance, we use a subset of the news articles released as part of an ESG issues classification task (Chen et al., 2023). We use the XLM and LF models in Table 2 as teacher models and make predictions on the English and French news articles in the ESG issues classification dataset. We generate pseudo labels in two ways: a) **direct**: use the label predicted by the multilingual teacher model directly, b) **avg. conf.**: for each article, take the average of the two confidence scores for each class predicted by the multi-lingual and mono-lingual teacher models and choose the label with the maximum average score. We sample articles based on these pseudo labels and combine them with the original training data to reduce class imbalance. This augmented data is used to fine-tune XLM and LF models. The weighted F1 scores for these models are reported in Table 2. The F1 scores of these models on the final test are reported in section A.2.

4.2.3. English Translation

We also consider converting the problem from multi-lingual to mono-lingual by adding translation as a prerequisite for training and testing. We use the Google Translate API[3] to translate all non-English samples to English. Post translation, we fine-tune a DeBERTa-v3-xsmall model (He et al., 2023) on the class labels using both the article text and title (if available). The model experiment reports a 10-

[3]https://translate.google.com/

fold average weighted F1 score of 62.37 and a maximum weighted F1 of 66.82 on fold 8. The fold 8 model (DBERT-en) was used in the second submission (<lang>_Jetsons_2) for all languages.

4.2.4. Ensemble

The final model for the impact duration classification task is an ensemble model. We consider an ensemble of the three models - XLM, LF, and the DeBERTa-v3. The class label with the highest total model label score sum is used as the final class label. The submitted ensemble models were - **English_Jetsons_3** (XLM-all-direct, LF-all-avg. conf, DeBERTa-v3), **French_Jetsons_3** (XLM-all, LF-fr, DeBERTa-v3), **Korean_Jetsons_3** (XLM-kr, XLM-all, DeBERTa-v3), and **Japanese_Jetsons_3** (XLM-all-direct, DeBERTa-v3).

5. Impact Level

We conduct experiments with the same two multilingual language models - XLM-RoBERTa-large and Longformer-large-4096 for the impact level task in French and English. We fine-tune the multilingual models in two ways: a) Using both languages, hoping that the data in one language can bolster the performance in the other, and b) separately in each language. First we compare models using only fold 0 data. Table 3 shows that the weighted F1-score for the model trained in combined languages is lower than single language. So we use data in single language to further fine-tune the two models using all 10 folds of data and calculate the average results for each language. It shows that the XLM model has better performance in both languages: XLM-en (**65.02**) vs. LF-en (59.27), and XLM-fr (**65.29**) vs. LF-fr (63.84). We pick the XLM models with the best performance among 10 folds for each language as our first submission. As our second submission, we randomly chose a fold and used the best model fine-tuned on that fold.

La-	XLM			LF		
ng	**en**	**fr**	**all**	**en**	**fr**	**all**
en	57.3	_	51.7	56.1	_	47.6
fr	_	71.5	69	_	72.9	59.4

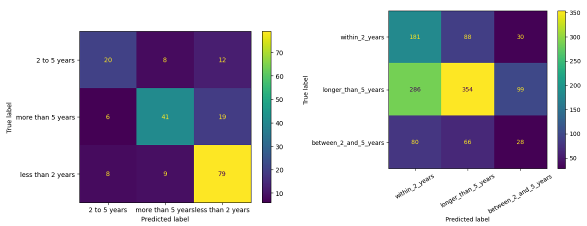
Table 3: Weighted F1 score for impact level classification over the data in the 0^{th} fold

6. Analysis

Table 4 shows the best micro and macro F1 scores on the test set for the submitted models. These models ranked best on 4 out of 7 tasks. Figures 2 and 3 show the confusion matrices for predictions generated using these four predictions. For impact duration, the models get most confused between 'less than 2 year' and 'More than 5 years' classes. For impact level, 'medium' and 'high' are the most confusing classes.

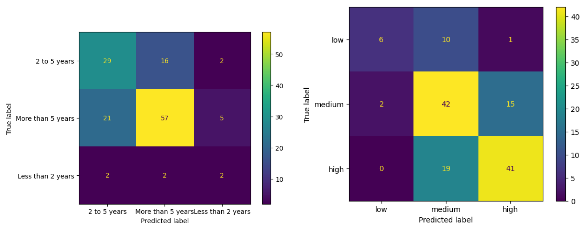
Submission Model	Micro F1	Macro F1
English_Jetsons_1	64.71	**52.47**
Korean_Jetsons_1	**70**	**66.24**
Japanese_Jetsons_2	36.5	**25.6**
French_Jetsons_1	47.95	37.06
English_Jetsons_1 (IL)	**65.44**	**60.90**

Table 4: F1 scores on the test set, **Bold faced** ones are top on the leaderboard. IL indicates impact level



(a) Korean_Jetsons_1 (b) Japanese_Jetsons_2

Figure 2: ID confusion matrices



(a) English_Jetsons_3 (b) English_Jetsons_1

Figure 3: Confusion matrices for: ID (a) and IL (b).

7. Conclusion

ESG is increasingly important for stakeholders who want to align their values with their investments, reduce risks, and enhance long-term returns. For the FinNLP shared task of impact duration and level classification, we find that finetuning BERT-style models, along with data augmentation techniques like translation and self-training, perform the best. For impact duration in Korean and impact level in English, we find that fine-tuning a BERT-based classifier with a custom strategy performs the best. An ensemble with BERT-style models fine-tuned for impact duration in English using self-training and on just English translations performs best. The DeBERTa-v3 model fine-tuned on only English translations performs best on the Japanese dataset for the impact duration task.

8. Bibliographical References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20:273–297.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. Multifin: A dataset for multilingual financial nlp. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 864–879.

Rasmus Kær Jørgensen, Mareike Hartmann, Xiang Dai, and Desmond Elliott. 2021. mdapt: Multilingual domain adaptive pretraining in a single model. *arXiv preprint arXiv:2109.06605*.

Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for ESG scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 45–54, Macao. -.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Marco Mandas, Oumaima Lahmar, Luca Piras, and Riccardo De Lisa. 2023. Esg in the financial industry: What matters for rating analysts? *Research in International Business and Finance*, 66:102045.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

Tim Nugent, Nicole Stelea, and Jochen L Leidner. 2021. Detecting environmental, social and governance (esg) topics using domain-specific language models and data augmentation. In *Flexible Query Answering Systems: 14th International Conference, FQAS 2021, Bratislava, Slovakia, September 19–24, 2021, Proceedings 14*, pages 157–169. Springer.

Joonbeom Park, Woojoo Choi, and Sang-Uk Jung. 2022. Exploring trends in environmental, social, and governance themes and their sentimental value over time. *Frontiers in Psychology*, 13:890435.

Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping esg trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction*, 2(4):453–468.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 5412–5416.

Liang Wang, Ivano Lauriola, and Alessandro Moschitti. 2023. Accurate training of web-based question answering systems with feedback from ranked users. In *ACL 2023*.

Artit Wangperawong. 2022. Multilingual search with subword tf-idf. *arXiv preprint arXiv:2209.14281*.

A. Additional experiment details

A.1. Traditional NLP methods

The weighted F1 score across different models and different languages is summarized in Figure 5.

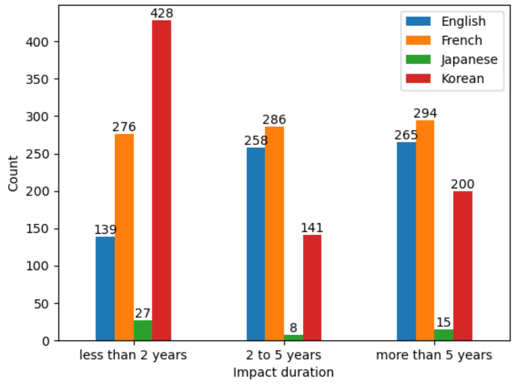


Figure 4: Data distribution across classes in the training set used in semi-supervised learning

A.2. Modern NLP Methods

The distribution of data across class labels in the dataset used to train the student models using semi-supervised learning is shown in figure 4. The micro and macro F1 scores achieved by the different fine-tuned language models on the test set are reported in table 5. In the case of the English, Korean, and Japanese data, we see that the models with the best 10-fold cross-validation scores also perform similarly on the test set. However, for the French news articles, while fine-tuning the Longformer model using only French data (LF-fr) gives maximum average weighted F1 during cross-validation, the same isn't reflected on the test set. XLM-Roberta fine-tuned on articles in all languages along with self-training (row 8 in table 5) gives the best macro F1 of **50.54** and micro F1 of **53.42**. The scores on the Japanese test data have been calculated after removing the 327 unlabelled news articles.

B. Hyperparameters

For the XLM-RoBERTa and Longformer fine-tuning experiments, the learning rates for the mono-lingual and multi-lingual models were $2e-5$ and $8e-6$, respectively, along with batch size -8 and epochs -10. The Longformer-large models were fine-tuned with gradient accumulation of 2 steps. For the DeBERTa-v3-xsmall model, the following hyperparameters were: learning rate - $2e-05$, epochs - 10, weight decay 0.01, and batch size - 2. The fine-tuning process was carried out on a GPU with 32 GB memory.

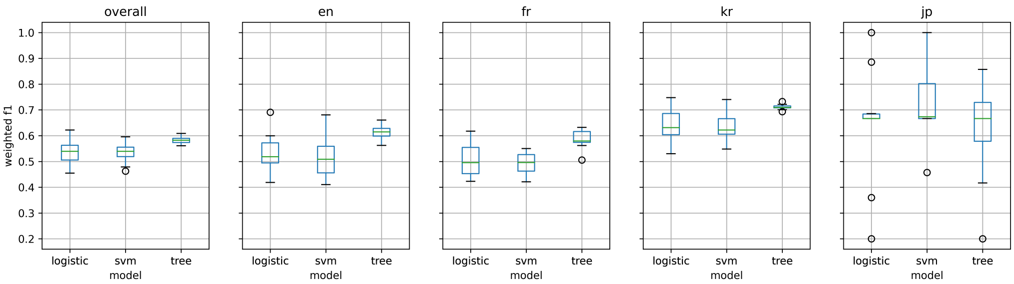


Figure 5: Weighted F1 scores across 10-folds for traditional NLP models for predicting impact length

Model-lang	**PSL**	**en**		**fr**		**kr**		**jp**	
		Mi. F1	Ma. F1	Mi. F1	Ma. F1	Mi. F1	Ma. F1	Mi. F1	Ma. F1
XLM-en		**61.03**	**48.79**						
XLM-fr				54.79	43.01				
XLM-kr						**70.0**	**66.24**		
LF-en		55.88	44.9						
LF-fr				47.95	37.06				
XLM-all		58.82	43.66	47.95	43.28	64.0	57.09	**38.19**	**32.99**
LF-all		59.56	38.3	43.84	36.86	48	21.62	24.81	13.25
XLM-all	direct	61.03	46.7	**53.42**	**50.54**	67	62.64	36.23	31.97
LF-all	direct	56.62	47.11	42.47	35.87	50.5	43.62	24.81	13.25
LF-all	avg. conf.	58.82	45.16	38.36	34.34	47.5	27.36	28.3	18.26

Table 5: Micro and Macro F1 for impact length classification task in the final test set.

ESG Classification by Implicit Rule Learning via GPT-4

Hyo Jeong Yun[1], Chanyoung Kim[2], Moonjeong Hahm[1], Kyuri Kim[3], Guijin Son[4*]
Chung-ang University[1], Konkuk University[2], Seoul Women's University[3], Yonsei University[4]
dbsgywjd@cau.ac.kr, spthsrbwls123@yonsei.ac.kr

Abstract

Environmental, social, and governance (ESG) factors are widely adopted as higher investment return indicators. Accordingly, ongoing efforts are being made to automate ESG evaluation with language models to extract signals from massive web text easily. However, recent approaches suffer from a lack of training data, as rating agencies keep their evaluation metrics confidential. This paper investigates whether state-of-the-art language models like GPT-4 can be guided to align with unknown ESG evaluation criteria through strategies such as prompting, chain-of-thought reasoning, and dynamic in-context learning. We demonstrate the efficacy of these approaches by ranking 2nd in the Shared-Task ML-ESG-3 *Impact Type* track for Korean without updating the model on the provided training data. We also explore how adjusting prompts impacts the ability of language models to address financial tasks leveraging smaller models with openly available weights. We observe longer general pre-training to correlate with enhanced performance in financial downstream tasks. Our findings showcase the potential of language models to navigate complex, subjective evaluation guidelines despite lacking explicit training examples, revealing opportunities for training-free solutions for financial downstream tasks.

Keywords: Large Language Model, Benchmark,Finance

1. Introduction

In recent years, there has been a noticeable increase in investors factoring environmental, social, and governance (ESG) considerations into their investment choices. Recent studies, through meta-analysis, have shown that improved ESG performance correlates with better corporate financial outcomes, potentially leading to higher investment returns (Cort and Esty, 2020; Friede et al., 2015). Assessing ESG performance involves nuanced analysis, and, as a result, the industry relies on rating agencies like MSCI[1], Sustainalytics[2], and Bloomberg[3] to evaluate and rank companies. Ongoing efforts to automate the ESG evaluation process exist, mainly through leveraging language models as substitutes for human analysts (Mehra et al., 2022). However, the specific methodologies used by each rating agency are not widely disclosed, leading to a lack of understanding of the detailed metrics necessary for evaluation. The closed nature of these agencies presents significant challenges when training language models to accurately replicate their evaluation criteria. This is particularly problematic for earlier language models, such as BERT (Devlin et al., 2018), which heavily rely on explicit training data on the output distribution to accurately approximate the underlying function. Without access to the specific criteria and data used by these agencies, it is difficult to teach language models to make judgments that align with past standards. Researchers have sought to enhance training datasets through synthetic data to address this issue (Glenn et al., 2023). Nonetheless, several hurdles exist. First, the lack of transparency in the evaluation methodologies used by rating agencies, which often include subjective assessments, makes it difficult for researchers to generate realistic datasets. Moreover, the creation of large-scale, high-quality labeled datasets is resource-intensive. Manually annotating extensive text collections requires considerable time and skilled professionals. Furthermore, the accurate classification of sentences poses challenges due to the subjective nature of interpretation, which can vary even among experts (Auzepy et al., 2023). Finally, the rapid evolution of ESG criteria requires regular updates on the training dataset and retraining the model to align with changing investor expectations, emerging trends, and new reporting standards.

In this paper, we investigate whether state-of-the-art language models can be guided to align with unknown values (specifically, ESG evaluation standards) without learning from explicit training data. We employ multiple strategies, such as prompting, Chain-of-Thought reasoning (Wei et al., 2022), and dynamic in-context learning (Dong et al., 2022) with *GPT-4* (OpenAI), to participate in the Shared-Task ML-ESG-3 and rank second place in the *Impact Type* track for Korean. Our findings underscore the efficacy of these strategies in approximating unknown guidelines, showcasing their potential in navigating the complexities of ESG criteria alignment. Furthermore, we extend our investigation

* Corresponding author.

[1] https://www.msci.com

[2] https://www.sustainalytics.com

[3] https://www.bloomberg.com

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 261–268
20 May, 2024.

[...] 스카버러 유전에서 25년간 이산화탄소 3억7000만톤이 배출될 것이라고 예측한 것. 이는 호주에서 연간 발생하는 이산화탄소 배출량보다 3배나 많은 양이다. 또 재단은 호주 신임 연방 환경장관인 타냐 플리버섹에게, 이번 프로젝트가 기후 변화를 악화시켜 그레이트 배리어 리프에 악영향을 미칠지 여부를 판단할 수 있을 때까지 사업 시작을 중단시켜 줄 것을 요구했다. 호주보존재단의 켈리 오샤나시 대표는 "스카버러 가스 프로젝트는 국가 환경법의 허점을 파고들어 승인을 따낸 사업"이라며, [...] , 호주에서는 노후화된 지역 석탄화력발전소에서도 정전이 발생한 바 있다. 이에 최근 선거에서 승리한 캔버라의 중도 좌파 정부는 "환경을 우선 생각해야 하지만 상업적인 부분도 무시할 수 없다"며 지속적으로 화석 연료 프로젝트를 지원할 것이라고 밝혔다.

[...] It is predicted that the Scarborough gas field will emit 370 million tons of CO2 over 25 years, three times more than Australia's annual emissions. The Australian Conservation Foundation has asked Environment Minister Tanya Plibersek to delay the project's start until its impact on climate change and the Great Barrier Reef can be assessed, criticizing the project for exploiting legal loopholes to gain approval. [...] ,In Australia, power outages have occurred at aging coal-fired power plants. The centrist-left government in Canberra, after winning the recent election, stated it would prioritize the environment but cannot ignore commercial aspects, pledging continued support for fossil fuel projects.

Figure 1: An example from the ML-ESG dataset. Sentences highlighted in red indicate negative implications for ESG, while those in blue denote positive ESG implications. The gold label for the ESG type of this text is "Opportunity." English translations are added for broader accessibility.

Category	Opp.	Risk	Cannot Dist.	Total.
Sustainable Econ.	160	57	41	258
Corporate Govern.	134	31	40	205
Env. & Society	71	79	6	156
Disclosure & Eval.	87	55	11	153
ESG Life	7	3	10	20
Opinion	3	4	1	8
Total	462	229	109	800

Table 1: Statistics on the *Impact Type* of Shared-Task ML-ESG-3 for Korean.

Category	< 2 Yrs	2-5 Yrs	> 5 Yrs	Total
Sustainable Econ.	101	54	103	258
Corporate Govern.	137	36	32	205
Env. & Society	67	26	63	156
Disclosure & Eval.	119	23	11	153
ESG Life	16	1	3	20
Opinion	6	2	0	8
Total	446	212	142	800

Table 2: Statistics on the *Impact Duration* of Shared-Task ML-ESG-3 for Korean.

to include two smaller models with publicly accessible weights, examining how slight modifications in prompts influence their performance and calibration. To the best of our knowledge, this study represents the first attempt to explore how adjustments in prompts can impact the ability of language models to address financial problems.

2. Shared Task ML-ESG-3

The Shared-Task ML-ESG-3 for Korean consists of two downstream tasks: *Impact Type* and *Impact Duration*. The *Impact Type* task involves classifying given ESG news articles to one of *Opportunity*, *Risk*, or *Cannot Distinguish*. The *Impact Duration* task involves classifying the impact duration of a news article as one of *Less than 2 years*, *2 to 5 years*, or *More than 5 years*. The dataset includes separate training and testing sets, with 800 Korean articles in the training set and 200 articles in the testing set.

In Table 1 we illustrate the distribution of impact types across categories in the training dataset. We observe significant data imbalance across multiple columns. For instance, while the largest category, "Sustainable Economics" feature 258 samples, the smallest category "Opinions," only include eight.

Furthermore, *Opportunity* category comprises 462 entries, roughly four times the count of the *Cannot Distinguish* category, which has 109 entries. The imbalance of data could potentially be attributed to either: 1) a sampling error arising from the small dataset size, or 2) the real-world distribution of ESG-related news being skewed, as press may be more reluctant to report negative issues due to associated risks. Regardless of the underlying cause, this imbalanced training set poses a critical challenge for traditional approaches to training language models, as they will inevitably learn skewed representations from the biased data distribution. Similar patterns can be found also for the *Impact Duration* subset as shown in Table 2. The *Less than 2 years* category is the largest with 446 entries, nearly three times more than the *More than 5 years* category, which is the least represented with 142 entries.

3. Main Results

In this section, we elaborate on our methodology(Section 3.1) and report observed performances (Section 3.2).

You will be given a text. Refer to the examples and the MSCI guideline for your decision. Classify it to either [cannot distinguish/risk/opportunity] based on the impact it will have on the company.

text: {exemplar1}
response: Based on the MSCI guideline the answer is {gold1}.

text: {text}
response:Based on the MSCI guideline the answer is [cannot distinguish/risk/opportunity].

Figure 2: An example prompt with one examplar (highlighted in red) and prompts to follow the MSCI guidelines (highlghted in blue). We calculate the chance for the gold answer to follow "the answer is".

3.1. Methodology

Predicting the ESG types and their impact duration from texts is a non-trivial task that traditionally relies on human experts. However, the criteria these experts use are mostly kept confidential. This ambiguity fence researchers from developing precise rules for LLMs to learn to perform such tasks. Accordingly, this leads to a question: **Can LLMs implicitly approximate unknown rules, without a comprehensive understanding of the task?** To address this question, we employ *GPT-4*, a state-of-the-art language model. To align the model with the implicit rules we leverage the following approaches:

In-Context Learning (Dong et al., 2022): In-context learning (ICL) is an approach where LLMs are provided with exemplars demonstrating the desired behavior. Instead of updating parameters through backpropagation, the model infers patterns from the examples and generalizes during inference. In our work, we dynamically alter the provided examples using the BM-25 algorithm. For a given input sample, we retrieve five relevant articles from the training set and provide them for ICL to the model during inference.

Chain-of-Thought (Wei et al., 2022):Chain-of-thought guide models to generate a series of intermediate reasoning steps while solving a task. In an autoregressive structure, one forward pass is calculated per generated token; accordingly, allowing a model to generate intermediate reasoning allows it to leverage more forward passes as needed.

Prompt Engineering (White et al., 2023):Prompt engineering involves creating prompts or prefixed to guide LLMs during inference. A prompt engineers the LLM to follow a desired behavior and output format. In this work, we prompt the language model to follow the MSCI guidelines for classification.

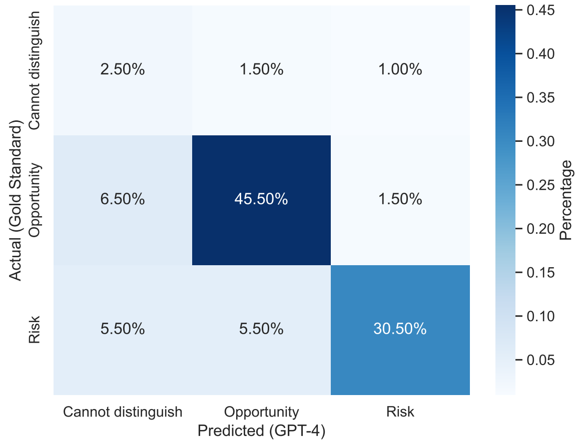


Figure 3: A confusion matrix analyzing the performance of *GPT-4* on the *Impact Type* subset.

Figure 4: A confusion matrix analyzing the performance of *GPT-4* on the *Impact Duration* subset.

3.2. Evaluation Results

Table 4 showcases the performance of selected models on the Korean subset for the Shared Task ML-ESG-3. Notably, our approach, which utilizes 5-shot exemplars and prompt engineering based on MSCI guidelines, ranks second in *Impact Type* classification. However, it falls short in accurately predicting *Impact Duration*. An initial analysis of the outputs, presented in Figures 3 and 4, reveals a tendency of *GPT-4* to incorrectly classify impact durations as *less than 2 years*. Further qualitative examination shows that articles containing multiple perspectives and events often mislead the model. This observation is consistent with findings that LLMs struggle with comprehending and referencing longer text inputs (Levy et al., 2024).

Task	Model	Min	Max	Mean	△ (Max - Min)
Impact Duration	EEVE-Korean-10.8B	38.0	48.5	44.9	10.5
Impact Type	EEVE-Korean-10.8B	35.0	55.5	48.9	20.5
Impact Duration	Yi-Ko-6B	44.0	51.5	47.9	7.5
Impact Type	Yi-Ko-6B	59.0	65.5	63.2	6.5

Table 3: Performance summary of *Yi-Ko-6B* and *EEVE-Korean-10.8B* with ten different prompts. We report the accuracy (%) of each models.

Submission	Impact Type	Impact Duration
Ours	<u>76.13</u>	43.98
3idiots_3	**79.85**	<u>61.54</u>
Jetsons_1	-	**66.24**
Tredence_2	75.95	58.18

Table 4: Performance of selected models. The highest-scoring model is highlighted in **bold**, and second-highest is <u>underlined</u>.

An example highlighting an instance with multiple implications is provided in Figure 1. Despite the challenges, SOTA LLMs like GPT-4 demonstrate a remarkable ability to implicitly identify patterns, surpassing traditional performance methods without requiring specific training.

4. Calibration

For a model's decisions to be considered trustworthy, they must be well-calibrated; this means that its confidence levels should accurately reflect the true likelihood of its predictions being correct. In this section, we will explore how various approaches influence models' calibration and accuracy.

4.1. Experimental Settings

Models Unfortunately, the *GPT-4* API does not provide enough information for the intended analysis. Therefore, we choose to use *Yi-Ko-6B* (Lee) and *EEVE-Korean-10.8B* (Kim et al., 2024) two pre-trained models with fewer than 14 billion parameters that demonstrate the highest performance on the KMMLU (Son et al., 2024) benchmark. See Appendix A for further details on the models.

Evaluation We evaluate ten distinct approaches, varying the number of in-context exemplars, the order of these exemplars, and the prompts themselves. See Appendix A for an explanation of each approach. For each approach, we append "The answer is" to a query and calculate the likelihood of each option following the query. Figure 2 provides an example of the query format.

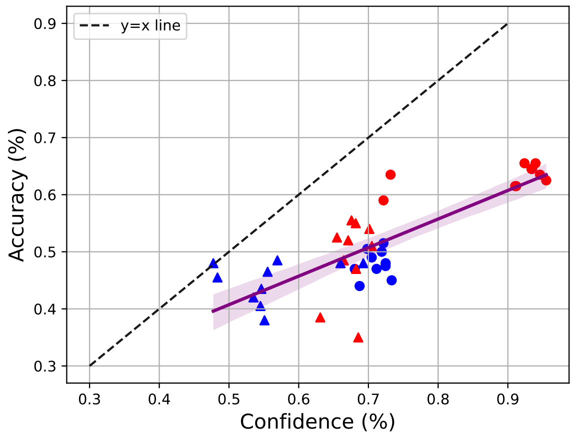


Figure 5: Relationship between accuracy and confidence of Yi-Ko-6B (circle) and EEVE-Korean-10.8B (triangle) for both subsets.(Red for 'Impact Type' and blue for 'Impact Duration'). Regression analysis exhibits a slope of 0.50.

4.2. Analysis

In Figure 5, we provide an overview of the calibration of models by testing how well the average confidence estimates the accuracy for each prompt. Surprisingly, both model appears to be well-calibrated, with a regression analysis exhibiting a slope of 0.5. In Table 3, we observe that *Yi-Ko-6B* outperforms *EEVE-Korean-10.8B* in both average and maximum scores. Additionally, *Yi-Ko-6B* exhibits a smaller delta, indicating greater robustness to prompt variations. This increased robustness may stem from extended continual pre-training, which is consistent with recent studies suggesting that the ICL capabilities of models are enhanced by encountering parallel structures in the training corpora (Chen et al., 2024b). Extended continual pre-training in Korean likely increases the model's exposure to parallel structures, thus improving its ability to capture implicit patterns robustly. Our analysis indicates that smaller, publicly available models can also effectively identify implicit patterns in ESG classification without prior training. Without needing task-specific fine-tuning, general pre-training seems to improve their robustness and overall performance.

5. Conclusion

In this work, we adopt multiple prompting, chain-of-thought reasoning, and in-context learning strategies to guide *GPT-4* in solving ESG classification tasks. We rank second in the Korean subset for Shared Task ML-ESG-3 in *Impact Type* prediction. Furthermore, we adopt open models to explain their calibration and robustness to different prompting strategies. The longer general pre-training correlates with enhanced performance in financial downstream tasks. While our work has been limited to the Korean language, we believe it will be equally applicable in different languages, especially in English, and leave for future works.

6. Bibliographical References

01-ai. 01-ai/yi-6b. `https://huggingface.co/01-ai/Yi-6B`. Accessed: 2024-03-08.

Alix Auzepy, Elena Tönjes, David Lenz, and Christoph Funk. 2023. Evaluating tcfd reporting: A new application of zero-shot analysis to climate-related financial disclosures. *arXiv preprint arXiv:2302.00326*.

Chang Heng Chen, Xiting Wang, Ting-En Lin, Ang Lv, Yuchuan Wu, Xin Gao, Ji-Rong Wen, Rui Yan, and Yongbin Li. 2024a. Masked thought: Simply masking partial reasoning steps can improve mathematical reasoning learning of language models.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2024b. Parallel structures in pre-training data yield in-context learning. *arXiv preprint arXiv:2402.12530*.

Todd Cort and Daniel Esty. 2020. Esg standards: Looming challenges and pathways forward. *Organization & Environment*, 33(4):491–510.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Gunnar Friede, Timo Busch, and Alexander Bassen. 2015. Esg and financial performance: aggregated evidence from more than 2000 empirical studies. *Journal of sustainable finance & investment*, 5(4):210–233.

Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using synthetic data and transfer learning for multilingual ESG issue classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 133–139, Macao. -.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024. Efficient and effective vocabulary expansion towards multilingual large language models. *arXiv preprint arXiv:2402.14714*.

Junbum Lee. beomi/yi-ko-6b. `https://huggingface.co/beomi/Yi-Ko-6B`. Accessed: 2024-03-08.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2024. Enhancing esg impact type identification through early fusion and multilingual models. *arXiv preprint arXiv:2402.10772*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C

Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

A. Additional details for Section 4

A.1. Adopted Models

We adopt the following models with openly-available weights for analysis in Section 4. Due to hardware limitations, all models are used in 4-bit quantization.

1. *EEVE-Korean-10.8B* (Kim et al., 2024): A Korean vocabulary-extended ver sion of *SOLAR-10.7B* (Kim et al., 2023) that has undergone continual pre-training on a total of 3.2M documents (or, 3.2B tokens).

2. *Yi-Ko-6B* (Lee): A Korean vocabulary-extended version of Yi-6B (01-ai) that has undergone continual pre-training on 60B tokens.

A.2. Prompts

In Table 5, we provide an overview of the ten prompts used for analysis in Section 4.

Prompt Name	# of In-Context Exemplars	Order of Exemplars	Prompted to follow MSCI Guidelines
1-shot-standard_order-msci	1	Similar First	O
1-shot-standard_order-standard	1	Similar First	X
3-shot-reverse_order-msci	3	Similar Last	O
3-shot-reverse_order-standard	3	Similar Last	X
3-shot-standard_order-msci	3	Similar First	O
3-shot-standard_order-standard	3	Similar First	X
5-shot-reverse_order-msci	5	Similar Last	O
5-shot-reverse_order-standard	5	Similar Last	X
5-shot-standard_order-msci	5	Similar First	O
5-shot-standard_order-standard	5	Similar First	X

Table 5: Entire list of prompt settins used in Section 4.

A.3. Performance Details

In Tables 6 and 6 we present the detailed per prompt perfomrnace for each models.

Prompt	Accuracy	Confidence	Model	Task
1-shot-standard_order-msci_simple	0.635	0.731760	Yi-Ko-6B	Impact Type
1-shot-standard_order-standard	0.590	0.721608	Yi-Ko-6B	Impact Type
3-shot-reverse_order-msci_simple	0.625	0.955045	Yi-Ko-6B	Impact Type
3-shot-reverse_order-standard	0.635	0.946185	Yi-Ko-6B	Impact Type
3-shot-standard_order-msci_simple	0.645	0.933864	Yi-Ko-6B	Impact Type
3-shot-standard_order-standard	0.655	0.923851	Yi-Ko-6B	Impact Type
5-shot-reverse_order-msci_simple	0.645	0.934855	Yi-Ko-6B	Impact Type
5-shot-reverse_order-standard	0.655	0.939728	Yi-Ko-6B	Impact Type
5-shot-standard_order-msci_simple	0.615	0.910514	Yi-Ko-6B	Impact Type
5-shot-standard_order-standard	0.615	0.912037	Yi-Ko-6B	Impact Type
1-shot-standard_order-msci	0.505	0.698373	Yi-Ko-6B	Impact Duration
1-shot-standard_order-standard	0.500	0.719090	Yi-Ko-6B	Impact Duration
3-shot-reverse_order-msci	0.470	0.680418	Yi-Ko-6B	Impact Duration
3-shot-reverse_order-standard	0.490	0.704762	Yi-Ko-6B	Impact Duration
3-shot-standard_order-msci	0.475	0.724632	Yi-Ko-6B	Impact Duration
3-shot-standard_order-standard	0.515	0.721509	Yi-Ko-6B	Impact Duration
5-shot-reverse_order-msci	0.440	0.687383	Yi-Ko-6B	Impact Duration
5-shot-reverse_order-standard	0.470	0.711635	Yi-Ko-6B	Impact Duration
5-shot-standard_order-msci	0.450	0.733333	Yi-Ko-6B	Impact Duration
5-shot-standard_order-standard	0.480	0.724686	Yi-Ko-6B	Impact Duration

Table 6: Detailed performance of *Yi-Ko-6B* on different prompts.

Prompt	Accuracy	Confidence	Model	Task
1-shot-standard_order-msci_simple	0.35	0.685465	EEVE-Korean-10.8B	Impact Type
1-shot-standard_order-standard	0.385	0.630959	EEVE-Korean-10.8B	Impact Type
3-shot-reverse_order-msci_simple	0.525	0.654941	EEVE-Korean-10.8B	Impact Type
3-shot-reverse_order-standard	0.54	0.701319	EEVE-Korean-10.8B	Impact Type
3-shot-standard_order-msci_simple	0.485	0.664646	EEVE-Korean-10.8B	Impact Type
3-shot-standard_order-standard	0.55	0.681784	EEVE-Korean-10.8B	Impact Type
5-shot-reverse_order-msci_simple	0.51	0.704919	EEVE-Korean-10.8B	Impact Type
5-shot-reverse_order-standard	0.555	0.675689	EEVE-Korean-10.8B	Impact Type
5-shot-standard_order-msci_simple	0.47	0.682284	EEVE-Korean-10.8B	Impact Type
5-shot-standard_order-standard	0.52	0.670969	EEVE-Korean-10.8B	Impact Type
1-shot-standard_order-msci	0.48	0.659873	EEVE-Korean-10.8B	Impact Duration
1-shot-standard_order-standard	0.48	0.692712	EEVE-Korean-10.8B	Impact Duration
3-shot-reverse_order-msci	0.435	0.546392	EEVE-Korean-10.8B	Impact Duration
3-shot-reverse_order-standard	0.465	0.555405	EEVE-Korean-10.8B	Impact Duration
3-shot-standard_order-msci	0.42	0.535136	EEVE-Korean-10.8B	Impact Duration
3-shot-standard_order-standard	0.485	0.569464	EEVE-Korean-10.8B	Impact Duration
5-shot-reverse_order-msci	0.405	0.545175	EEVE-Korean-10.8B	Impact Duration
5-shot-reverse_order-standard	0.48	0.477536	EEVE-Korean-10.8B	Impact Duration
5-shot-standard_order-msci	0.38	0.55096	EEVE-Korean-10.8B	Impact Duration
5-shot-standard_order-standard	0.455	0.483521	EEVE-Korean-10.8B	Impact Duration

Table 7: Detailed performance of *EEVE-Korean-10.8B* on different prompts.

Leveraging Semi-Supervised Learning on a Financial-Specialized Pre-trained Language Model for Multilingual ESG Impact Duration and Type Classification

Jungdae Kim, Eunkwang Jeon, Sanghyun Jeon
KakaoBank Corp.
{jd.kim, weezy.j, ali.jeon}@kakaobank.com

Abstract

This paper presents the results of our participation in the Multilingual ESG Impact Duration Inference (ML-ESG-3) shared task organized by FinNLP-KDF@LREC-COLING-2024. The objective of this challenge is to leverage natural language processing (NLP) techniques to identify the impact duration or impact type of events that may affect a company based on news articles written in various languages. Our approach employs semi-supervised learning methods on a finance-specialized pre-trained language model. Our methodology demonstrates strong performance, achieving 1st place in the Korean - Impact Type subtask and 2nd place in the Korean - Impact Duration subtask. These results showcase the efficacy of our approach in detecting ESG-related issues from news articles. Our research shows the potential to improve existing ESG ratings by quickly reflecting the latest events of companies.

Keywords: ESG, ESG Rating, NLP, SSL

1. Introduction

The importance of Environmental, Social, and Governance (ESG) factors in the investment decision-making process has been increasingly emphasized. ESG factors have emerged as key considerations for corporate sustainability and long-term success, leading to the proposal of various frameworks and approaches to evaluate and quantify companies ESG-related activities. However, existing ESG evaluation methods primarily rely on fixed materials such as annual reports, limiting their ability to promptly reflect the dynamic changes in the market. In this context, an approach has been proposed to infer the impact of the latest events and news articles on companies ESG ratings (Tseng et al., 2023; Kannan and Seki, 2023). Tseng et al. (2023) introduced a new dataset that can identify the ESG impact type and impact duration of corporate events using ESG-related news articles. This dataset has become an important foundation for the Multi-Lingual ESG Impact Duration Inference (ML-ESG-3) shared task proposed at FinNLP-KDF@LREC-COLING-2024. The goal of the ML-ESG-3 shared task is to identify the impact duration or impact type of events that may affect companies using natural language processing (NLP) techniques on news articles written in various languages.

To achieve this goal, we utilized a finance-specialized pre-trained language model and applied semi-supervised learning (SSL) methods using unlabeled data collected through web crawling. This approach achieved 1st and 2nd place in the Korean impact type and impact duration identification tasks, respectively. As part of the research exploring the modernization and dynamic update possibilities of ESG evaluation, this paper presents an NLP-based methodology that can improve ESG evaluation by promptly reflecting the latest corporate events. This is expected to enable investors to make investment decisions considering ESG factors based on more accurate and timely information.

2. Dataset

The Korean task consists of two sub-tasks: Impact Type Identification and Impact Duration Inference. The datasets for these sub-tasks were annotated following the methodology proposed by Tseng et al. (2023).

Impact Type identification is a single-choice question that aims to determine the type of impact a news article might have on a company. The possible labels are "opportunity", "risk", and "cannot distinguish". The "opportunity" label indicates that the news article discusses a potential positive impact or benefit to the company, while the "risk" label suggests that the article highlights a potential negative impact or threat. The "cannot distinguish" label is assigned when the impact type is unclear.

Impact Duration inference is a single-choice question that seeks to determine the duration of the impact a news article might have on a company. Based on the distinction between short-term and long-term, three labels are presented: “less than 2 years”, “2 to 5 years”, and “more than 5 years”. These labels provide a temporal context for the impact, allowing for a better understanding of the Immediate and long-term implications of the news content on the company.

The news articles In the dataset vary In length, with an average of 733 characters per article. The shortest article has 173 characters, while the longest article has 1,768 characters. This variation in article length presents a challenge for the models, as they need to effectively understand texts of different sizes.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 269–273
20 May, 2024.

To provide a clear understanding of the dataset composition, Tables 1 and 2 show the distribution of labels for the Impact Type and Impact Duration sub-tasks, respectively, within the training set. For both sub-tasks, the dataset provides a train set containing 800 examples and a test set with 200 examples.

Labels	Count
opportunity	462
risk	229
cannot distinguish	109
Total	**800**

Table 1: Label counts in Korean – Impact Type train set.

Labels	Count
less than 2 years	446
2 to 5 years	142
more than 5 years	212
Total	**800**

Table 2: Label counts in Korean – Impact Duration train set.

3. Methods

We first designate a model that has been fine-tuned using supervised learning with KF-DeBERTa (Jeon et al., 2023), a Korean language model specialized for the financial domain, as our baseline model. Subsequently, to improve performance compared to the baseline model, we collect additional ESG-related news articles from the web and conduct semi-supervised learning using the collected data.

3.1 Finance-specialized Pre-trained Language Model

KF-DeBERTa (Jeon et al., 2023) is trained on a large-scale Korean financial corpus and follows the architecture and methods of DeBERTa (He et al., 2020). KF-DeBERTa is suitable for ESG-related tasks because it showed state-of-the-art performance in most evaluations of general and financial domains. In particular, the DeBERTa architecture has a significant advantage in understanding long sequences like in this dataset because it uses relative position embeddings, compared to BERT (Devlin et al., 2018) architecture models that use absolute position embeddings. To take advantage of this, we used the number of max position embeddings used for relative position embedding allocation as a hyperparameter during fine-tuning. Table 3 shows the performance of the validation set of Korean - Impact Type according to the number of max position embeddings. We chose 1,792 as the max position embeddings to be used for all future experiments.

Max Position Embeddings	Micro-F1	Macro-F1
512	0.8197	0.7417
768	0.8279	0.7553
1024	**0.8361**	0.7466
1280	0.8279	0.7613
1536	**0.8361**	0.7555
1792	**0.8361**	0.7814
2048	0.8179	**0.7881**

Table 3: Effects of max position embeddings on performance in Korean – Impact Type validation set.

3.2 Semi-supervised Learning

Semi-supervised learning has been shown to be effective in improving model performance when labeled data is scarce (Tarvainen and Valpola, 2017; Bertheolot et al., 2019; Xie et al., 2020; Shon et al., 2020). In the case of this task, we believed that semi-supervised learning utilizing unlabeled data would be effective since the number of labeled data is only 800. We collected 2,916 unlabeled data by crawling ESG-related news articles from the web and applied the ideas of UDA (Xie et al., 2020) and FixMatch (Shon et al., 2020), which are consistency training-based semi-supervised learning methods. Consistency training methods regularize model predictions to be invariant to noise injected into input examples or hidden states. UDA utilizes high-quality augmentation methods that have traditionally been effective in supervised learning as noise to be injected into unlabeled data. In each iteration, UDA calculates the supervised loss for a mini-batch of labeled data and the consistency loss for a mini-batch of unlabeled data using the model prediction of the unlabeled example as a soft pseudo-label for the augmented unlabeled example. It then calculates the final loss by summing the two losses. Generally, a larger batch size is used for consistency loss than for supervised loss.

FixMatch employs both weak and strong augmentation techniques for processing unlabeled data. Weak augmentation is applied to unlabeled examples to create hard pseudo-labels, and strong augmentation is applied to unlabeled examples to create model predictions and calculate consistency loss.

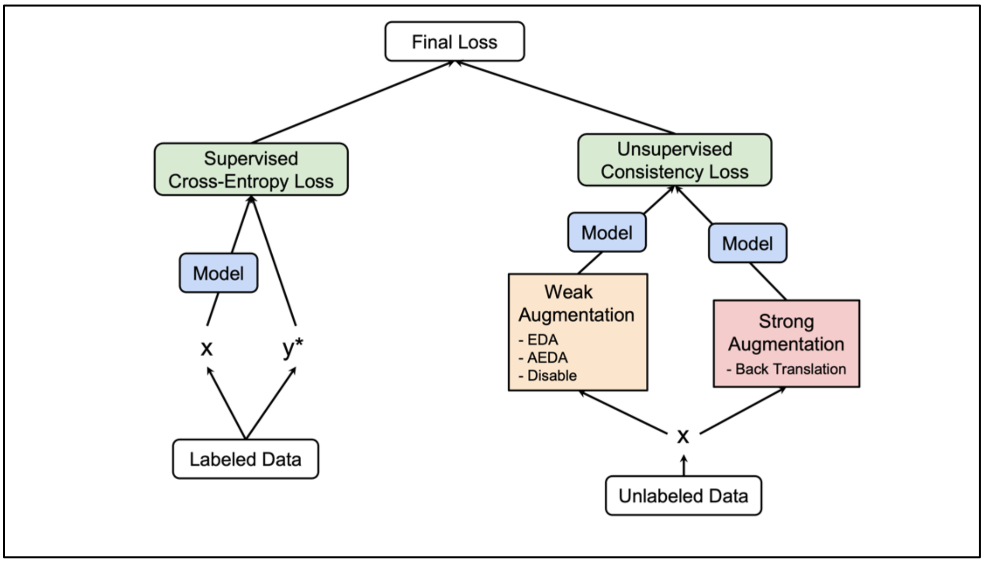


Figure 1: The entire process of the semi-supervised learning we used.

We chose the idea of using both weak augmentation and strong augmentation from FixMatch for augmentation diversity and the idea of using soft pseudo-labels from UDA to mitigate the model's overconfidence in unlabeled data. We used EDA (Wei and Zou, 2019) and AEDA (Karimi et al., 2021) for weak augmentation and also considered not using weak augmentation. When weak augmentation is not used, it is the same as UDA. EDA augmentation applies Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD) to some of the words in a sentence. We only used SR and RS for augmentation in EDA, as they were empirically suitable for Korean data. AEDA augmentation randomly selects some of all positions between words in a sentence and inserts one of the six punctuation marks {".", ";", "?", ":", "!", ","} randomly selected at each position. We also used back translation for strong augmentation, where we first translated the Korean unlabeled data into English using machine translation and then back into Korean.

To summarize our method, we calculate the supervised loss using labeled data, create soft pseudo-labels by applying weak augmentation to unlabeled data, and calculate consistency loss by applying strong augmentation to create model predictions. We then calculate the final loss by summing the two losses. Figure 1 shows the entire process of the semi-supervised learning we used. The loss used for training can be formulated as follows:

$$L = L_s + L_c \tag{1}$$

$$L_s = \frac{1}{B}\sum_{i=1}^{B} CE\left(y^*, p_\theta\left(y \middle| x_i^l\right)\right) \tag{2}$$

$$L_c = \frac{1}{\mu B}\sum_{i=1}^{\mu B} CE\left(p_{\tilde{\theta}}\left(y \middle| \alpha(x_i^u)\right), p_\theta\left(y \middle| \mathcal{A}(x_i^u)\right)\right) \tag{3}$$

where L is the total loss, L_s is the supervised loss, L_c is the consistency loss, $p_\theta\left(y \middle| x_i^l\right)$ is the model's predicted probability distribution for the target given the i-th labeled example x_i^l, $\tilde{\theta}$ is a fixed copy of the current parameters θ indicating that the gradient is not propagated through $\tilde{\theta}$, x_i^u is the i-th unlabeled example, B is the batch size of labeled data, μ is the ratio of unlabeled data to labeled data, μ is the multiplier used to determine the batch size of unlabeled data μB by multiplying it with the batch size of labeled data B. CE is the cross-entropy loss function, y^* is the one-hot encoded label for labeled example, α is the weak augmentation function, $\mathcal{A}$ is the strong augment-ation function.

Table 4 shows the performance on the Korean-Impact Type validation set for each configuration. The batch size of the unlabeled data was most effective when it was 4 to 5 times the batch size of the labeled data. In the weak augmentation setting, AEDA led to decreased performance.

μ	weak aug.	strong aug.	Micro-F1	Macro-F1
4	-	BT	0.8361	**0.7901**
4	EDA	BT	**0.8443**	0.7525
4	AEDA	BT	0.8279	0.7506
5	-	BT	**0.8443**	0.7603

Table 4: Performance on the Korean - Impact Type validation set by augmentation methods. "BT" stands for Back Translation, and "aug." is short for

augmentation. μ is the multiplier used to determine the batch size of unlabeled data μB by multiplying it with the batch size of labeled data B.

4. Experiments

4.1 Training Setup

We used 120 samples out of the 800 samples in the train set as a validation set. For training, we used the AdamW optimizer (Loshchilov and Hutter, 2017) with a linear learning rate schedule having a warmup of 100 steps and an initial learning rate of 2.5×10^{-5}. Batch size was set to 4, weight decay to 0.01, and gradient clipping to 1.0. We conducted training for 5 to 12 epochs and also utilized the exponential moving average (EMA) of weights with decay rates of 0.99 and 0.999.

4.2 Results

We evaluated the Korean - Impact Type dataset and the Korean - Impact Duration dataset using the Micro-F1 and Macro-F1 performance metrics.

Our SSL method worked well on the Korean - Impact Type dataset. The model trained with SSL showed improved Micro-F1 and Macro-F1 performance on the validation set compared to the supervised learning baseline model. On the other hand, the model with EMA applied did not show performance improvement compared to the baseline. We submitted the baseline model and two SSL models based on validation set performance. In the final results, one of the SSL models achieved 1st place with Test Micro-F1 of 0.8400 and Test Macro-F1 of 0.7985. Table 5 shows the experimental results on the Korean - Impact Type dataset.

The EMA technique was effective on the Korean - Impact Duration dataset. EMA is a technique that calculates the exponential moving average of model weights to reduce noise and decrease variability, thereby stabilizing the learning process (Izmailov et al., 2018). It helps prevent overfitting and improves generalization performance. The model with EMA applied showed improved Micro-F1 performance on the validation set compared to the supervised learning baseline model, and some models also showed improved Macro-F1 performance. In contrast, the model trained with SSL did not show performance improvement over the baseline. We submitted three EMA models based on validation set performance. In the final results, one of the EMA models achieved 2nd place with Test Micro-F1 of 0.6750 and Test Macro-F1 of 0.6198. Table 6 shows the experimental results on the Korean - Impact Duration dataset.

Model	Valid. Micro-F1	Valid. Macro-F1	Test Micro-F1	Test Macro-F1
baseline	0.8361	0.7814	0.8050	0.7343
EMA	0.8279	0.7522	-	-
SSL #1	0.8361	**0.7901**	0.8150	0.7398
SSL #2	**0.8443**	0.7603	**0.8400**	**0.7985**

Table 5: Experimental results in Korean - Impact Type.

Model	Valid. Micro-F1	Valid. Macro-F1	Test Micro-F1	Test Macro-F1
baseline	0.7869	0.7438	-	-
EMA #1	**0.7951**	0.7579	**0.6750**	**0.6198**
EMA #2	**0.7951**	**0.7608**	0.6650	0.6102
EMA #3	**0.7951**	0.7339	**0.6750**	0.6154
SSL	0.7705	0.7164	-	-

Table 6: Experimental results in Korean - Impact Duration.

5. Conclusion

In this paper, we presented our approach and results for the Multilingual ESG Impact Duration Inference (ML-ESG-3) shared task at FinNLP-KDF@LREC-COLING-2024. Our methodology, which employed semi-supervised learning and exponential moving average of weights on a finance-specialized pre-trained language model, demonstrated strong performance in the Korean - Impact Type and Korean - Impact Duration subtasks. Our model achieved 1st place in the Korean - Impact Type subtask and the 2nd place in the Korean - Impact Duration subtask. These results highlight the potential of our methodology in identifying ESG-related issues from news articles.

6. Bibliographical References

Tseng, Y. M., Chen, C. C., Huang, H. H., & Chen, H. H. (2023, October). DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (pp. 5412-5416).

Jeon, E., Kim, J., Song, M., & Ryu, J. (2023). KF-DeBERTa: Financial Domain-specific Pre-trained Language Model. In *Proceedings of the 35th Annual Conference on Human and Cognitive Language Technology* (pp. 143-148).

He, P., Liu, X., Gao, J., & Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled

attention. *arXiv preprint arXiv:2006.03654*.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, *33*, 6256-6268.
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., ... & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, *33*, 596-608.
Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, *30*.
Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.
Karimi, A., Rossi, L., & Prati, A. (2021). AEDA: an easier data augmentation technique for text classification. *arXiv preprint arXiv:2108.13230*.
Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.
Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.
Kannan, N., & Seki, Y. (2023). Textual Evidence Extraction for ESG Scores. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting* (pp. 45-54).
Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, *32*.

Adapting LLM to Multi-lingual ESG Impact and Length Prediction using In-context Learning and Fine-Tuning with Rationale

Pawan Kumar Rajpoot, Ashvini Jindal, Ankur Parikh
SCB DataX Thailand, LinkedIn AI USA, UtilizeAI Research India

Abstract

The prediction of Environmental, Social, and Governance (ESG) impact and duration (length) of impact from company events, as reported in news articles, hold immense significance for investors, policymakers, and various stakeholders. In this paper, we describe solutions from our team "Upaya" to ESG impact and length prediction tasks on one such dataset ML-ESG-3. We employed two different paradigms to adapt Large Language Models (LLMs) to predict both ESG impact level and length of events. In the first approach, we leverage GPT-4 within the In-context learning (ICL) framework where a retriever identifies top K-relevant in-context learning examples for a given test example. The second approach involves instruction-tuning Mistral (7B) LLM to predict impact level and duration, supplemented with rationale generated using GPT-4. Our models secured second place in both French tasks where for one task fine-tuned Mistral model outperformed and for other task, GPT-4 with ICL outperformed. These results demonstrate the potential of different LLM-based paradigms for delivering valuable insights within the ESG investing landscape.

Keywords: ESG Impact, ESG Length, Large Language Models, FinNLP, Q-LoRA, In-Context Learning, Rationale Generation, Chain of Thoughts

1. Introduction

Environmental, Social, and Corporate Governance (ESG) factors have become pivotal in assessing the long-term sustainability and ethical impact of businesses, investments, and policy decisions. The integration of ESG criteria in investment strategies aims to mitigate risks, identify opportunities aligned with responsible practices, and foster positive change.

The advent of large language models (LLMs), exemplified by GPT-4 (Brown et al., 2020) (Thoppilan et al., 2022), marks a significant breakthrough in natural language processing (NLP). These models exhibit proficiency across various domains and can be readily applied to multiple NLP tasks. Traditionally, language models follow distinct pre-training and fine-tuning pipelines (Devlin et al., 2018) (Beltagy et al., 2019) (Raffel et al., 2020) (Lan et al., 2019) (Liu et al., 2021b), where fine-tuning occurs after pre-training on task-specific datasets in a fully-supervised manner.

A recent paradigm, In-context Learning (ICL) (Brown et al., 2020) (Thoppilan et al., 2022), reshapes NLP tasks, enabling LLMs to make predictions by learning from demonstrations presented within the context prompt. Under the ICL framework, LLMs achieve remarkable performance, rivaling fully-supervised methods, even with a limited number of demonstrations. The retrieval of contextually relevant examples plays a crucial role in overall performance, as LLMs benefit from examples similar to the "to be predicted" data point, reducing hallucination and improving performance.

This paper explores two approaches within the ML-ESG-3 dataset for English and French datasets: 1. Guiding GPT-4 under the ICL framework to predict ESG impact and event duration, using a learning-free dense retriever to identify top K relevant In-context learning examples. 2. Instruction-tuning the open-source LLM, Mistral, with 7B parameters to predict ESG impact and duration, incorporating rationale. Efficient fine-tuning is achieved through Parameter Efficient Fine Tuning (PEFT), specifically QLoRA 4-bits quantization.

2. Preliminary Background

2.1. Task Definition

As per the challenge "ESG Impact Level and Length Prediction" (Chen et al., 2024) is the task of automatically determining the ESG impact level - opportunity or risk and the duration (length) of the impact an event in the news article might have on the company". This shared task is a part of the Fifth Workshop on Knowledge Discovery from Unstructured Data in Financial Services, co-located with LREC-COLING 2024.

Let x denote the news article. Given a set of predefined impact level classes, Level=*Low*, *Medium*, *High* and a set of predefined impact length classes, Length=*Short-Term*, *Medium-Term*, *Long-Term*, the task aims to predict the class c_1 in level and c_2 in length for input x.

2.2. Data

The English dataset released with this task contains 545 train and 136 test (evaluation) instances. While the French dataset had 661 training examples and 146 test (evaluation) examples.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 274–278

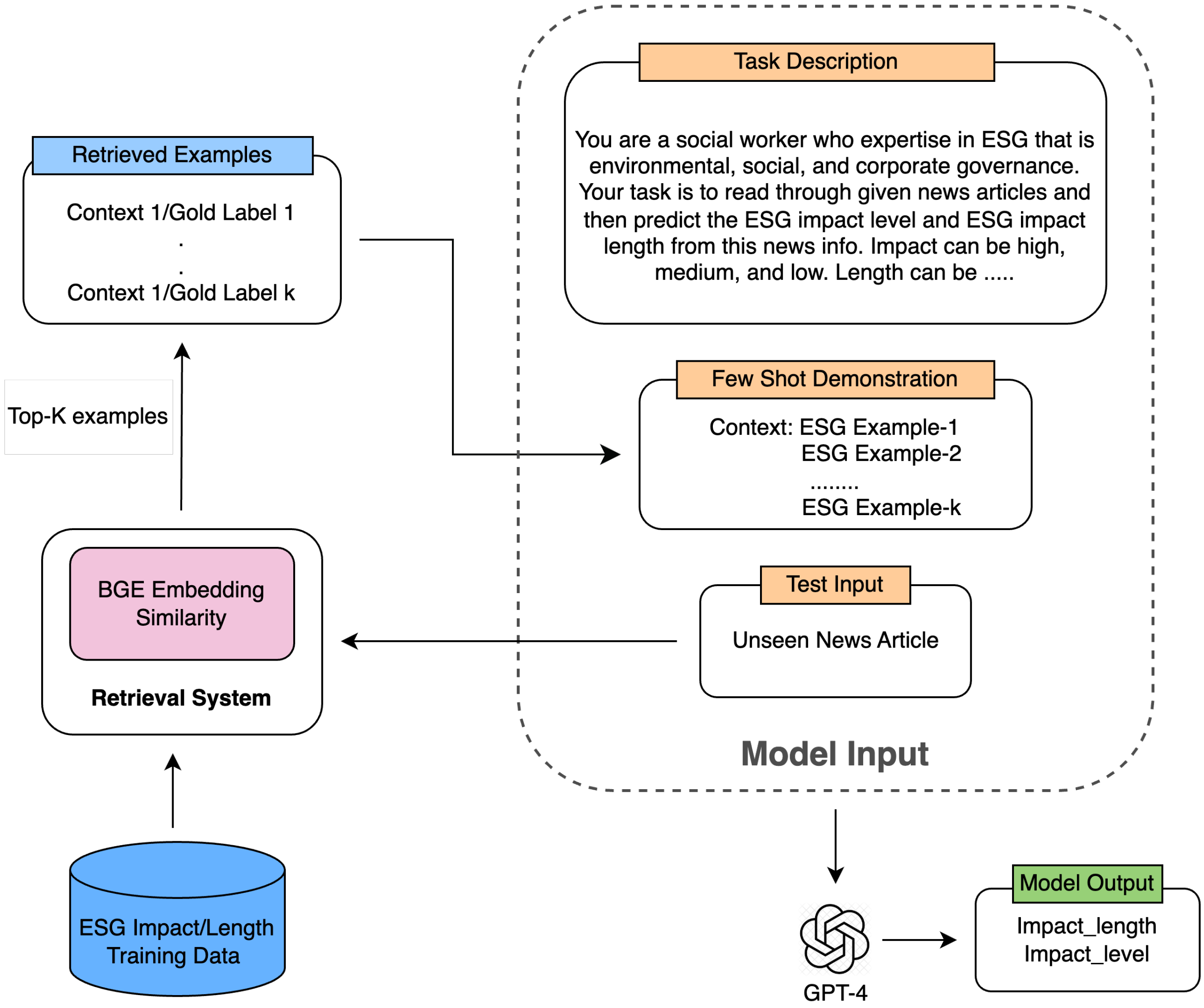


Figure 1: Approach 1: In-context Learning

2.3. In-context Learning

In-context learning (ICL) is a key emergent ability of language models (Wei et al., 2023), allowing them to infer tasks from context. Unlike gradient-based 'in-weights learning' (which updates model parameters), ICL is gradient-free, adapting directly from the context ((Brown et al., 2020). Formally, each training instance is first linearized into an input text x and an output text y. Given a test input text x_{test}, in-context learning defines the generation of output y as $y_{test} \sim PLM\ (y_{test}|x_1, y_1, \dots x_k, y_k, x_{test})$, where k refers to number of in-context examples and $\sim$ refers to decoding strategies(e.g., greedy decoding and nuclear sampling (Li et al., 2022)), and each in-context example $e_i = (x_i, y_i)$ is sampled from a training set D. The generation procedure is especially attractive as it eliminates the need for updating the parameters of the language model when encountering a new task, which is often expensive and impractical. Notably, the performance of ICL on downstream tasks can vary from almost random to comparable with state-of-the-art systems, depending on the quality of the retrieved in-context examples (Rubin et al., 2021) (Liu et al., 2021a).

3. Adapting LLM for ESG Impact Level and Length Prediction

We employed two paradigms to adapt LLMs for the specific task. 1. In-context Learning and 2. Instruction Fine-Tuning.

3.1. In-context Learning

The formalization of the task under the ICL framework, using GPT-4 is shown in figure 1.

3.1.1. Prompt Construction

For each test example, a prompt is meticulously constructed and subsequently input to GPT-4. The prompt encompasses the following key components:

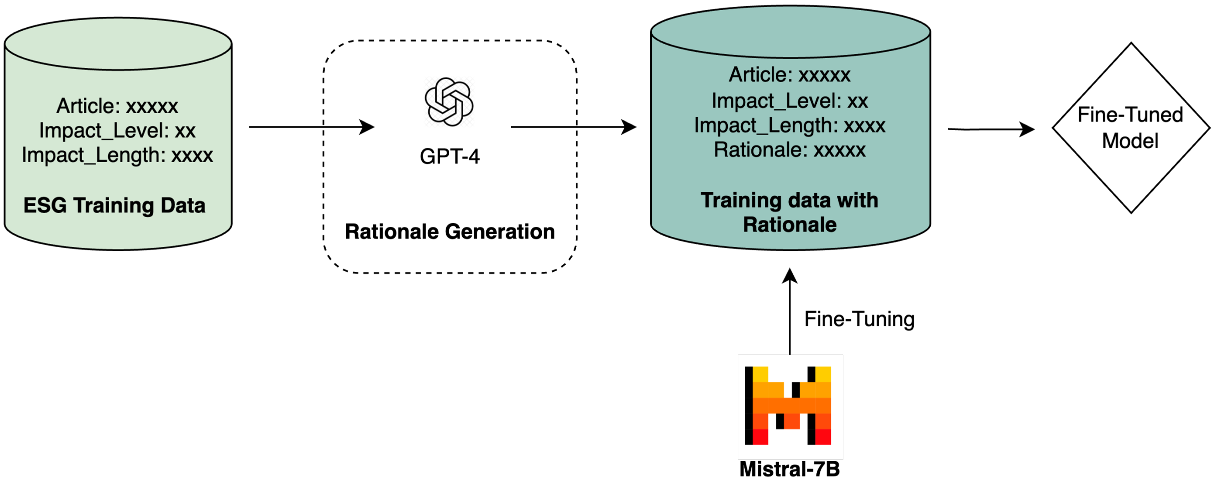


Figure 2: Approach 2: Instruction Fine-tuning

- **Task Description** - A concise overview is presented, outlining the task and the predefined classes for impact level and length prediction tasks.
- **K-shot Demonstrations** - The demonstration section involves the reformulation of each example, displaying the input x_{demo} and corresponding class labels $c1_{demo}$ and $c2_{demo}$ where c_1 refer to level and c_2 refers to length.
- **Test Input** - Test input x_{test} is provided, and GPT model is tasked with generating the corresponding class $c1_{test}$ and $c2_{test}$

3.1.2. Retrieval with Cosine Similarity over BGE Embedding

In-context learning (ICL) using demonstrations that are closer to the test sample within the embedding space tend to perform better than random selection. We used cosine similarity to find the most relevant examples from the training set. To represent these examples, we used BGE Embeddings (Xiao et al., 2023), and FAISS for fast similarity search (Johnson et al., 2019)

3.2. Instruction Fine-Tuning with Rationale

The formalization of the task under the Instruction Fine-Tuning with Rationale framework, using Mistral Base Model (7B) is shown in figure 2.

3.2.1. Rationale Generation

Since we are using relatively smaller model (7B) for fine-tuning, we employed Chain-of-Thought (CoT) based paradigm. Instead of directly generating only label, the LLM should generate both rationale and label. To generate both rationale and labels as output, the system needs to be fine-tuned with rationale as part of the output. Since the rationale behind labels wasn't available in the training dataset, we used GPT-4 to generate the rationale for each training data sample. Refer to section 8 for the description used in the prompt and generated rationale. For the English task in the dataset, this approach worked well. For the French language, we translated training data from French to English using GPT-4 and then used the same methodology to generate rationale.

3.2.2. Instruction Fine-Tuning

We fine-tuned the Mistral-7B base model using an English training set with GPT-4 rationale, plus a French set (translated to English) with GPT-4 rationale. Due to memory limits, we used 4-bit QLoRA (Dettmers et al., 2024) with rank 128 and alpha 256. Quantized LoRA was applied to self-attention Query, Key, Value matrices and Linear layers. We used gradient accumulation (steps=2), paged Adamw 32bit optimizer, cosine schedule (LR=2e-5), decay rate 0.01, and 5 warmup steps. Fine-tuning was done using axolotl [1]

4. Experiments and Results

Maximum of 3 submissions were allowed for each language subtask. We submitted for both English and French subtasks as shown in Table 1. Specifically, we submitted one entry with instruction-tuned Mistral model and two entries with ICL with different values of K (number of demonstrations retrieved).

[1] https://github.com/OpenAccess-AI-Collective/axolotl

Submission	Language	Approach
E1	English	Fine-tune
E2	English	ICL 10-shot
E3	English	ICL 20-shot
F1	French	Fine-tune
F2	French	ICL 10-shot
F3	French	ICL 20-shot

Table 1: Our Submission details

4.1. Impact Level

Table 2 shows results for Impact Level prediction task. For English language, our fine-tuned Mistral 7B based model outperformed GPT-4 with K-shot learning. For French language, the performance of fine-tuned Mistral and GPT-4 with 20-shot is comparable.

Submission	Micro-F1	Macro-F1
E1	54.41	48.40
E2	53.68	45.93
E3	51.47	46.09
F1	58.22	56.78
F2	58.22	56.69
F3	42.47	37.64

Table 2: Overall scores on Impact level prediction

4.2. Impact Length

Table 3 shows results for Impact Length prediction task. For English language, GPT-4 with 20-shot learning performs better than fine-tuned Mistral. However, for the French language, GPT-4 with 10-shot learning performs better than the fine-tuned Mistral model. In summary, for the Impact Level task, fine-tuned Mistral model outperformed GPT-4 with ICL. However, for Impact Length task, GPT-4 with ICL outperformed fine-tuned Mistral model.

Submission	Micro-F1	Macro-F1
E1	57.35	42.75
E2	51.47	38.55
E3	60.29	44.23
F1	46.58	42.86
F2	52.05	48.73
F3	41.10	32.09

Table 3: Overall scores on Impact length prediction

Overall, in the context of the shared-task, for the Impact Length prediction (French Language), our submission F2 got 2nd rank. For the Impact Level prediction (French Language), our submission F1 got 2nd rank. For the Impact Length prediction (English Language), our submission E3 got 7th rank and for the Impact Level prediction (English Language), our submission E1 got 17th rank.

5. Conclusion

This work explores the potential of GPT + ICL and Mistral (7B) + Fine-Tuning with Rationale on ESG Impact Level and Length Prediction task. For Impact Level prediction, the fine-tuned model performed better. Form Impact Length prediction, the GPT + ICL combination performed better. We achieved 2nd rank in both sub-tasks for French Langauge.

6. Acknowledgements

We thank the organizing committee of FinNLP-2024 along with the task-setting team of Shared Task ML-ESG-3 for giving us this opportunity to work on this problem. We also would like to thank Lambda Labs[2] for providing the compute resources required for this work.

7. Bibliographical References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Hanwool Lee, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2024. Multi-lingual esg impact duration inference. In *Proceedings of Joint Workshop of the 7th Financial Technology and Natural Language Processing and the 5th Knowledge Discovery from Unstructured Data in Financial Services*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient fine-tuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

[2]https://lambdalabs.com/

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Zhuang Liu, Wayne Lin, Ya Shi, and Jun Zhao. 2021b. A robustly optimized bert pre-training approach with post-training. In *China National Conference on Chinese Computational Linguistics*, pages 471–484. Springer.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2021. Learning to retrieve prompts for in-context learning. *arXiv preprint arXiv:2112.08633*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

8. Appendix

8.0.1. Rationale Generation

GPT4 Prompt for Rationale Generation

You are an annotator who has already annotated ESG IMPACT_LEVEL and IMPACT_LENGTH on news articles. Now you have to justify your annotations for each data point, why you gave the annotation? Give an answer in minimum words.

Article: High risk of strain on the power grid: an energy-by-energy status report The grid operator RTE has updated its forecasts for the winter. While these forecasts now predict a fall in electricity consumption, they also anticipate a drop in nuclear electricity generation. This should almost balance out. Despite everything, RTE warns that the risk of a power cut cannot be ruled out, particularly in January. Here is a status report on each energy source.
impact_level: high
impact_length: More than 5 years

GPT4 Rationale Generation Output

RATIOANLE_IMPACT_LEVEL: high - Risk of power cuts affects national energy security, economy, public safety, and has cascading effects on all sectors.
RATIOANLE_IMPACT_LENGTH: More than 5 years - Energy infrastructure and generation capacity issues are complex and typically take many years to address and improve.

ESG-GPT:GPT4-based Few-Shot Prompt Learning for Multi-lingual ESG News Text Classification

Ke Tian, Hua Chen
Rakuten Group Inc, JiangXi Normal University
Tokyo, Japan, Jiangxi Province, China
tianke0711@gmail.com, hua.chen@jxnu.edu.cn

Abstract

Environmental, Social, and Governance (ESG) factors for company assessment have gained great attention from finance investors to identify companies' risks and growth opportunities. ESG Text data regarding the company like sustainable reports, media news text, and social media text are important data sources for ESG analysis like ESG factors classification. Recently, FinNLP has proposed several ESG-related tasks. One of the tasks is Multi-Lingual ESG Issue Identification 3(ML-ESG-3) which is to determine the duration or impact level of the impact of an event in the news article regarding the company. In this paper, we mainly discussed our team: KaKa's solution to this ML-ESG-3 task. We proposed the GPT4 model based on few-shot prompt learning to predict the impact level or duration of the impact of multi-lingual ESG news for the company. The experiment result demonstrates that GPT4-based few-shot prompt learning achieved good performance in leaderboard quantitative evaluations of ML-ESG-3 tasks across different languages.

Keywords: ESG, GPT-4, Few-shot Learning, Prompt Learning

1. Introduction

Recently "Environment, Social, and Governance (ESG)" related issues in the financial domain have gained more and more attention with the goal of building a sustainable environment. ESG evaluation is considered an essential tool for investors to assess a company's sustainability and ethical performance. ESG Text data regarding the company like sustainable reports, media news text, and social media text are important data sources for ESG evaluation like ESG impact level or ESG score prediction. Recently, the FinNLP organizers proposed several ESG-related shared tasks for this topic. In FinNLP-2022 (FinNLP-2022, 2022), they proposed a FinSim4-ESG shared task which is related to the ESG topic detection of ESG term words and sustainable sentence classification. Moreover, multi-lingual ESG identification tasks: ML-ESG-1(FinNLP-2023, 2022). ML-ESG-1 is to classify the ESG-related news into 35 ESG key issues. Furthermore, In real-world application scenarios such as making financial decisions, the opportunity (risk) of short-term or long-term impact of ESG news regarding the company should be taken into account. The ML-ESG task organizers defined the following definitions and categories regarding the impact type and impact duration.

- Impact Type Identification:This single-choice question aims to ascertain the type of impact a news article might have on the company. The possible labels are "Opportunity", "Risk", and "Cannot Distinguish".
- Impact Duration (Length) Inference:This single-choice question seeks to determine the duration of the impact a news article might have on the company. Based on the distinction between short-term and long-term defined above, we present three labels: "Less than 2 years", "2 to 5 years", and "More than 5 years".

Considering the importance of impact type and duration, the FinNLP organizer proposed ML-ESG-2(FinNLP-2023, 2023a) and ML-ESG-3(FinNLP-2023, 2023b) subsequently. ML-ESG-2 is to detect the ESG impact type (opportunity or risk) of ESG news regarding the company. In ML-ESG-3, the goal of this task is to determine the duration or length of the impact an event in the multi-lingual ESG news might have on the company.

To challenge these ESG tasks, a variety of methods have been proposed. NLP and deep learning are the dominant techniques(Ke Tian, 2019)(Ke Tian, 2021)(Ke Tian, 2022). However, it is difficult to apply one trained deep learning or ML model across different language ESG texts, it is required to train multi-models to solve the multi-lingual dataset task. Recently, the emergence of Large Language Models (LLM), represented by ChatGPT, has exhibited great performance in general Natural Language Processing (NLP) tasks and across different language texts. These LLMs can complete various tasks by transforming them into generative paradigms using prompt learning. For example, ChatGPT using prompt learning can perform well on text classification, text generation, sentiment detection, NER extraction, etc. As for the ML-ESG-3 task, there are 5 languages of ESG news tasks, we listed the details of each sub-task in Table 1.

In this paper, we presented our solution to the ML-ESG-3 task. Considering the multi-lingual datasets

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 279–282

Language	Task goal
Chinese	Classify the news text into impact duration labels:"Less than 2 years", "2 to 5 years", and "More than 5 years"
English	There are two tasks in this dataset: Impact Level and Impact Length classification.
French	There are two tasks in this dataset: Impact Level and Impact Length classification.
Japanese	Classify the news text into impact duration labels: "Less than 2 years", "2 to 5 years", and "More than 5 years".
Korean	The are two tasks in this dataset: Impact Type and Impact Length classification.

Table 1: Details of ML-ESG-3 task.

in this task, we created firstly four prompts for each language task goal, then utilized the GPT-4(OpenAI) few-shot learning method to predict the impact type or duration of news text in each task. Our approach achieved 1st place for the French language's impact level and performed well in other sub-tasks.

2. Method

2.1. In-Context Learning

Recently, much research work on large language models (LLMs) has explored the phenomenon of in-context learning (ICL). In this paradigm, an LLM learns to solve a new task at inference time (without any change to its weights) by being fed a prompt with examples of that task. For example, if you ask ChatGPT to categorize different, you might first give it example pieces with their correct categorizations in the prompt, then ask ChatGPT to classify the input text with provided label categories in the prompt. In our proposed method, the GPT-4(Open AI) model is utilized, and GPT-4 is a multi-modal model able to consume 32,768 tokens.

2.2. Few-shot Prompt Learning

Few-Shot Learning is the method where a machine learning model is trained with a minimal set of data to shape its predictions, using only a handful of examples at the time of inference, unlike traditional fine-tuning methods that demand a considerable volume of data for the pre-trained model to fine-tune itself accurately to a new task. Recently, with the advent of cutting-edge Language Models such as OpenAI's GPT-3 and GPT-4, its application has broadened to encompass Natural Language Processing (NLP). Within NLP, Few-Shot Learning is applicable to Large Language Models which have, during their pre-training phase on extensive textual datasets, inherently acquired the capability to undertake a diverse array of tasks. This pre-training equips the models with the ability to generalize, or understand and perform tasks that are similar yet not previously encountered, with a few examples serving as guidance. The method is exemplified through the task of translating between English and French, as illustrated in Fig 1.

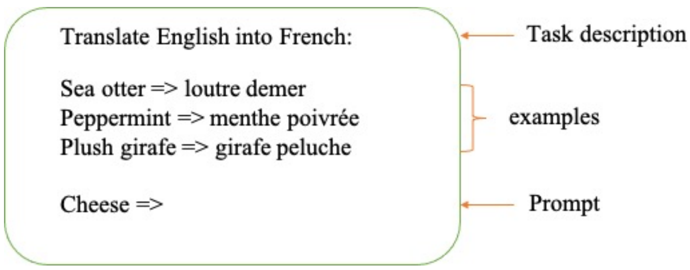


Figure 1: Few-shot prompt learning Example.

2.3. Proposed Method

We proposed the GPT4-based In-context learning with few shot learning for our task, the overall procedure of the proposed method is shown in Fig 2.

Firstly, we utilized the training dataset to create few-shot learning examples for GPT4 model learning. Secondly, the prompt engineering integrates the few-shot learning example, instruction, and test dataset (news title & content) for creating a prompt. Finally, the prompt text as input to call GPT-4 API to predict the input prompt to obtain the result.The schema of the prompt consists of the following components:

- Task description: explain the purpose of the task content.

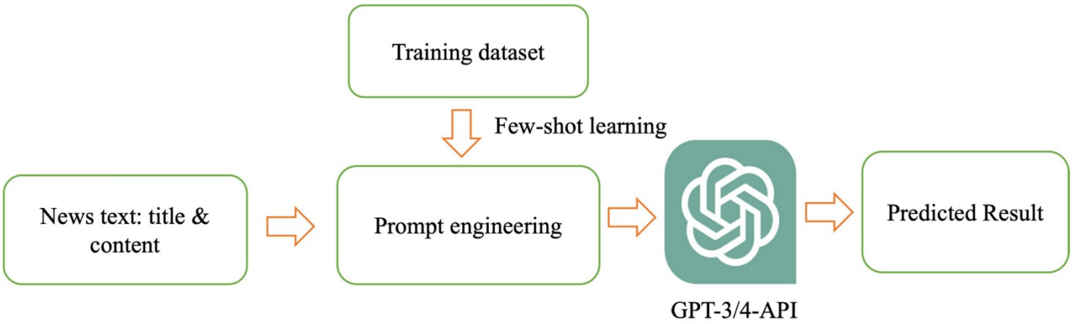


Figure 2: Overall procedure of method.

- Instruction prompt: The sentences that describe the key information regarding the task that the model needs to complete based on required constraints. Moreover, also includes the few-shot learning example based on the train label data.
- Input: the input text for the model predicting Response: the predicted result regarding the input text.

We take the English dataset as an example to explain the above prompt.

As the task description: " Below is an instruction that describes the English ESG text classification task. Write a response that appropriately completes the request. Instruction: " In this task, you are presented with English news titles and their corresponding content, all of which are focused on Environmental and Social (ES) Governance (ESG) themes. Your objective is to evaluate the potential future impact level of these news items to assist in predicting the viability of investments based on their environmental and social implications. There are three categories to classify the impact level: low, medium, and high. Below are 9 examples of English news texts, including both the title and content, related to impact level in the end as follows: created few-shot examples. Your task is to classify the input text, which consists of the news title and content, into one of the three impact level categories: low, medium, and high. Please respond with only one of the three category labels: low, medium, and high, based on your analysis of the news' future impact level. Do not include any additional words or explanations in your response."

As examples in the above string, we used the following example format: input: news title: Guest Post: Carbon Trading and Transfer Pricing; News Content: In order to meet overall carbon emissions ... impact length: long.

input: news title: Guest Post: Eaton Appoints Harold Jones as Chief; News Content: Eaton Appoints Harold Jones as Chief Sustainability Officer; impact length: medium.

3. Result

To understand the model's performance, the organizer used micro-F1 and macro-F1 indicators to evaluate the performance of each submission. The performance of our submitted result is displayed in Table 2.

As for the English task result, the impact level and impact length result are obviously different which is caused by the different number of few-shot learning examples. The result shows that the more few-shot learning examples are better for helping the GPT4 model to understand the semantic meaning of the task. The French submission result achieved the best performance for the impact level in the leaderboard. Our proposed method also achieved good performance in the Korean and Japanese dataset tasks. However, there is still a gap between our solution results and other top results in the leaderboard.

4. Conclusion

We presented GPT-4 mode-based few-shot text classification for the ML-ESG-3 task. We demonstrated that generative LLMs, like GPT-3.5 and GPT-4, can perform well in solving the multi-lingual ESG text classification. Although the French submission result achieved the best performance for the impact level in the leaderboard. There is still a gap between our solution results and other top results in the leaderboard. The following direction will be conducted in the next step: Firstly optimizing the prompt engineering to obtain a better result. Secondly, fine-tuning the GPT-3/4 model or other LLMs to make the GPT model understand well the training dataset.

5. Acknowledgements

This research has been supported by the Jiangxi Double Thousand Plan (JXSQ2019101077) fund, China.

Language	Impact Length		Impact Level		Impact Type	
	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1
English	52.94%	36.36%	51.47%	51.07%	-	-
French	46.58%	47.42%	63.70%	63.29%	-	-
Japanese	34.90%	25.50%	-	-	-	-
Korean	56.00%	52.94%	-	-	63.00%	55.53%

Table 2: Result of sub-task submission in leaderboard.

6. References

FinNLP-2022. 2022. Shared task finsim4-esg. Accessed: 2024-03-08.

FinNLP-2023. 2022. Multi-lingual esg issue identification (ml-esg) shared task. Accessed: 2024-03-08.

FinNLP-2023. 2023a. Multi-lingual esg impact type identification (ml-esg-2) shared task. Accessed: 2024-03-08.

FinNLP-2023. 2023b. Multi-lingual esg impact type identification (ml-esg-3) shared task. Accessed: 2024-03-08.

Hua Chen Ke Tian. 2021. aiai at the finsim-2 task: Finance domain terms automatic classifica- tion via word ontology and embedding. In *In The 1st Work- shop on Financial Technology on the Web (FinWeb) The Web Conference*, Ljubljana, Slovenia.

Hua Chen Ke Tian, ZePeng Zhang. 2022. Automatic term and sentence classification via augmented term and pre-trained language model in esg taxonomy texts. In *The 4th Workshop on Financial Technology and Natural Language Processing (FinNLP) In conjunction with IJCAI-ECAI-2022*, Vienna, Austria.

Zi Jun Peng Ke Tian. 2019. aiai at finsbd task: Sentence boundary detection in noisy texts from financial documents using deep attention model. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, Macao, China.

OpenAI. 2023. *GPT-4 technical report*. Open AI.

Shared Task for Cross-lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics

Yola Nayekoo[1], Sophia Katrenko[1], Véronique Hoste[2], Aaron Maladry[2], Els Lefever[2]
[1]EcoVadis, Avenue de la Grande Armée 43, Paris, France
[2]Language and Translation Technology Team, Ghent University, Groot-Brittanniëlaan 45, Ghent, Belgium
{ynayekoo, skatrenko}@ecovadis.com
{veronique.hoste, aaron.maladry, els.lefever}@UGent.be

Abstract

This paper provides an overview of the Shared Task for Cross-lingual Classification of CSR Themes and Topics. We framed the task as two separate sub-tasks: one cross-lingual multi-class CSR theme recognition task for English, French and simplified Chinese and one multi-label fine-grained classification task of CSR topics for Environment (ENV) and Labor and Human Rights (LAB) themes in English. The participants were provided with URLs and annotations for both tasks. Several teams downloaded the data, of which two teams submitted a system for both sub-tasks. In this overview paper, we discuss the set-up of the task and our main findings.

Keywords: multilingual CSR, multi-label classification, CSR theme detection

1. Introduction

Today, business organizations are expected to report on matters that affect environment, the economy and people (impact materiality) as well as matters that influence enterprise value (financial materiality). Corporations are held accountable for impacts across the entire value chain and recognize the need for sustainable procurement, to reduce the risk of supply chain disruption, protect their brands and reputation, and facilitate access to capital. As a consequence, there is a growing need and interest in processing Corporate Social Responsibility content originating from both business organizations and media. Laws and regulations such as FCPA in the US, Sapin II and the UK Bribery Act have made companies even more liable for knowing about sustainability infractions, yet the information is difficult to uncover, anticipate, and manage.

For over 16 years, EcoVadis has been measuring the quality of a company's sustainability management system through its policies, actions and results. It has been screening a large variety of specialized sources and newspapers to identify CSR-related content and assess it with respect to CSR themes and criteria (topics). A key distinguishing element of EcoVadis' sustainability monitoring platform is the integration of this external input to augment company-provided documentation and data sources. Sustainability analysts assess news items in a variety of languages (e.g., English, Spanish, French) on how they impact the quality and effectiveness of the sustainability management system or reflect positive innovation. The analyzed results are then integrated as part of the EcoVadis sustainability rating, and are displayed on the EcoVadis scorecard, which allows businesses to monitor the sustainability performance of their trading partners as well as their continuous improvement actions.

Despite the progress in automatic information extraction in the last decades, no datasets or methodologies are available yet aiming at automatic CSR theme detection. This shared task, which is co-organized by EcoVadis[1], a business sustainability ratings provider, and by the Language and Translation Technology Team[2] (LT3) from Ghent University provides the NLP community with data sets in multiple languages (English, French, and simplified Chinese) for CSR news analysis and will shed light on the feasibility of cross-lingual CSR theme detection. In addition, we also provide data sets to gain insights into fine-grained topic classification for two large CSR themes, viz. Environment (ENV) and Labor and Human Rights (LAB) in English.

The remainder of the paper is organized as follows. In Section 2, we discuss related research linked to the analysis of financial and social responsibility information. The shared task setup is described in depth in Section 3 and includes details on the two sub-tasks, dataset annotation and the experimental data selected for both tasks. In Section 4, we list the results of the participating teams for both task A and B. We end the paper with the main conclusions of the shared task and some prospects for follow-up research.

2. Related Research

The task of detecting corporate social responsibility themes and topics is operationalized as a classification task (cf. supra). Text classification is by nature the most fundamental task in NLP (Li et al., 2020). With the advent of deep neural networks,

[1]www.ecovadis.com
[2]www.lt3.ugent.be

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 283–291

and transformer-based language models in particular, approaches to text classification have drastically changed. More traditional machine learning models were feature-based, and incorporated manually crafted features relying on linguistic insights and knowledge from experts. Neural networks, however, utilize the text data itself to derive the embeddings used as input for the model. Deep learning models “integrate feature engineering into the model fitting process by learning a set of nonlinear transformations that serve to map features directly to outputs” (Li et al., 2020). These neural models have shown to perform very well for a wide range of NLP tasks, and they automatically provide meaningful semantic representations without the need of human-designed features or rules. They do, however, also come with important drawbacks: they require huge amounts of data and computational resources, and they are *black boxes*, viz. it is hard to investigate what information is really captured by the model, or to trace back why certain predictions (or errors) are made.

Detecting fine-grained CSR topics for Environment and Labor and Human Rights in English (Task B) is a multi-label classification task: the topic labels are nonexclusive and there are no restrictions about the number of classes that need to be assigned to the instances. Most recent multi-label text classification research has addressed this uncertainty of the number of labels, mainly by recasting the multi-label classification task into a multi-task problem (Lin et al., 2022). Another challenge, however, is to construct a better semantic representation space when feeding multi-label instances. As mentioned by Lin et al. (2023), the semantic space becomes “susceptible to distractions” when confronted with multi-label samples, and the boundaries between the classes become “blurred”. They propose to deploy contrastive learning techniques to improve multi-label classification tasks.

While a large body of literature has already been devoted to the topic of CSR and CSR communication (Crane and Glozer, 2016), the application of natural language processing techniques in the domain of corporate social responsibility is fairly new and recently also gained some further visibility through the organization of the First Computing Social Responsibility Workshop which was held in collocation with LREC-2022 (Wan and Huang, 2022). Current NLP work on corporate social responsibility often deals with the collection and (automatic) analysis of CSR reports, i.e. regular reports published by a company or an organization about the economic, environmental and social impacts caused by its activities. The work on CSR reports among others describes the collection of corpora of CSR reports (Händschke et al., 2018; Purver et al., 2022), the analysis of financial and corporate social responsibility reports with respect to the Task Force on Climate-related Financial Disclosures (TCFD) questions that guide sustainability reporting (Luccioni et al., 2020), Global Reporting Initiative (GRI) topics detection from CSR reports (Polignano et al., 2022), the development of a Word Embedding-based Inclusion Model (WEIM) in CSR reports (Lu et al., 2022), etc.

CSR-related topics have furthermore also been investigated in social media, and among others, deal with sentiment analysis of Environmental, Social and Governance (ESG)-related social media posts (Park et al., 2022), such as for example the detection of human rights on social media (Pilankar et al., 2022).

However, to the best of our knowledge, there are no publicly available datasets that would enable CSR theme detection or a more fine-grained CSR topic detection per theme. Furthermore, the fact that the majority of studies have also been conducted on English with limited experimentation on other languages, motivated us to set up a shared task for cross-lingual classification of corporate social responsibility (CSR) themes and topics.

3. Shared Task Setup

3.1. Pilot study

To assess the feasibility of the task, we conducted a pilot study, resulting in over 1,034 annotated news items in English, 54 items in Spanish, over 250 items in French, and 24 articles in simplified Chinese. CSR theme detection includes the classification of news into one of four CSR themes:

1. Environment (ENV), which deals with factors that affect the natural environment such as carbon emissions, natural resources, energy efficiency, waste management, and raw material sourcing.

2. Labor and Human Rights (LAB), discussing topics such as human rights, labor standards, diversity and inclusion or career management and training.

3. Fair Business Practices (FBP), reflecting on anti-competitive practices, corruption, and responsible information management.

4. Sustainable Procurement (SUP), which includes supplier environmental and social practices

The pilot study showed that the ENV and LAB themes were predominant followed by FBP and SUP for all four languages. The most frequent topics within the ENV theme were *Materials, Chemicals, & Waste*, and *Environmental Services & Ad-*

vocacy, while *Employee Health & Safety* and *Labor Practices and Human Rights* were the most reported topics within the LAB theme. An overview of the different topics in both ENV and LAB themes is given in Table 1. In the case of CSR topic detection, we observed that articles may be assigned two labels, while the assignment of three or more labels was less common.

Since 2001, the company has reduced its CO2 emissions rate by 52 percent, and it plans to continue this commitment by establishing a target to reduce the rate more than 65 percent by 2021.
"NextEra Energy is committed to creating a sustainable energy future and providing customers with electricity that is affordable, reliable and clean," said Jim Robo, chairman and chief executive officer of NextEra Energy. "We're one of the cleanest energy companies in America, and the world's largest generator of renewable energy from the wind and sun. We've been reducing emissions for decades through the development of renewable energy and modernizing our generation fleet. Through our significant investments in energy infrastructure, we're shaping how energy is produced and delivered, putting tens of thousands of Americans to work, providing significant economic benefits to the communities we serve and delivering value for our customers, employees and shareholders – all while protecting and conserving the environment."

Figure 1: An example of CSR news (**CSR theme**: ENV, **CSR topic**: ENERGY CONSUMPTION & GHG).

An example of CSR news for the Environment CSR theme is given in Fig. 1. The parts in yellow are indicated by the annotators as triggers for the chosen label, but are not part of the shared task.

3.2. Task Description

The shared task includes two sub-tasks:

- **Task A:** Cross-lingual CSR theme recognition (English, French, simplified Chinese): cross-lingual, multi-class classification task with the following labels: Environment (ENV), Labor and Human Rights (LAB), Fair Business Practices (FBP), Sustainable Procurement (SUP).
- **Task B:** Fine-grained multi-label classification of CSR topics (English) for Environment (ENV) and Labor and Human Rights (LAB) themes.

Task A is framed as a multi-class classification task, for which participants output for each news article in the different languages a CSR label. **Task B** is a multi-label classification problem whereby an article may be assigned multiple topics from Table 1 within the specified theme (e.g., an article with two topics, *Air Pollution* and *Customer Health and Safety*, within the ENV theme). While we encouraged participants to contribute to both sub-tasks, they could also decide to participate in **Task A** or **Task B** only.

CSR theme	topic
ENV	Air Pollution
	Biodiversity
	Customer Health & Safety
	Energy Consumption & GHGs
	Environmental Services & Advocacy
	Materials, Chemicals & Waste
	Product End-of-Life
	Product Use
	Water
LAB	Career Management & Training
	Child Labor, Forced Labor & Human Trafficking
	Diversity, Equity, and Inclusion
	Employee Health & Safety
	External Stakeholder Human Rights
	Labor Practices and Human Rights
	Social Dialogue
	Social Discrimination
	Working Conditions

Table 1: List of CSR topics for the ENV and LAB themes.

3.3. Dataset Construction for the Shared Task

We aimed at collecting and annotating at least 1,500 publicly available English news articles with CSR themes for the training set and at least 500 news items per LAB and ENV CSR theme. Articles covering the two largest themes (ENV, LAB) were annotated with underlying CSR topics to produce the dataset for **Task B**. The datasets for both sub-tasks were constructed from publicly available content. As no personal data was used, we did not anticipate risks with respect to ethics, privacy or security.

Dataset quality and annotators In line with the pilot study, dataset quality was ensured by engaging highly qualified CSR experts as annotators, monitoring inter-annotator agreement and resolving disagreements. Every document was independently annotated by two trained CSR analysts and disagreements were resolved through discussion in pairs to arrive at the final list of annotations (Oortwijn et al., 2021).

Annotation scheme For cross-lingual CSR theme recognition, the annotation was done at the news item level whereby each URL was classified into one of four CSR themes: ENV, LAB, FBP, or SUP. The subset of news items labeled with the ENV and LAB themes was subsequently further annotated into one or more CSR topics. The data set shared with participants included news item URLs and the corresponding labels.

3.4. Experimental data

The annotated data was split using stratified random sampling to build training and test sets for English. For the remaining languages in **Task A**, only test sets were made available. The label distribution for the training (English) and test data (English, French and Chinese) from Task A is presented in Table 2 and the corresponding figures for Task B are given in Table 3. Recall that Task B was set up as a multi-label classification task. When we consider the distribution of the labels across both themes (Figure 2), we can observe that one or two labels were assigned to the large majority of instances, whereas up to 10% of the instances received three or even more labels.

	TRAIN	TEST		
	English	English	French	Chinese
ENV	708	164	70	70
FBP	197	48	21	25
LAB	662	149	70	40
SUP	41	2	1	0
Total	1608	363	162	135

Table 2: Label distribution for Task A

ENV	TRAIN	TEST
Air pollution	36	6
Biodiversity	62	11
Customers Health and Safety	62	19
Energy Consumption, GHGs	366	80
Env. Services & Advocacy	242	79
Materials,Chemicals, Waste	112	32
Product End of Life	73	20
Product Use	44	7
Water	71	16
LAB	TRAIN	TEST
Career Mgmt & Training	77	18
Child Labor, Forced Labor, Human Trafficking	7	1
Diversity, Equity, Inclusion	149	35
Employee Health, Safety	138	37
Ext. Stakeh. Human Rights	14	3
Labor Pract. & Human Rights	47	24
Social Dialogue	52	14
Social Discrimination	18	5
Working Conditions	201	60

Table 3: Label distribution for Task B

4. Methodology of Participating Teams

For this shared task, two teams submitted results for both sub-tasks: Team Kosar & Van Nooten (Van Nooten et al., 2024) and Team TredenceAICoE (Sharma et al., 2024).

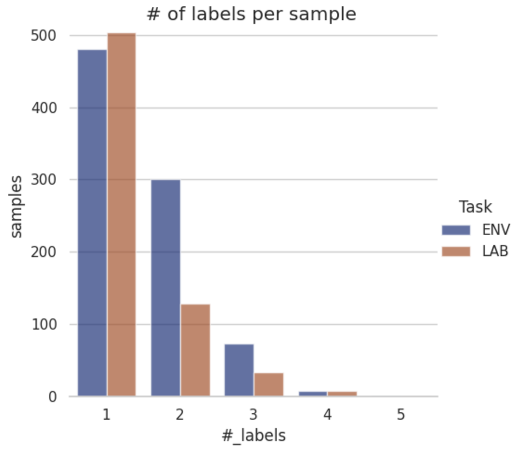


Figure 2: Number of samples with 1,2,3,4 or 5 labels for Task B (test and training data combined)

4.1. Data Collection, Cleaning and Augmentation

The two participating teams employed different libraries to scrape the content of the provided URLs. Kosar & Van Nooten used the Trafilatura library (Barbaresi, 2021), which was not able to scrape the content of all of the provided URLs. TredenceAICoE, on the other hand, used Newspaper3K[3] and was able to scrape not only the content of the pages but also the titles of the pages, which they leveraged as additional input information. After collecting the data, only Team Kosar & Van Nooten cleaned the data, using GPT 3.5[4].

Since the organizers only provided training data for English for both tasks, but Task A also involves testing on Chinese and French, both teams made use of data augmentation to obtain training and development data. Again, the two teams used different systems to translate the training data. Whilst Team Kosar & Van Nooten used the Google Translate API, Team TredenceAICoE used two language-specific transformer models to translate to French Helsinki-NLP/opus-mt-tc-big-en-fr (Tiedemann and Thottingal, 2020) and Chinese (Helsinki-NLP/opus-mt-en-zh).

In addition to generating data in the other languages, both teams used data augmentation to create additional samples to address class and label imbalance. Team TredenceAICoE used GPT 4[5] to generate samples for the minority classes and Team Kosar & Van Nooten used Mixtral[6] to

[3] https://github.com/codelucas/newspaper
[4] https://platform.openai.com/docs/models/overview
[5] https://platform.openai.com/docs/models/overview
[6] https://docs.together.ai/docs/

paraphrase each sample in the train set.

4.2. Methodologies

Both teams tested a wide variety of systems, with the most notable being zero-shot prompting with advanced prompts for GPT 3.5 and GPT 4 (Kosar & Van Nooten). Regardless, the best approach for both teams was still fine-tuning pre-trained transformer models.

For their final systems, Kosar & Van Nooten use the XLM-RoBERTa-large model for Task A and monolingual RoBERTa-large model for Task B (Conneau et al., 2020). Similarly,Team TredenceAICoE used MDeBERTa (He et al., 2023) for Task A (as it is multilingual) and Longformer (Beltagy et al., 2020) for Task B. As most scraped articles extend beyond the standard token length of 512, Longformer is a sensible model choice. However, this was not the only measure the team took to address the exceeding text length. They also divided the dataset into multiple sequences or chunks, and experimented with a Variable Selection Network (VSN) (Lim et al., 2021) to provide selected additional information to the classification layer for improved predictions.

In addition to creating augmented data to combat class imbalance, both teams also modified the training procedure. While Team TredenceAICoE made use of Dynamic Weighted Loss to increase the probabilities of underrepresented classes, Team Kosar & Van Nooten employed an advanced variation of contrastive learning that is specifically aimed at dealing with multi-label classification.

5. System Evaluation & Results

5.1. Evaluation

To evaluate system performance for Tasks A and B, the prediction of coarse-grained CSR themes and fine-grained CSR topics for environment and Labor and Human rights, we used standard evaluation measures, including accuracy, precision, recall and F1-score. The results are ranked according to weighted F1-score to account for the difference in sample sizes for English (363), Chinese (135) and French (162) and the different class distributions across the CSR themes and topics. However, in addition to the weighted F1-score, we also provide macro-averaged F1-scores to describe the performance on the labels with low sample counts. One of the participating teams (Kosar & Van Nooten) reported encountering issues while scraping the text for some of the test samples. As a result, they could not provide any predictions for 43 samples for Task A and 19 samples for Task B. Naturally, the results for Team Kosar & Van Nooten are better when only considering the samples they had access to. However, as the performance difference is small and does not impact the ranking for either of the tasks, we present the results on the complete test set. More concretely, if we assume the missing predictions are wrong for Task A, this results in a drop of 3% across all labels. For Task B, we assume that none of the labels are present in the prediction, resulting in a drop of 0.4% across all tasks and subsets for Team Kosar & Van Nooten.

inference-models

	team	acc.	prec.	rec.	f-m	f-w
EN	A-J	0.90	0.96	0.90	0.61	0.93
	TRED	**0.95**	**0.97**	**0.95**	**0.77**	**0.96**
ZH	A-J	0.58	0.67	0.58	0.35	0.60
	TRED	**0.78**	**0.88**	**0.78**	**0.61**	**0.81**
FR	A-J	0.82	0.93	0.82	0.65	0.87
	TRED	**0.94**	**0.95**	**0.94**	**0.87**	**0.94**
avg.	A-J	0.76	0.86	0.76	0.54	0.80
	TRED	**0.89**	**0.93**	**0.89**	**0.75**	**0.90**

Table 4: Summary of the results for Task A with detailed information on the performance per language. Precision (prec.), recall (rec.) and F1 scores (f-w) are weighted averages across all classes. To describe the performance on minority classes, we also show macro-averaged F1 (f-m).

	team	acc.	prec.	rec.	f1
ENV	A-J	0.76	0.91	0.75	0.82
	TRED	**0.89**	**0.97**	**0.86**	**0.91**
FBP	A-J	0.76	0.88	0.66	0.75
	TRED	**0.89**	**0.81**	**0.98**	**0.88**
LAB	A-J	0.76	0.77	0.84	0.79
	TRED	**0.89**	**0.95**	**0.89**	**0.91**
SUP	**A-J**	0.86	**0.35**	**0.75**	**0.48**
	TRED	**0.95**	0.32	**0.75**	0.44
avg.	A-J	0.76	0.85	0.76	0.80
	TRED	**0.89**	**0.93**	**0.89**	**0.90**

Table 5: Summary of the results for Task A with detailed information on the performance per label. The concluding row with averaged values (*avg.*) reports the weighted averaged F1-score.

5.2. Results for Task A

As shown in Table 5, Team TredenceAICoE attained the highest overall scores for Task A with a macro-averaged F1-score of 75% and a weighted F1-score of 90% across all labels. Except for the SUP category - for which the test data, depending on the language, merely contains between 0 and 2 instances -, their system consistently outperformed the system of their competitors on the sustainability labels (as illustrated in Figure 4), but mostly attained these increased scores by performing better on French and Chinese (illustrated in Figure 3).

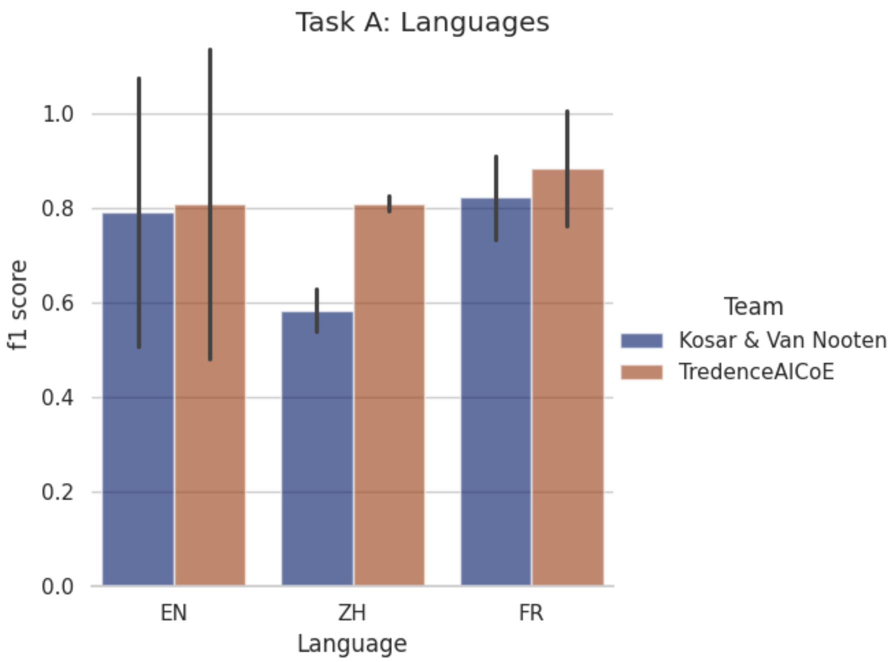


Figure 3: Weighted F1-scores on the test set for Task A (theme multi-class classification) per language. The vertical lines describe the standard deviation for the different labels.

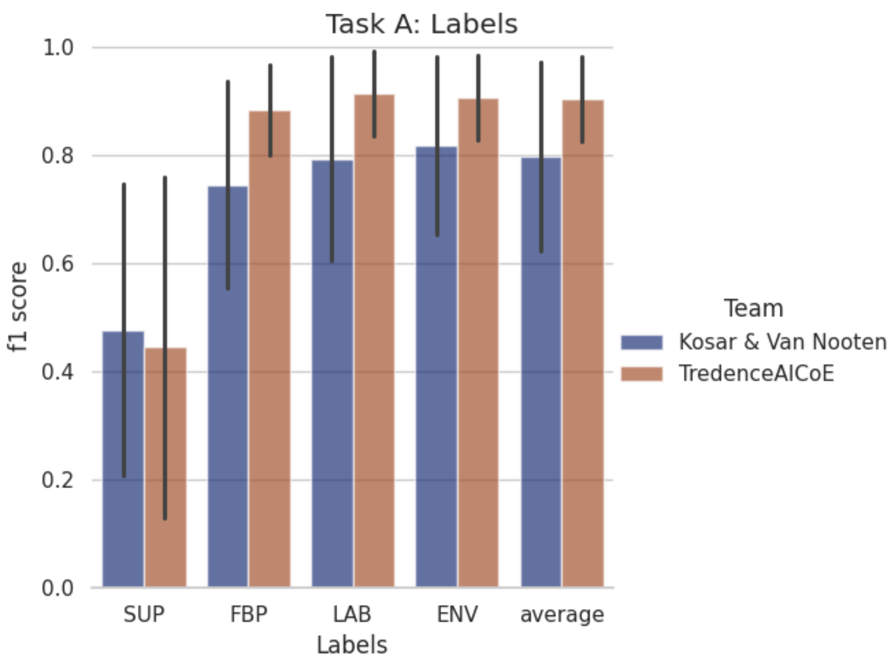


Figure 4: Weighted F1-scores on the test set for Task A (theme multi-class classification) for each individual label. The vertical lines describe the standard deviation for the different languages.

5.3. Results for Task B

For Task B, the fine-grained multi-label classification for ENV and LAB, Team Kosar & Van Nooten reached the highest overall weighted F1-score (*f-w*) of 88.1%. From their macro-averaged F1-score (*f-m*) of 73.8% (shown in Table 6), we could derive that across the two CSR themes, Team Kosar & Van Nooten is better at identifying the less prominent labels.

On the ENV sub-task, there is no significant difference in performance between the two teams, with weighted F1-scores of 87.3% vs 87.7% and an equal accuracy score of 87.5%. As shown in

	team	acc.	prec.	rec.	f-m	f-w
ENV	A-J	**0.875**	0.877	**0.875**	**0.745**	0.873
	TRED	**0.875**	**0.884**	**0.875**	0.725	**0.877**
LAB	**A-J**	**0.892**	0.888	**0.892**	**0.731**	**0.889**
	TRED	0.875	**0.901**	0.875	0.682	0.879
avg.	**A-J**	**0.884**	0.882	**0.884**	**0.738**	**0.881**
	TRED	0.875	**0.893**	0.875	0.704	0.878

Table 6: Summary of the results for Task B. Precision (prec.), Recall (rec.) and F1 scores (f-w) weighted averages across all labels. To describe the performance on minority classes, we also show macro-averaged F1 (f-m).

Figure 5, the teams had varying scores depending on the label, with Team Kosar & Van Nooten scoring 8% higher for “Biodiversity” and 10% lower for “Energy Consumption & GHGs”. However, along with the other labels, the score difference averaged out to 0.3% (weighted F1-score).

On the LAB sub-task, Team Kosar & Van Nooten did achieve a weighted F1-score of 88.9%, which is higher compared to the 87.9% score of the other team. On the LAB subset of Task B, the winning team scored 7.5% higher on the label for “Labor Practices and Human Rights”, while scoring 5% lower on the label for “Working Conditions” (illustrated in Figure 6). Along with some minor differences (both ups and downs) on the other labels, the notable performance difference on the label for “Labor Practices and Human Rights” seems to be the game-changer for Task B.

6. Conclusion

Both participating teams developed highly advanced systems that were directly modified to deal with the two specific sub-tasks. The modifications of Team Kosar & Van Nooten address class/label imbalance, cross-lingual transfer and multi-label co-occurence with data augmentation, machine translation and a special variant of contrastive learning for multi-label classification. Their competitors, Team TredenceAICoE, also employed data augmentation to create additional samples for the unseen languages and underrepresented classes. However, their efforts specifically address the exceeding text length of the articles using a Variable Selection Network for chunking.

For Task A, TredenceAICoE attained the highest score across all three classes. Their fine-tuned MDeBERTa most notably outperformed the other system on the Chinese subset of the data. Likely, their choice of using a transformer model for translation aided them in exceeding the results of the other team, who used the Google Translate API for translation.

For Task B, the advanced multi-label approach of Kosar & Van Nooten with their particular variant of contrastive learning allowed them to beat the performance of their opponents on the LAB subset of Task B.

7. Acknowledgments

The authors thank Clay Liu, Yi-Shiuan Iza Sun, Manisha Babubudjnauth, Ravi Ramruttun, Hans Lachanah, Viduchy Goorvadoo, Swastee Pomny for making the data sets available. This work was supported by Ghent University under grant BOF.24Y.2021.0019.01

8. Bibliographical References

Adrien Barbaresi. 2021. Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131, Online. Association for Computational Linguistics.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Andrew Crane and Sarah Glozer. 2016. Researching corporate social responsibility communication: Themes, opportunities and challenges. *Journal of Management Studies*, 53:7.

Sebastian G.M. Händschke, Sven Buechel, Jan Goldenstein, Philipp Poschmann, Tinghui Duan, Peter Walgenbach, and Udo Hahn. 2018. A corpus of corporate annual and social responsibility

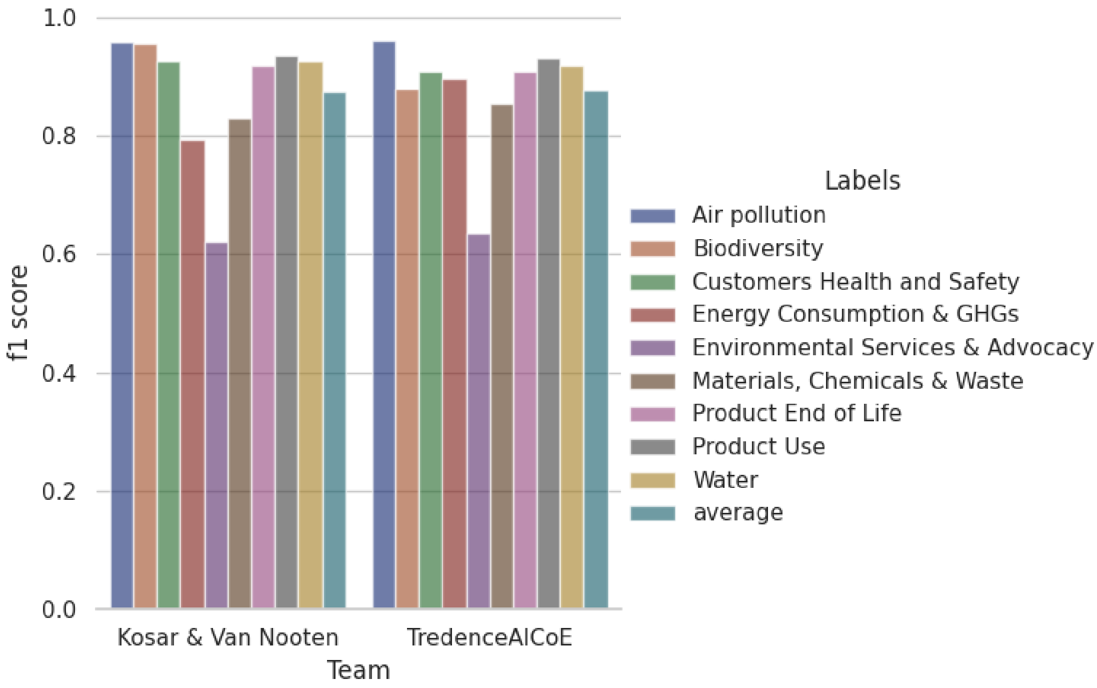


Figure 5: Weighted F1-scores on the test set for Task B (multi-label classification) for each individual label in the ENV subset.

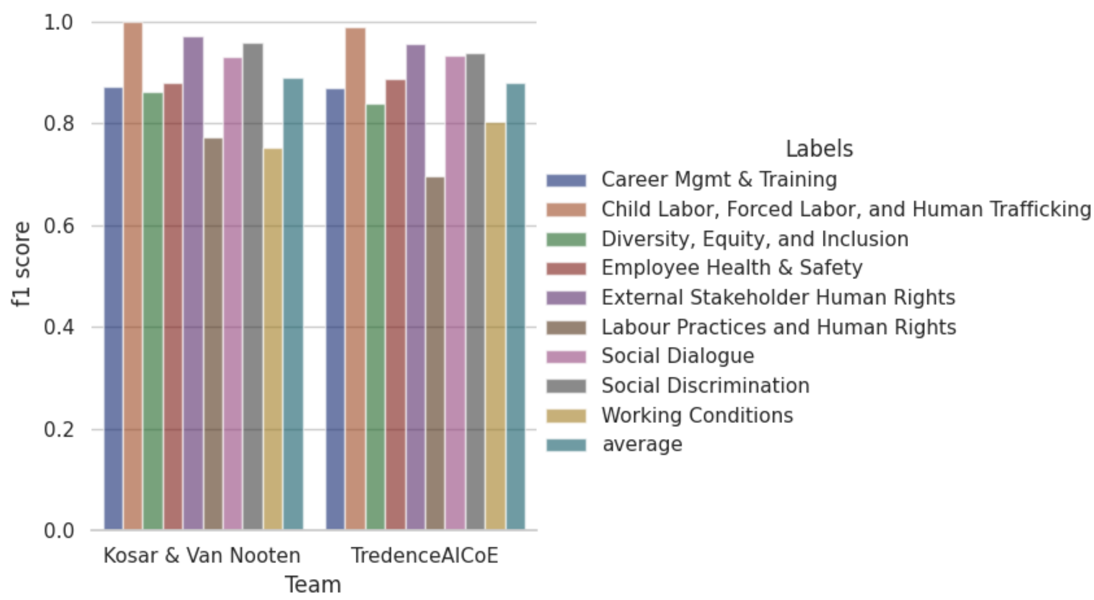


Figure 6: Weighted F1-scores on the test set for Task B (multi-label classification) for each individual label in the LAB subset.

reports: 280 million tokens of balanced organizational writing. In *Proceedings of the First Workshop on Economics and Natural Language Processing*, pages 20–31, Melbourne, Australia. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Qian Li, Hao Peng, Jianxin Li, Congyin Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2020. A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13:1–41.

Bryan Lim, Sercan Ö. Arık, Nicolas Loeff, and Tomas Pfister. 2021. Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Nankai Lin, Sihui Fu, Xiaotian Lin, and Lianxi Wang. 2022. Multi-label emotion classification based on adversarial multi-task learning. *Information Processing and Management*, 59(6).

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, Toronto, Canada. Association for Computational Linguistics.

Lu Lu, Jinghang Gu, and Chu-Ren Huang. 2022. Inclusion in CSR reports: The lens from a data-driven machine learning model. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 46–51, Marseille, France. European Language Resources Association.

Alexandra Luccioni, Emily Baylor, and Nicolas Duchene. 2020. Analyzing Sustainability Reports Using Natural Language Processing. *CoRR*, abs/2011.08073.

Yvette Oortwijn, Thijs Ossenkoppele, and Arianna Betti. 2021. Interrater Disagreement Resolution: A Systematic Procedure to Reach Consensus in Annotation Tasks. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 131–141, Online. Association for Computational Linguistics.

Joonbeom Park, Woojoo Choi, and Sang-Uk Jung. 2022. Exploring Trends in Environmental, Social, and Governance Themes and their Sentimental Value over Time. *Frontiers in Psychology*, 13.

Yash Pilankar, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes, and Pramod Pathak. 2022. Detecting violation of human rights via social media. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 40–45, Marseille, France. European Language Resources Association.

Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. 2022. An NLP Approach for the Analysis of Global Reporting Initiative Indexes from Corporate Sustainability Reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 1–8, Marseille, France. European Language Resources Association.

Matthew Purver, Matej Martinc, Riste Ichev, Igor Lončarski, Katarina Sitar Šuštar, Aljoša Valentinčič, and Senja Pollak. 2022. Tracking changes in ESG representation: Initial investigations in UK annual reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pages 9–14, Marseille, France. European Language Resources Association.

Shubham Sharma, Himanshu Janbandhu, and Ankush Chopra. 2024. Improving Cross-Lingual CSR Classification using Pretrained Transformers with Variable Selection Networks and Data Augmentation . In *Proceedings of the Joint Workshop of FinNLP-KDF-ECONLP@LREC-COLING 2024*, Torino, Italy.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Jens Van Nooten, Andriy Kosar, Guy De Pauw, and Walter Daelemans. 2024. Advancing CSR Theme and Topic Classification: LLMs and Training Enhancement Insights. In *Proceedings of the Joint Workshop of FinNLP-KDF-ECONLP@LREC-COLING 2024*, Torino, Italy.

Mingyu Wan and Chu-Ren Huang, editors. 2022. *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

9. Language Resource References

Advancing CSR Theme and Topic Classification: LLMs and Training Enhancement Insights

Jens Van Nooten[*], Andriy Kosar[|*], Guy De Pauw[**], Walter Daelemans[*]**
[*]University of Antwerp (CLiPS), [**]Textgain
Antwerp, Belgium
{jens.vannooten, walter.daelemans}@uantwerpen.be
{andrew, guy}@textgain.com

Abstract

In this paper, we present our results of the classification of Corporate Social Responsibility (CSR) Themes and Topics shared task, which encompasses cross-lingual multi-class classification and monolingual multi-label classification. We examine the performance of multiple machine learning (ML) models, ranging from classical models to pre-trained large language models (LLMs), and assess the effectiveness of Data Augmentation (DA), Data Translation (DT), and Contrastive Learning (CL). We find that state-of-the-art generative LLMs in a zero-shot setup still fall behind on more complex classification tasks compared to fine-tuning local models with enhanced datasets and additional training objectives. Our work provides a wide array of comparisons and highlights the relevance of utilizing smaller language models for more complex classification tasks.

Keywords: multi-class classification, multi-label classification, cross-lingual classification, CSR

1. Introduction

The landscape of Corporate Social Responsibility (CSR) is increasingly becoming a pivotal aspect of how businesses operate and are perceived in the global market (Wen and Deltas, 2022). Significant regulations have been instrumental in shaping the CSR framework. For a comprehensive history of CSR regulation, consult Wen and Deltas (2022).

These regulations have increased the liability of companies regarding sustainability non-compliance, making it imperative for them to not only be aware of but also manage and anticipate such issues effectively. However, even with mandatory or voluntary reporting, not all pertinent information is disclosed or reported and consequently leveraged for company evaluation due to CSR-related information being scattered across different media sources, languages and formats. This leads to challenges in its identification and analysis. As a result, there is a critical need for efficient methods to detect and classify this diverse information in order to reinforce corporate compliance and enhance stakeholder decision-making.

In response to this growing need and interest in processing and analyzing CSR content, our study addresses the complexities of detecting and classifying CSR content through participation in the "Cross-lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics" shared task (Nayekoo et al., 2024). The task facilitates cross-lingual CSR theme detection and fine-grained topic classification, specifically targeting the Environment (ENV) and Labour and Human Rights (LAB) themes across English, French, and simplified Chinese. The theme classification is approached as a multi-class problem, and the topic classification within these themes is framed as a multilabel classification task. Our evaluation extends to various text representations and ML models, encompassing both traditional approaches and Large Language Models (LLMs), utilizing pretrained models for ZS classification and Fine-tuning. Additionally, we explore the potential of enhancement techniques like Data Augmentation (DA) and Contrastive Learning (CL) to improve performance.

In the following sections, we delve into the methodology employed in our study, the experimental setup, the results and analysis of our findings, and the implications of our research for the field of CSR content processing and classification.

2. Previous Work

Text Classification The field of text classification, encompassing both multi-class and multi-label types, has experienced significant evolution over the past decade. This evolution has been particularly notable in three key areas: model types, text representation, and training methods. The advent of LLMs, starting with BERT, has transformed the landscape by introducing advanced model architectures, enhancing text representation through context-aware embeddings, and pioneering efficient training methodologies that leverage pre-trained models for fine-tuning or even enable zero-shot learning capabilities. For an overview of the diverse approaches and developments in multi-

These authors contributed equally to this work.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 292–305
20 May, 2024.

class and multi-label text classification, we refer to the comprehensive surveys conducted by Li et al. (2022), Gasparetto et al. (2022), Chen et al. (2022) and Bogatinovski et al. (2022), which cover both existing approaches and the latest advancements.

NLP for CSR In addition to the broad advancements in text classification, there has not been much research conducted on applying these techniques to the Corporate Social Responsibility (CSR) domain, with a few exceptions. Most of the work was conducted for the automatic analysis of Corporate Sustainability Reports: Shahi et al. (2011, 2014) applied multi-label text classification to classify reports according to the Global Reporting Initiative Index. Castellanos et al. (2015) applied neural networks, decision trees, and a memory-based learning algorithm, to classify parts of the report according to the five dimensions of the Sustainability Accounting Standards Board.

CSR has recently attracted more attention in the context of NLP, exemplified by the First Computing Social Responsibility Workshop (CSR-NLP I 2022) (Wan and Huang, 2022). However, to the best of our knowledge, there has been limited progress in classifying publicly accessible information on the internet across diverse textual genres, including but not limited to news articles, company briefs/newsletters, and industry reports.

LLMs for Text Classification LLMs have been used widely in the field of text classification ever since the advent of BERT. In the following years, countless new models have been released that yield state-of-the-art results on a multitude of classification and generation tasks, such as LLAMA (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Gemini (Team et al., 2023) and GPT (Radford et al., 2018). Recently, GPT-3.5 (Ouyang et al., 2022) and GPT-4 (Achiam et al., 2023) have sparked the interest of many researchers, leading to a great deal of work being dedicated to their applications for classification, besides generation. For a more comprehensive overview, consult Minaee et al. (2024). As evidenced in Peskine et al. (2023) and De Langhe et al. (2024), prompting generative models for more complex classification tasks such as multi-label classification can be quite challenging, leading to inferior performance compared to fine-tuned encoders.

Data Augmentation To address the issue of data imbalance and data scarcity, multiple data augmentation techniques have been leveraged, including generating synthetic data with state-of-the-art GPT models (Van Nooten and Daelemans, 2023; Sufi, 2024; Kumar et al., 2020; Zhang et al., 2020), reaching superior performance compared to other data augmentation methods.

Contrastive Learning Contrastive Learning aims to maximize the distance between dissimilar texts and minimize the distance between similar pairs in the embeddings space. Some studies explore contrastive losses (CL) for multi-class classification (Pan et al., 2022) and multi-label classification (U et al., 2023; Lin et al., 2023) using variants of Supervised Contrastive Loss (SCL) (Khosla et al., 2021) or NT-XENT (Sohn, 2016).

In this work, we aim to provide a wide range of baselines for multi-class and multi-label CSR text classification. We hypothesise that fine-tuning smaller language models can outperform more recent generative LLMs for more complex classification tasks. Moreover, we also hypothesise that CL can further improve performance and that generative LLMs produce useful synthetic data to further enhance performance of classification models, as previous work indicates.

3. Datasets

Shared Task The shared task is divided into two subtasks: cross-lingual, multi-class classification for CSR theme recognition (one dataset) and monolingual multi-label text classification (two datasets) of CSR topics for Environment (ENV) and Labour and Human Rights (LAB) themes. These datasets comprise lists of URLs for English texts, each associated with relevant labels. Table 1 provides statistics for each dataset.

Data Collection and Cleaning The texts in the training dataset were scraped using the Trafilatura library (Barbaresi, 2021). URLs that could not be successfully scraped were excluded from the training dataset. Given that a significant portion of the data contained artifacts potentially detrimental to the training of the models (such as URLs, external links or other irrelevant text), we employed GPT-3.5[1] for data cleaning. The resulting cleaned texts were checked manually. After cleaning the data and removing duplicates, 675 of 699 texts remained. The specific prompt that was used is described in Appendix B.1. The test data were scraped using the Boilerpy library[2], following the organizers' recommendation.

4. Methodology

Classification Models In our study, we evaluated the performance of a wide range of classifi-

[1] *gpt-3.5-turbo-0613*, `https://platform.openai.com/docs/models/gpt-3-5-turbo`

[2] `https://pypi.org/project/boilerpy3/`

Dataset	type	n classes	labels per text	n train	n test
Themes	Multi-class	4	1	1,515	618
ENV	multi-label	9	1.53	675	157
LAB	multi-label	9	1.35	500	149

Table 1: Datasets' statistics.

cation models, which differ significantly in terms of model complexity, to identify those most suitable for tasks with limited training data. As a baseline, we chose the SVM model combined with TF-IDF and OpenAI text embeddings. Additionally, we included Zero-shot (ZS) text classification with GPT-3.5 and GPT-4 models (Radford et al., 2018) as baselines. Examples of prompts for ZS text classification are provided in Appendix B.3 and B.4.

We further expand our model repertoire by incorporating models with more complex architectures, specifically, the Multi-Layer Perceptron (MLP) and an LSTM, while maintaining the same text representation strategies to ensure a consistent comparison basis. Detailed information on the optimal hyperparameters identified for these models can be found in Appendix D.

For the themes dataset, we utilized multilingual pre-trained language models, such as Multi-Lingual DistilBERT[3] (Sanh et al., 2020), XLM-RoBERTa, and XLM-RoBERTa-large[4] (Conneau et al., 2020). For the multi-label datasets, we employed DistilBERT, BERT (Devlin et al., 2019), RoBERTa, and RoBERTa-large (Liu et al., 2019). All models were trained with a batch size of 8 and 2 gradient accumulation steps. Where appropriate, we repeated each experiment with three random seeds. The optimal hyperparameters for all models are detailed in Appendix D.

Data Paraphrasing and Translation The training data for each task was expanded using various methods. Every entry in the datasets was paraphrased using Mixtral[5], effectively doubling the size of the training data. A detailed description of the prompt used can be found in Appendix B.2. Additionally, for the cross-lingual multi-class task, we opted for translating the data to French and simplified Chinese using the Google Translate API[6]. These languages were selected because the model was to be tested on them. By translating the data into these two additional languages, the size of the training data was tripled. The synthetic data was incorporated into the training dataset.

[3] *distilbert/distilbert-base-multilingual-cased*

[4] *FacebookAI/xlm-roberta-(base/large)*

[5] *mistralai/Mixtral-8x7B-Instruct-v0.1*, https://docs.together.ai/docs/inference-models

[6] https://cloud.google.com/translate/docs/basic/translating-text

Contrastive Learning We train additional models with a contrastive loss. For the multi-label variant, we follow Lin et al. (2023) and select positive and negative in-batch samples for a given anchor by calculating the Jaccard Index (JI) between binary label vector pairs. If JI is greater than the threshold hyperparameter, the sample is considered a positive. To allow CL to work with relatively small batches, all possible pair combinations are constructed in a batch to maximize the information gained from the contrastive loss.

We devise a variant of NT-Xent (Sohn, 2016) that allows for multiple positives to be taken into account per batch. In essence, we calculate the Binary Cross-Entropy loss between two vectors with length n, where n is the number of possible combinations in a batch: vector α, which is a binary vector that denotes whether a pair is positive or negative, and vector β, which is a vector that contains the cosine distance between the two in-batch samples in a pair (cf. Eq. 1 and 2 in Appendix C). The goal is to minimize the cosine distance between samples in positive pairs and maximize the distance between samples in negative pairs, which leads to a decrease in BCE. The resulting loss is then weighted and added to the classification loss.

Evaluation All models are evaluated in a five-fold stratified cross-validation setup. To stratify the multi-label splits, we employ the strategy described in Sechidis et al. (2011). The corresponding paraphrases and translations of a certain training fold were added during training so no indirect data leakage would occur. As evaluation metrics, micro-averaged and macro-averaged F1 are used.

5. Results

Subtask A The cross-validation results for Subtask A are summarised in Table 2[7]. It can be observed that smaller models with less complex training methods, such as keyword-based learning with SVM, already achieve respectable results[8], though the larger models and more complex models generally achieve the best results. Both CL and adding translations to the training data generally yields improvements in terms of macro- and micro-averaged F1 performance. However, data translation alone yields the best results, which is especially beneficial for learning the minority class "SUP".

Interestingly, we observe that the ZS experiments with GPT-3.5 and GPT-4 yield inferior results, thus indicating that the fine-tuned models benefit from learning the text-specific features during training.

[7] Confusion matrices for all models and datasets can be found in Appendix F.

[8] These models are solely trained on English data and are not to be deployed on the multilingual test set.

Th.	tf-idf + SVM	ada-003 + SVM	ada-003 + MLP	ada-003 + LSTM	GPT-3	GPT-4	DB	DB + CL	DB + DA	DB + CL + DA	BERT	BERT + CL	BERT + DA	BERT + CL + DA	RB-Lg	RB-Lg + CL	RB-Lg + DA	RB-Lg + CL + DA
ENV	94.59 (± 0.88)	94.59 (± 0.88)	95.93 (± 0.71)	96.07 (± 0.58)	92.44 (± 0.79)	93.69 (± 1.6)	94.4 (± 1.4)	94.54 (± 1.62)	95.42 (± 0.98)	95.73 (± 0.89)	95.34 (± 1.39)	95.62 (± 1.33)	96.45 (± 0.96)	96.46 (± 0.89)	95.58 (± 1.1)	95.28 (± 2.65)	97.01 (± 0.63)	97.06 (± 0.77)
FBP	85.17 (± 4.01)	85.17 (± 4.01)	91.02 (± 3.86)	91.88 (± 2.64)	74.76 (± 3.78)	81.52 (± 5.18)	79.04 (± 6.63)	77.01 (± 7.24)	86.8 (± 3.62)	87.61 (± 2.39)	87.07 (± 4.97)	89.07 (± 5.97)	91.43 (± 3.58)	92.03 (± 3.28)	84.0 (± 31.26)	82.9 (± 36.58)	94.25 (± 2.1)	94.37 (± 1.98)
LAB	94.42 (± 1.66)	94.42 (± 1.66)	96.12 (± 0.88)	96.4 (± 0.72)	90.23 (± 1.5)	94.02 (± 0.98)	93.7 (± 1.45)	93.73 (± 1.48)	94.58 (± 0.76)	95.05 (± 0.8)	95.77 (± 0.69)	96.08 (± 0.73)	96.43 (± 1.09)	96.28 (± 1.26)	94.94 (± 3.85)	95.25 (± 3.66)	97.15 (± 0.88)	96.98 (± 0.71)
SUP	52.5 (± 23.09)	52.5 (± 23.09)	56.92 (± 18.8)	67.06 (± 17.62)	42.1 (± 23.49)	47.72 (± 15.69)	0.0 (± 0.0)	0.0 (± 0.0)	59.89 (± 24.61)	58.29 (± 15.72)	0.0 (± 0.0)	0.0 (± 0.0)	67.0 (± 13.42)	61.54 (± 10.27)	0.0 (± 0.0)	2.0 (± 7.13)	72.54 (± 13.49)	70.71 (± 15.06)
mic	92.81 (± 1.28)	92.81 (± 1.28)	94.72 (± 0.93)	95.16 (± 0.55)	88.09 (± 0.93)	90.96 (± 1.77)	91.27 (± 1.45)	91.11 (± 1.55)	93.36 (± 0.96)	93.64 (± 0.91)	93.47 (± 1.35)	93.97 (± 1.46)	95.18 (± 1.27)	95.05 (± 1.36)	92.96 (± 3.8)	92.87 (± 4.78)	96.22 (± 0.83)	96.13 (± 0.78)
mac	81.67 (± 6.27)	81.67 (± 6.27)	85.0 (± 4.98)	87.85 (± 4.36)	74.88 (± 5.06)	79.24 (± 2.36)	66.79 (± 2.04)	66.32 (± 2.23)	84.17 (± 6.16)	84.17 (± 3.94)	69.55 (± 1.62)	70.19 (± 1.83)	87.83 (± 3.59)	86.58 (± 3.24)	68.63 (± 8.8)	68.86 (± 10.96)	90.24 (± 3.45)	89.78 (± 3.63)

ENV																		
0	47.62 (± 13.88)	47.97 (± 20.94)	49.86 (± 17.46)	50.91 (± 16.13)	16.77 (± 5.14)	23.03 (± 5.35)	52.07 (± 19.2)	51.85 (± 24.23)	57.14 (± 20.91)	56.65 (± 20.04)	53.93 (± 16.06)	52.57 (± 20.2)	61.02 (± 16.66)	56.41 (± 12.48)	57.78 (± 13.44)	65.99 (± 15.79)	60.0 (± 9.82)	62.83 (± 9.43)
1	40.97 (± 11.46)	56.96 (± 7.34)	57.66 (± 3.61)	55.29 (± 7.31)	54.07 (± 7.34)	54.69 (± 6.41)	55.42 (± 8.52)	58.58 (± 10.3)	56.44 (± 13.75)	63.28 (± 11.69)	56.19 (± 11.46)	58.36 (± 8.96)	58.93 (± 12.66)	60.34 (± 9.02)	63.0 (± 10.05)	66.47 (± 9.59)	60.31 (± 10.01)	64.46 (± 11.25)
2	49.39 (± 14.21)	71.01 (± 16.22)	75.77 (± 14.38)	74.27 (± 12.1)	36.84 (± 10.9)	59.41 (± 10.04)	63.79 (± 12.61)	66.0 (± 16.33)	67.09 (± 10.97)	66.23 (± 15.39)	62.63 (± 16.61)	64.95 (± 13.24)	64.67 (± 15.9)	64.88 (± 16.22)	75.57 (± 7.73)	78.47 (± 6.6)	73.93 (± 9.23)	77.78 (± 5.06)
3	79.1 (± 2.62)	80.72 (± 1.34)	81.44 (± 2.24)	82.4 (± 1.63)	80.77 (± 3.39)	84.55 (± 3.84)	83.77 (± 3.51)	83.45 (± 2.95)	82.62 (± 2.98)	83.22 (± 3.12)	83.57 (± 2.38)	83.43 (± 2.35)	82.78 (± 2.65)	82.83 (± 1.98)	83.74 (± 3.59)	85.71 (± 2.41)	84.25 (± 3.36)	84.4 (± 2.43)
4	42.37 (± 4.25)	47.02 (± 6.69)	47.37 (± 6.84)	49.29 (± 6.49)	26.64 (± 6.58)	53.65 (± 3.19)	50.04 (± 6.48)	51.52 (± 4.1)	51.53 (± 4.33)	53.54 (± 2.97)	52.57 (± 6.2)	51.8 (± 4.78)	54.94 (± 4.41)	54.37 (± 4.07)	50.48 (± 6.25)	53.33 (± 3.34)	55.52 (± 5.34)	55.09 (± 5.05)
5	44.92 (± 6.27)	49.42 (± 11.34)	49.02 (± 8.64)	51.32 (± 10.25)	30.68 (± 4.76)	37.84 (± 4.96)	62.59 (± 9.21)	64.14 (± 6.91)	61.43 (± 10.0)	59.17 (± 9.02)	58.22 (± 11.28)	59.44 (± 7.96)	55.51 (± 7.94)	55.81 (± 8.23)	60.76 (± 7.94)	56.2 (± 8.81)	59.1 (± 9.56)	58.77 (± 11.03)
6	62.5 (± 8.63)	60.23 (± 8.87)	62.54 (± 4.75)	65.21 (± 5.52)	18.75 (± 13.37)	54.11 (± 7.32)	66.67 (± 8.74)	67.46 (± 10.21)	68.59 (± 8.41)	66.04 (± 8.96)	64.3 (± 6.79)	65.21 (± 9.37)	63.11 (± 9.51)	62.87 (± 11.18)	69.14 (± 6.99)	71.0 (± 6.7)	70.78 (± 4.94)	68.57 (± 6.92)
7	7.27 (± 8.91)	13.16 (± 12.49)	13.27 (± 6.87)	1.59 (± 5.54)	5.21 (± 6.43)	28.41 (± 4.99)	0.0 (± 0.0)	1.57 (± 5.54)	17.34 (± 9.16)	10.87 (± 11.01)	4.51 (± 7.68)	11.19 (± 12.88)	16.83 (± 9.88)	18.18 (± 9.67)	1.55 (± 4.54)	4.14 (± 7.86)	22.34 (± 13.47)	19.57 (± 10.65)
8	61.07 (± 6.49)	60.44 (± 3.41)	65.36 (± 7.77)	71.17 (± 8.65)	37.02 (± 15.8)	61.51 (± 7.65)	77.57 (± 8.5)	77.75 (± 7.68)	78.01 (± 8.15)	77.37 (± 9.01)	77.09 (± 7.22)	77.66 (± 8.71)	79.08 (± 7.92)	77.44 (± 5.91)	78.76 (± 5.33)	77.42 (± 4.8)	75.66 (± 5.84)	75.86 (± 6.52)
mic	58.53 (± 0.62)	62.22 (± 1.02)	63.55 (± 0.94)	64.96 (± 0.89)	43.6 (± 1.19)	56.12 (± 1.53)	66.55 (± 2.28)	67.16 (± 2.55)	66.29 (± 2.59)	66.84 (± 2.23)	66.28 (± 2.15)	66.51 (± 2.47)	66.24 (± 2.05)	66.24 (± 1.75)	67.85 (± 2.21)	69.19 (± 1.85)	68.66 (± 3.17)	68.85 (± 2.43)
mac	48.36 (± 1.93)	54.1 (± 1.74)	55.81 (± 1.43)	55.71 (± 1.72)	34.08 (± 1.34)	50.8 (± 1.52)	56.88 (± 3.55)	58.04 (± 4.28)	60.02 (± 3.55)	59.6 (± 3.79)	57.0 (± 3.46)	58.29 (± 3.76)	59.65 (± 4.06)	59.24 (± 3.2)	60.09 (± 2.4)	62.08 (± 2.36)	62.43 (± 3.77)	63.04 (± 3.03)

LAB																		
0	22.33 (± 15.79)	39.97 (± 12.38)	42.4 (± 9.98)	46.98 (± 9.51)	26.85 (± 7.38)	21.62 (± 4.13)	47.18 (± 12.43)	43.04 (± 12.28)	46.61 (± 15.42)	48.65 (± 11.72)	54.26 (± 12.79)	52.73 (± 12.25)	52.02 (± 12.63)	57.22 (± 14.71)	57.39 (± 9.02)	62.69 (± 8.95)	58.64 (± 7.22)	59.36 (± 12.0)
1	0.0 (± 0.0)	20.0 (± 40.0)	60.0 (± 32.66)	0.0 (± 0.0)	4.87 (± 0.57)	13.33 (± 26.67)	0.0 (± 0.0)	0.0 (± 0.0)	12.5 (± 24.94)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	21.05 (± 33.99)	30.0 (± 40.0)
2	72.61 (± 2.64)	69.84 (± 5.78)	72.7 (± 5.9)	74.96 (± 5.18)	53.74 (± 9.78)	70.01 (± 3.1)	75.78 (± 6.0)	77.93 (± 4.81)	77.99 (± 3.62)	77.17 (± 3.77)	78.28 (± 3.8)	77.41 (± 3.47)	75.89 (± 6.69)	77.61 (± 5.01)	77.34 (± 2.89)	79.22 (± 2.87)	79.78 (± 4.14)	77.82 (± 3.46)
3	66.66 (± 3.46)	73.57 (± 5.03)	75.6 (± 2.04)	80.36 (± 3.39)	56.72 (± 4.34)	76.46 (± 3.47)	81.75 (± 3.57)	80.72 (± 5.72)	81.32 (± 4.65)	82.04 (± 4.24)	83.02 (± 3.99)	80.71 (± 2.99)	82.48 (± 4.03)	81.55 (± 3.3)	85.76 (± 3.93)	84.83 (± 3.5)	85.63 (± 4.94)	85.04 (± 5.14)
4	0.0 (± 0.0)	13.33 (± 26.67)	68.0 (± 19.04)	0.0 (± 0.0)	16.97 (± 16.49)	47.05 (± 12.6)	23.53 (± 29.48)	33.33 (± 32.66)	55.81 (± 31.89)	66.67 (± 19.12)	6.45 (± 16.63)	6.45 (± 16.63)	42.11 (± 33.26)	46.15 (± 36.24)	23.53 (± 29.48)	0.0 (± 0.0)	28.57 (± 31.43)	42.11 (± 33.26)
5	25.33 (± 17.71)	19.28 (± 11.35)	12.87 (± 11.71)	20.98 (± 15.9)	20.2 (± 2.45)	18.88 (± 1.39)	3.15 (± 6.5)	5.76 (± 9.77)	11.11 (± 11.99)	25.29 (± 12.99)	12.08 (± 12.06)	10.6 (± 10.16)	23.96 (± 14.08)	24.86 (± 16.78)	15.28 (± 14.07)	7.75 (± 10.03)	25.29 (± 15.32)	23.46 (± 13.24)
6	48.28 (± 18.2)	56.03 (± 15.74)	57.74 (± 14.75)	55.71 (± 14.23)	28.81 (± 15.34)	40.89 (± 6.38)	59.68 (± 9.6)	58.87 (± 9.15)	55.97 (± 6.77)	53.94 (± 10.21)	58.68 (± 6.1)	63.24 (± 9.91)	56.3 (± 9.68)	57.03 (± 7.81)	61.21 (± 10.23)	65.81 (± 10.24)	61.41 (± 10.97)	60.0 (± 9.05)
7	38.0 (± 19.39)	10.0 (± 20.0)	10.0 (± 20.0)	0.0 (± 0.0)	31.02 (± 8.76)	24.26 (± 5.95)	0.0 (± 0.0)	0.0 (± 0.0)	0.0 (± 0.0)	4.0 (± 12.47)	0.0 (± 0.0)	0.0 (± 0.0)	17.54 (± 22.74)	11.32 (± 18.79)	0.0 (± 0.0)	0.0 (± 0.0)	8.0 (± 17.0)	11.54 (± 20.0)
8	60.92 (± 4.07)	74.64 (± 4.94)	76.5 (± 3.78)	80.64 (± 3.5)	57.34 (± 4.02)	64.71 (± 2.46)	74.18 (± 4.89)	72.41 (± 4.34)	71.86 (± 8.45)	71.94 (± 5.21)	75.81 (± 4.72)	76.05 (± 4.11)	75.64 (± 5.62)	76.88 (± 3.77)	81.1 (± 3.32)	82.66 (± 3.83)	79.27 (± 5.29)	78.71 (± 5.42)
mic	57.15 (± 2.82)	63.44 (± 3.47)	65.84 (± 2.87)	69.22 (± 2.34)	36.16 (± 1.76)	47.68 (± 1.58)	67.53 (± 2.93)	66.73 (± 2.03)	66.85 (± 2.65)	67.67 (± 2.48)	69.28 (± 2.45)	68.84 (± 1.89)	68.53 (± 2.96)	69.99 (± 2.69)	72.13 (± 2.16)	73.7 (± 1.62)	72.26 (± 2.13)	71.59 (± 3.59)
mac	37.13 (± 4.8)	41.85 (± 7.22)	52.87 (± 7.49)	39.96 (± 2.26)	32.95 (± 1.94)	41.91 (± 3.57)	40.58 (± 4.35)	41.34 (± 4.52)	45.91 (± 5.67)	47.74 (± 3.86)	40.95 (± 2.47)	40.8 (± 1.97)	47.33 (± 5.76)	48.07 (± 5.13)	44.62 (± 4.83)	42.55 (± 2.01)	49.74 (± 5.57)	52.0 (± 6.61

Table 2: Mean results (F1) and standard deviations across folds and random seeds (if applicable) on the Themes (Th.), ENV and LAB datasets respectively. DB = (multilingual) DistilBERT, RB = (XLM-)RoBERTa. Red = worst score across models, green = best score across models. Consult Appendix E for a label index - label name mapping.

Subtask B The cross-validation results for Subtask B are summarised in Table 2. We observe that the tf-idf approach yields the worst results and that the larger models yield the best results. Additionally, we observe that Contrastive Learning and Data Augmentation generally yield improvements for each base model, indicating that the better separation between class-wise instances in the embedding space is beneficial for learning the task. Moreover, the added paraphrases aid the models especially in predicting uncommon classes. For both of these methods, an increase in true positives, but also false positives is observed across several models. The best macro-averaged results on the ENV dataset are achieved when a combination of the two is used with RoBERTa-large, while the best micro-averaged performance is achieved by training the model with CL.

The LAB dataset is more challenging to classify, as evidenced by the relatively lower scores. We found that RoBERTa-large trained with CL yielded the best micro-averaged performance. However, this model fails to predict some infrequent classes (*Child Labor*, *Ext. Stakeh. Human Rights* and *Soc.l Discr.*), as opposed to the ada-003 + MLP model or models trained with extra data, which each yield a superior macro-averaged performance.

Similar to the results from Subtask A, we observe that the generative models underperform compared to fine-tuned language models on both datasets. This is to be expected, since multi-label classification is challenging with regards to the number of labels that are assigned to a single instance. Such patterns could only be learned by models by including annotation guidelines in the prompt (which we did not have access to) or from the training data itself, which the generative models did not have access to. Fine-tuning generative LLMs on multi-label data could address this issue.

6. Conclusion

In this study, we examined several LLMs for classifying CSR themes and fine-grained CSR topics. We found that even though some smaller, less complex models yield respectable results for both multi-class and multi-label CSR classification, larger fine-tuned models are more successful at performing tasks. ZS experiments with GPT models showed that those models still fall behind on fine-tuned models for multi-label classification. This shortfall can be largely attributed to the complexities of multi-label classification, which demands an understanding of either explicit annotation guidelines or implicit annotator knowledge – insights that are not accessible to LLMs in a ZS context.

7. Acknowledgements

This research was funded by Flanders Innovation & Entrepreneurship (VLAIO), grant HBC.2021.0222 and by the Flemish government under FWO IRI project CLARIAH-VL.

8. Limitations

This study's findings are subject to several limitations. Firstly, the computational cost of running models on a large scale was not considered, which is crucial in practical applications due to resource constraints. Secondly, the choice of prompts for ZS classification and label interpretation may have affected the results, suggesting that exploring different prompting strategies could enhance performance. Thirdly, despite no significant impact observed from testing truncated texts, models capable of processing longer sequences might inherently benefit from more contextual information. These limitations highlight the need for continuous research to refine evaluation methodologies for LLMs, particularly in classifying CSR themes and topics.

9. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

Jasmin Bogatinovski, Ljupčo Todorovski, Sašo Džeroski, and Dragi Kocev. 2022. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215.

Arturo Rodríguez Castellanos, Carlos M. Parra, and Monica Chiarini Tremblay. 2015. Corporate social responsibility reports: Understanding topics via text mining. In *Americas Conference on Information Systems*.

Xiaolong Chen, Jieren Cheng, Jingxin Liu, Wenghang Xu, Shuai Hua, Zhu Tang, and Victor S. Sheng. 2022. A survey of multi-label text classification based on deep learning. In *Artificial Intelligence and Security: 8th International Conference, ICAIS 2022, Qinghai, China, July 15–20, 2022, Proceedings, Part I*, page 443–456, Berlin, Heidelberg. Springer-Verlag.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Loic De Langhe, Aaron Maladry, Bram Vanroy, Luna De Bruyne, Pranaydeep Singh, and Els Lefever. 2024. Benchmarking zero-shot text classification for dutch. *Computational Linguistics in the Netherlands Journal*, 13:63–90.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrea Gasparetto, Matteo Marcuzzo, Alessandro Zangari, and Andrea Albarelli. 2022. A survey on text classification algorithms: From text to predictions. *Information*, 13(2).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2021. Supervised contrastive learning.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. 2022. A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).

Nankai Lin, Guanqiu Qin, Gang Wang, Dong Zhou, and Aimin Yang. 2023. An effective deployment of contrastive learning in multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8730–8744, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey.

Yola Nayekoo, Sophia Katrenko, Veronique Hoste, Aaron Maladry, and Els Lefever. 2024. Shared task for cross-lingual classification of corporate social responsibility (csr) themes and topics. In *Proceedings of the Joint Workshop of FinNLP-KDF-ECONLP @ LREC-COLING 2024*, Torino, Italy.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Youri Peskine, Damir Korenčić, Ivan Grubisic, Paolo Papotti, Raphael Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4054–4063, Singapore. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer Berlin Heidelberg.

Amir Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2014. Automatic analysis of corporate sustainability reports and intelligent scoring. *International Journal of Computational Intelligence and Applications*, 13:27 pages.

Amir Mohammad Shahi, Biju Issac, and Jashua Rajesh Modapothala. 2011. Analysis of supervised text classification algorithms on corporate sustainability reports. In *Proceedings of 2011 International Conference on Computer Science and Network Technology*, volume 1, pages 96–100.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Fahim Sufi. 2024. Generative pre-trained transformer (gpt) in research: A systematic review on data augmentation. *Information*, 15(2).

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Simon Chi Lok U, Jie He, Víctor Gutiérrez-Basulto, and Jeff Z. Pan. 2023. Instances and labels: Hierarchy-aware joint supervised contrastive learning for hierarchical multi-label text classification.

Jens Van Nooten and Walter Daelemans. 2023. Improving Dutch vaccine hesitancy monitoring via multi-label data augmentation with GPT-3.5. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 251–270, Toronto, Canada. Association for Computational Linguistics.

Mingyu Wan and Chu-Ren Huang, editors. 2022. *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France.

Hui Wen and George Deltas. 2022. Global corporate social responsibility reporting regulation. *Contemporary Economic Policy*, 40(1):98–123.

Danqing Zhang, Tao Li, Haiyang Zhang, and Bing Yin. 2020. On data augmentation for extreme multi-label classification.

Appendix

A. Class Counts per Dataset

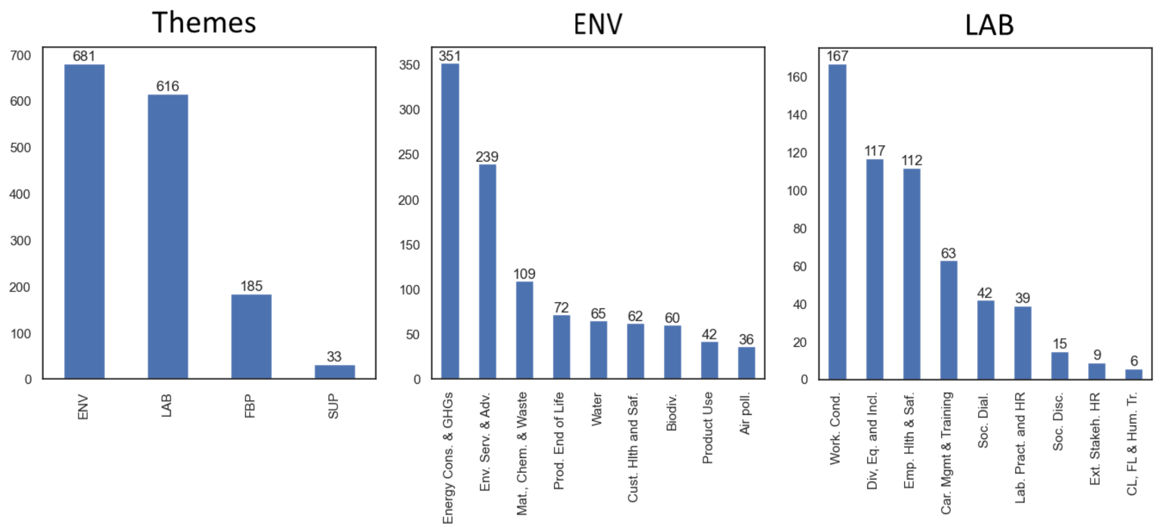


Figure 1: Class counts per (cleaned) dataset.

B. Prompts

B.1. Data Cleaning

"You are a human annotator extracting relevant parts of messy, unstructured texts. Your task is to extract useful parts of texts, like titles, subtitles and paragraphs of texts. The texts will be used for a classification task. It is very important that you ONLY remove the parts of the text that are not useful. The text that you will have to process will be provided in the next message. The output should just be the text, without any other information. Do not generate anything else besides the provided text."

B.2. Data Augmentation

"You are a helpful assistant tasked with creating synthetic data by translating or paraphrasing texts. Paraphrase the input text to approximately 300 words, aiming to retain the essential information. Here is the text: {INPUT_TEXT}."

B.3. Zero Shot Classification Themes

Prompt text:

"You are tasked with the role of a human annotator, responsible for carefully classifying texts into specific categories related to corporate social responsibility (CSR). Your role involves analyzing the content of various texts, including news articles, reports, and company statements, to identify their alignment with CSR topics. The classification categories are as follows:

1. ENV (Environment): Texts related to environmental sustainability, conservation efforts, impacts of corporate activities on the environment, climate change initiatives, and pollution control.

2. SUP (Sustainable Procurement): Texts discussing sustainable procurement practices, including ethical sourcing, supply chain sustainability, fair trade, and the environmental footprint of products and services.

3. LAB (Labour and Human Rights): Texts detailing labor conditions, human rights issues in business operations, employee welfare, workplace safety, and fair treatment practices within organizations.

4. FBP (Fair Business Practices): Texts focusing on corporate ethics, anti-corruption efforts, transparency, consumer rights, and fair competition in the business landscape.

{INPUT_TEXT}"

Function calling:

```
"functions": [{
    "name": "annotate_text",
    "description": "Analyzes the content of the text, determining its relevance to
        corporate social responsibility topics, and classifies it into one of the
        specified categories",
    "parameters": {
        "type": "object",
        "properties": {
            "text_category": {
                "type": "string",
                "enum": ["ENV", "SUP", "LAB", "FBP"],
                "description": "Corporate social responsibility topic assigned to the text
                    ."
            }
        },
        "required": ["text_category"]
    }
}]
```

B.4. Zero Shot Classification LAB

Prompt text:

"As a human annotator specializing in corporate social responsibility (CSR) with a focus on labor and human rights, your task is to classify texts into detailed categories that reflect various aspects of labor and human rights issues. This role involves a binary relevance classification, meaning for each category listed, you need to decide whether the text is relevant or not. A comprehensive examination of a variety of texts, such as news articles, reports, company statements, and more, is required to identify their relevance to specific topics within the realm of labor and human rights in CSR. A single text may cover multiple aspects of labor and human rights issues, allowing for multiple binary classifications as appropriate:

1. Career Mgmt & Training: Is the text relevant to career development, employee training, and professional growth within organizations?
2. Child Labor, Forced Labor, and Human Trafficking: Does the text address child labor, forced labor, and human trafficking issues?
3. Diversity, Equity, and Inclusion: Is the text focusing on diversity, equity, and inclusion efforts in the workplace?
4. Employee Health & Safety: Does the text concern workplace health and safety policies and practices?
5. External Stakeholder Human Rights: Is the text on human rights issues affecting external stakeholders impacted by corporate activities?
6. Labour Practices and Human Rights: Does the text detail labor practices and human rights considerations within organizations?
7. Social Dialogue: Is the text related to dialogue between employees and management aimed at improving working conditions and relations?
8. Social Discrimination: Does the text deal with social discrimination issues within the workplace or business operations?
9. Working Conditions: Is the text related to employment conditions, such as work hours, pay, and overall work environment?

{INPUT_TEXT}"

Function calling:

```
"functions": [{
    "name": "annotate_text",
    "description": "Analyze text content to determine its binary relevance to labor and
        human rights topics within CSR. For each of the specified categories, the
        annotator will classify the text as either 'relevant' or 'not relevant', based on
        the issues it addresses.",
    "parameters": {
        "type": "object",
        "properties": {
            "text_categories": {
                "type": "object",
                "properties": {
                    "Career Mgmt & Training": {"type": "boolean", "description": "
                        Indicates if the text is relevant to career management and
                        training."},
                    "Child Labor, Forced Labor, and Human Trafficking": {"type": "boolean
                        ", "description": "Indicates if the text addresses child labor,
                        forced labor, and human trafficking issues."},
                    "Diversity, Equity, and Inclusion": {"type": "boolean", "description":
                         "Indicates if the text focuses on diversity, equity, and
                        inclusion efforts."},
                    "Employee Health & Safety": {"type": "boolean", "description": "
                        Indicates if the text is relevant to employee health and safety
                        ."},
                    "External Stakeholder Human Rights": {"type": "boolean", "description
                        ": "Indicates if the text discusses external stakeholder human
                        rights issues."},
                    "Labour Practices and Human Rights": {"type": "boolean", "description
                        ": "Indicates if the text details labor practices and human rights
                         considerations."},
                    "Social Dialogue": {"type": "boolean", "description": "Indicates if
                        the text is related to social dialogue for improving working
                        conditions."},
                    "Social Discrimination": {"type": "boolean", "description": "Indicates
                         if the text deals with social discrimination issues."},
                    "Working Conditions": {"type": "boolean", "description": "Indicates if
                         the text is relevant to working conditions."}
                },
                "required": ["Career Mgmt & Training", "Child Labor, Forced Labor, and
                    Human Trafficking", "Diversity, Equity, and Inclusion", "Employee
                    Health & Safety", "External Stakeholder Human Rights", "Labour
                    Practices and Human Rights", "Social Dialogue", "Social Discrimination
                    ", "Working Conditions"]
            }
        },
        "required": ["text_categories"]
    }
}]
```

C. Contrastive Loss Formulas

Given are vectors α and β, where α contains binary labels indicating whether an in-batch text pair is positive (similar) or negative (dissimilar). Consult Section 4 for a description on positive and negative sample selection. Eq. 1 describes the normalization procedure of the cosine distance values. Eq. 2 denotes the calculation of the contrastive loss, which is the BCE loss between α and β'.

$$\beta' = \text{sig}\left(\frac{\beta}{\Theta}\right) \tag{1}$$

$$\text{CL}(\alpha, \beta) = -\sum i \left[\alpha i \log(\beta' i) + (1 - \alpha i)\log(1 - \beta' i)\right] \tag{2}$$

D. Model Hyperparameters

dataset	tf-idf + SVM	ada-003 + SVM	ada-003 + MLP	ada-003 + LSTM
Themes	ngram = (1,1) C = 50 max iter = 1.000 max features = 1.000	C = 10 max iter = 100	LR = 1e-4 n-layers = 2 n iter = 500	LR = 1e-3 epochs = 50 n-layers = 1 dropout = 0.1 hidden dim = 700
ENV	ngram = (1,1) C = 10 max iter = 100 max features = 10000	C = 50 max iter = 1000	LR = 1e-3 n-layers = 2 n iter = 500	LR = 1e-3 epochs = 100 n-layers = 1 dropout = 0.3 hidden dim = 700
LAB	ngram = (1,3) C = 10 max iter = 500 max features = 1000	C = 100 max iter = 500	LR = 1e-3 n-layers = 2 n iter = 500	LR = 1e-3 epochs = 100 n-layers = 2 dropout = 0.3 hidden dim = 700

Table 3: Optimal hyperparameters for the baseline models, obtained by performing gridsearch experiments.

dataset	DB	DB + CL	DB + DA	DB + CL + DA	RB / BERT	RB / BERT + CL	RB/BERT + DA	RB/BERT + CL + DA	RB-Large	RB-Large + CL	RB-Large + DA	RB-Large + CL + DA
Themes	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 1.0 JI = 1.0 alpha = 0.1	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 1.0 JI = 1.0 alpha = 0.1	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 0.5 JI = 1.0 alpha = 0.2	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 0.5 JI = 1.0 alpha = 0.2	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 0.5 JI = 1.0 alpha = 0.1	LR = 2e-5 epochs = 10	LR = 2e-5 epochs = 10 temp = 1.0 JI = 1.0 alpha = 0.2
ENV	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 2e-5 epochs = 10 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 2e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1
LAB	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 5e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 0.5 JI = 0.5 alpha = 0.1	LR = 2e-5 epochs = 15	LR = 5e-5 epochs = 15 temp = 1.0 JI = 0.5 alpha = 0.1

Table 4: Optimal hyperparameters for the LLMs used in this study.

E. Label Names

Idx	ENV	LAB
0	'Air pollution'	'Employee Health & Safety'
1	'Biodiversity'	'Career Mgmt & Training'
2	'Customers Health and Safety'	'Working Conditions'
3	'Energy Consumption & GHGs'	'External Stakeholder Human Rights'
4	'Environmental Services & Advocacy'	'Diversity Equity and Inclusion'
5	'Materials Chemicals & Waste'	'Child Labor Forced Labor and Human Trafficking'
6	'Product End of Life'	'Labour Practices and Human Rights'
7	'Product Use'	'Social Dialogue'
8	'Water'	'Social Discrimination'

Table 5: Label indices and their corresponding names per dataset.

F. Confusion Matrices

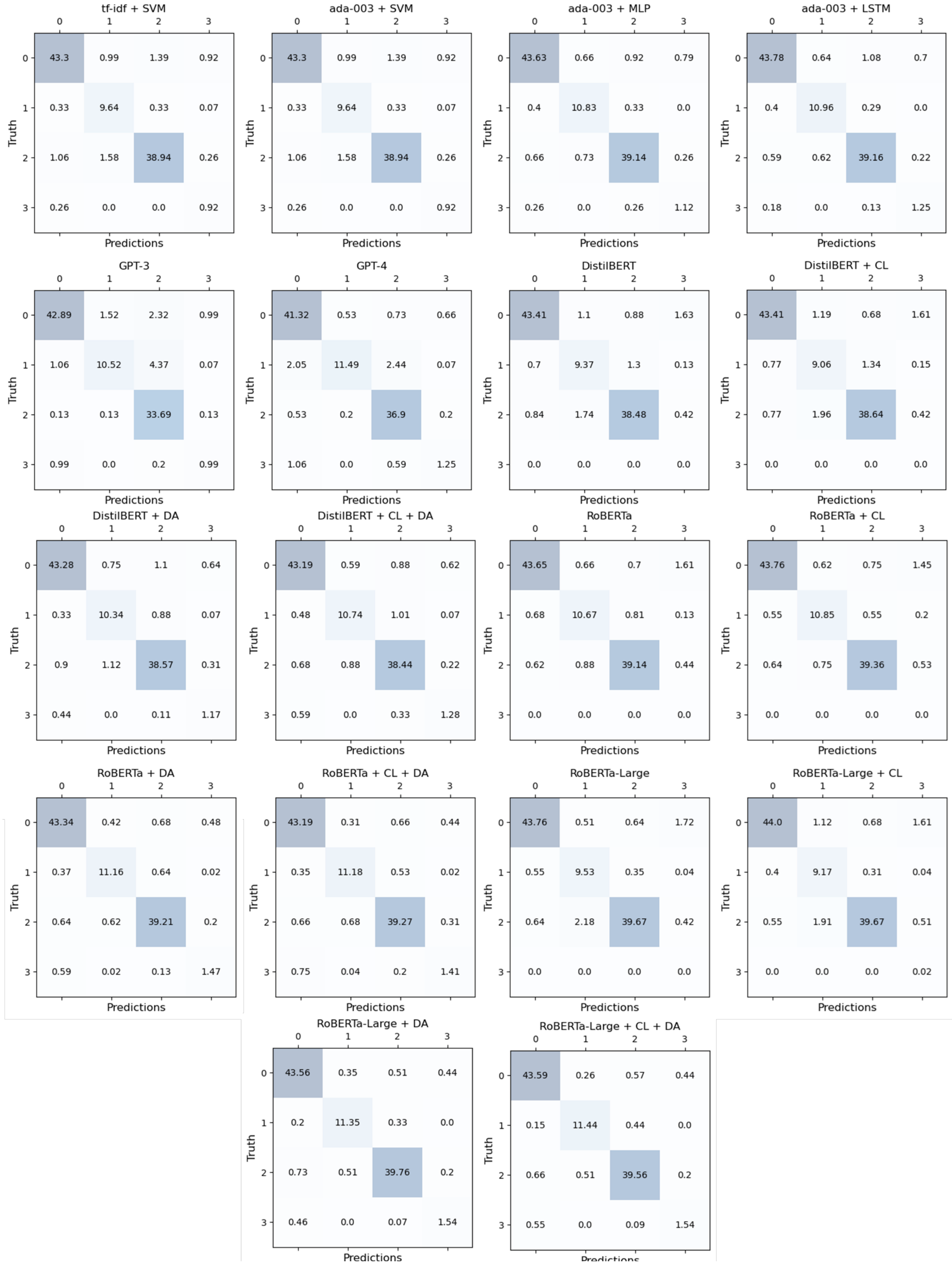


Figure 2: Confusion matrices from all models trained on the themes dataset.

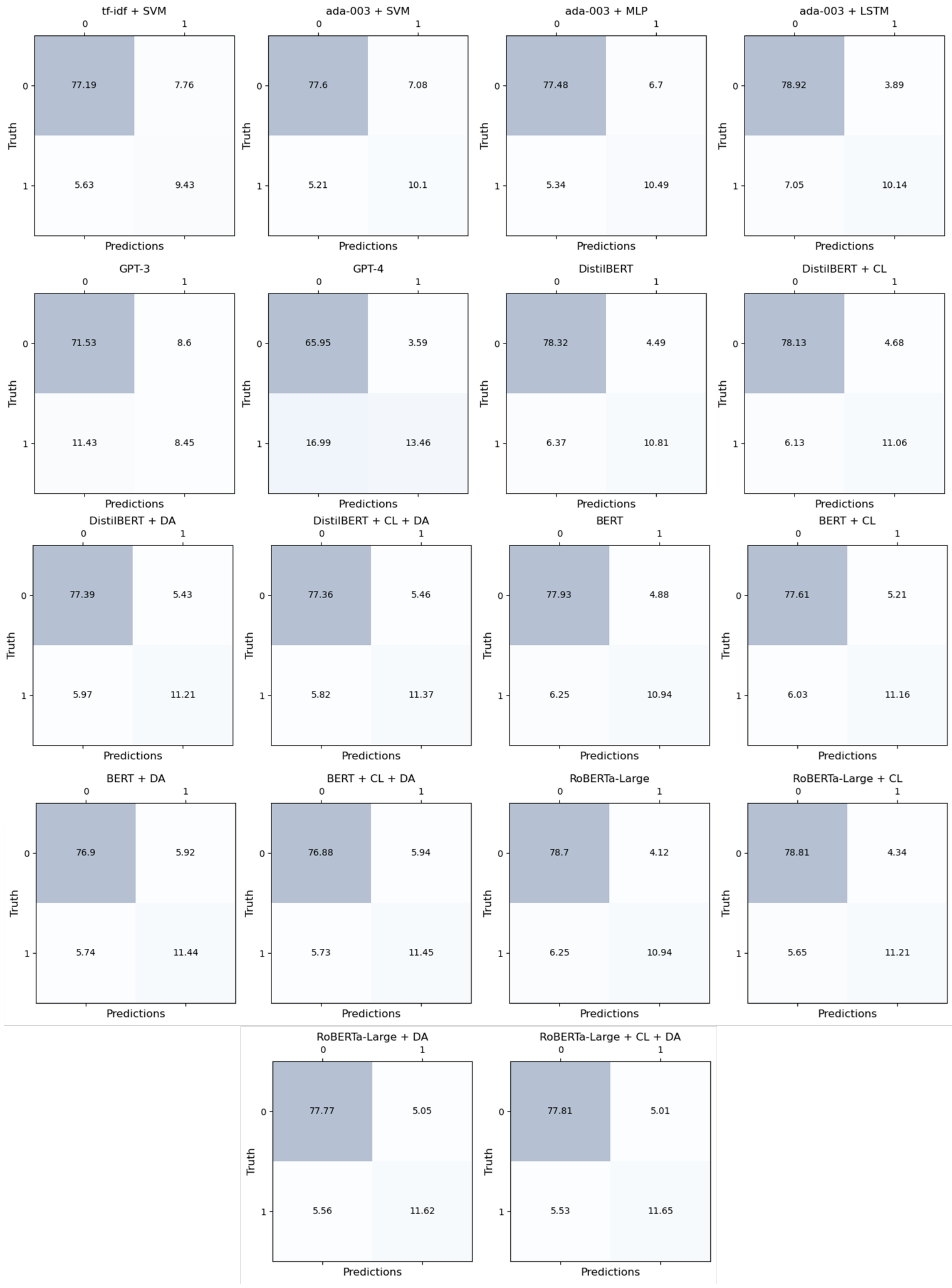


Figure 3: Confusion matrices from all models trained on the ENV dataset.

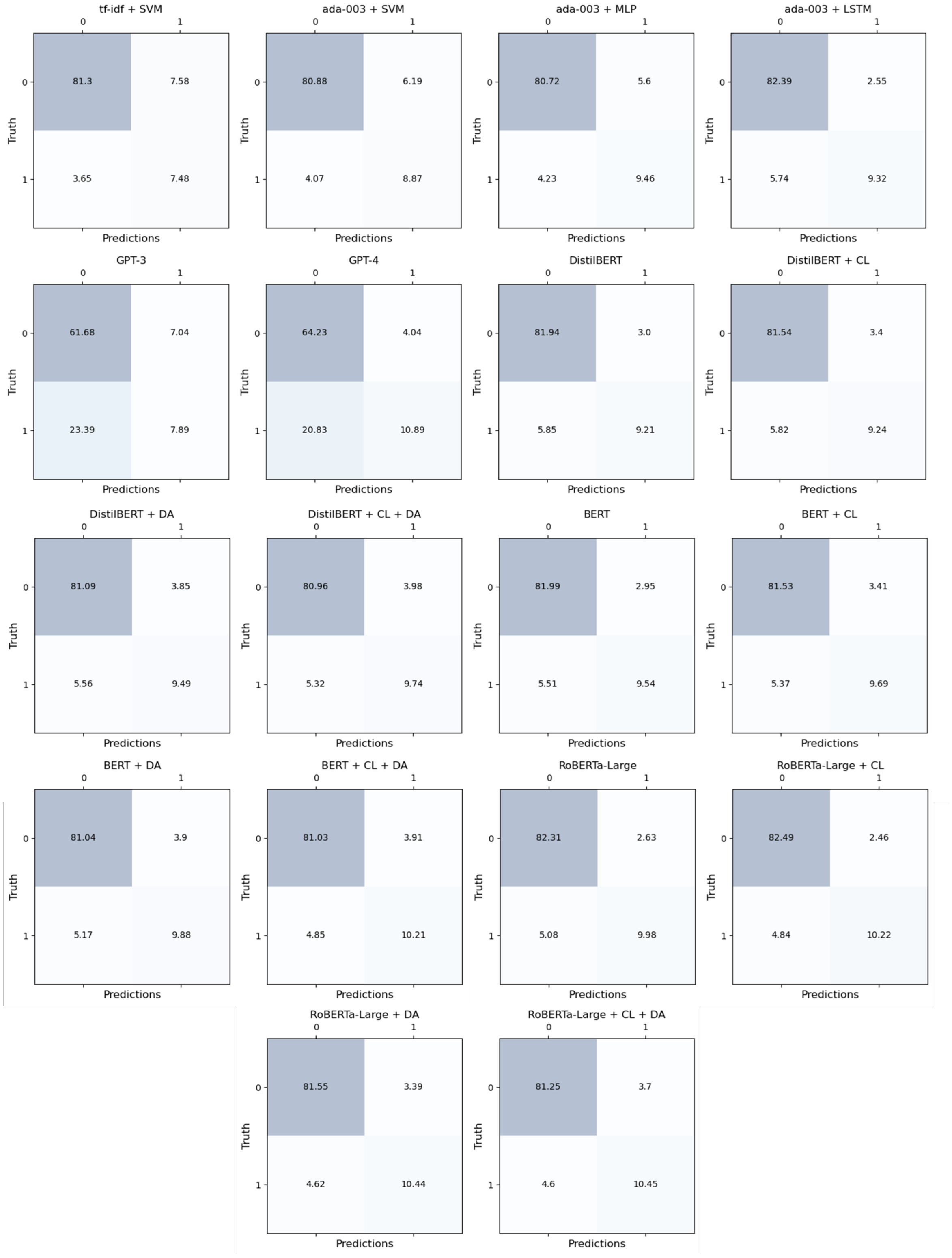


Figure 4: Confusion matrices from all models trained on the LAB dataset.

Improving Cross-Lingual CSR Classification using Pretrained Transformers with Variable Selection Networks and Data Augmentation

Shubham Sharma, Himanshu Janbandhu, Ankush Chopra
Tredence, Whitefield, Bengaluru,India
{shubham.sharma,himanshu.jandbandhu, ankush.chopra}@tredence.com

Abstract

This paper describes our submission to the Cross-Lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics shared task, aiming to identify themes and fine-grained topics present in news articles. Classifying news articles poses several challenges, including limited training data, noisy articles, and longer context length. In this paper, we explore the potential of using pretrained transformer models to classify news articles into CSR themes and fine-grained topics. We propose two different approaches for these tasks. For multi-class classification of CSR themes, we suggest using a pretrained multi-lingual encoder-based model like microsoft/mDeBERTa-v3-base, along with a variable selection network to classify the article into CSR themes. To identify all fine-grained topics in each article, we propose using a pretrained encoder-based model like Longformer, which offers a higher context length. We employ chunking-based inference to avoid information loss in inference and experimented with using different parts and manifestation of original article for training and inference.

Keywords: News Classification, CSR Classification, Multi-lingual Model

1. Introduction

In recent years, Corporate Social Responsibility (CSR) and Sustainability reporting have become significantly important for businesses. Consumers and investors are increasingly conscious of the social and environmental impacts of the companies they engage with or invest in. They often expect businesses to demonstrate a commitment to CSR, sustainability, and ethical business practices. The media plays a crucial role in highlighting CSR issues and holding companies accountable for their actions.

This work is an outcome of a shared task organized in the EcoNLP workshop Nayekoo et al. (2024) which had two subtasks. These subtasks focus on automatically identifying themes and topics present in news articles. The first subtask involves multi-class classification of multi-lingual news articles. We experimented with various multi-lingual encoder-only transformer models. We also experimented with various transformations of input text such as translation, summarization and chunking for accurately classifying the news articles. The best performing model was mDeBERTa (He et al., 2023) in conjunction with variable selection network trained on article text and the titles.

The second subtask is a multi-label classification for fine-grained labels among two CSR themes: ENV and LAB. This subtask was conducted with only English language news articles. The best performing model for this task was Longformer (Beltagy et al., 2020) with multi-label head and custom threshold for each class head.

2. Related Work

The field of multi-lingual text classification, integral to natural language processing (NLP), has seen significant advancements with the introduction of pre-trained transformer-based models such as mBERT.Gürel and Emin (2021) have effectively utilized mBERT for multi-lingual text classification, showcasing its proficiency across various languages. Complementing this, Pujari et al. (2021) have explored the use of a transformer-based multi-task model for multi-label classification, employing a novel approach that trains a dedicated classifier for each node by merging transformers with hierarchical algorithms.

In a related development, Wang et al. (2021) have adopted graph convolutional networks (GCN) for cross-lingual text classification. Their method involves constructing a heterogeneous graph where documents and words serve as nodes. These nodes are interconnected through a network of relationships defined by part-of-speech roles, semantic similarity, and document translations, facilitating a comprehensive understanding of language nuances.

Parallel to these developments, the area of Corporate Social Responsibility (CSR) topic identification has been flourishing. Chae and Park (2018) have applied a probabilistic topic modeling-based computational text analysis framework to examine the prevalence, evolution, and correlation of CSR topics. Further advancing the research, Salvatore et al. (2022) have employed a structural topic model to detect CSR topics and assess the impact of time

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 306–318
20 May, 2024.

and sector on the proportion of discussions surrounding these topics.

3. Dataset and Problem statements

3.1. Data Scraping

The datasets provided for both subtasks consisted of URLs of the articles, using which we had to scrape text. To accomplish this, we utilized a Python library called Newspaper3K (Ou-Yang, 2018), which provided us with the text of the articles (excluding boilerplate), article titles, and other metadata. This information was extracted from the HTML of the URLs by the library itself. Due to challenges in web scraping, we were unable to scrape all instances provided in the original data of both the tasks.

3.2. Subtask A

This subtask aims to identify the theme of news articles, which could be in English, French, or simplified Chinese. The themes are categorized into four labels: environment, labor and human rights, fair business practices, and sustainable procurement. Therefore, it becomes a multi-lingual and multi-class classification problem.

3.2.1. Dataset

In our experiments, we utilized the Corporate Social Responsibility theme recognition dataset provided by the event organizers. This dataset included URLs and CSR themes for each news article. Since the dataset did not contain the actual text of the news articles, we resorted to web scraping to extract it. The four main themes were categorized as follows: 1. ENV (Environment), 2. LAB (Labor and Human Rights), 3. FBP (Fair Business Practices), and 4. SUP (Sustainable Procurement).
The provided dataset was only available in English, but predictions needed to be made in three languages: English, French, and simplified Chinese. To deal with the small sample size, we also employed a synthetic data generation method for the SUP class as described in section 4.4. The class distribution of the data is provided in Table 1. The tokenized length distriution of scraped articles is provided in Table 2.

Data	ENV	LAB	FBP	SUP
Original Data	706	652	197	39
Data Post Scraping	633	547	179	33
With Synthetic Data	633	547	179	153

Table 1: Label Distribution

Max	Min	Mean	Std
13876	23	566	712

Table 2: Length Distribution in scraped articles for Task A.

3.3. Subtask B

This subtask aims to classify fine-grained CSR topics within the Environment (ENV) and Labor and Human Rights (LAB) themes (English), allowing for multiple topics to be assigned to an article within each specified theme. Therefore, it presents a multi-label multi-class classification challenge.

3.3.1. Dataset

For this subtask, we received two separate datasets for ENV and LAB, each with its own set of labels. These datasets were also scraped as discussed in section 3.1. The label distribution before and after scraping for the ENV dataset is provided in Table 4. The tokenized length distribution of scraped articles is provided in Table 3.

Dataset	Max	Min	Average	Std
ENV	4866	35	539	542
LAB	11867	36	655	1256

Table 3: Length Distribution in scraped articles for Task B.

Similarly, for the LAB class, the label distribution before and after scraping is detailed in Table 5. We observed that the labels "External Stakeholder Human Rights" and "Social Discrimination" had a smaller number of training examples provided. Additionally, they were not highly co-occurring with other classes. Consequently, we created artificial data for these two labels using the method described in section 4.4.

ENV	Original Data	Records Post Scraping
Air Pollution	36	31
Biodiversity	62	51
Customer Health and Safety	62	48
Energy Consumption and GHGs	366	313
Environmental Services and Advocacy	242	204
Materials, Chemical and Waste	112	92
Product End of Life	73	64
Product Use	44	36
Water	71	57

Table 4: Number of records for ENV dataset

LAB	Original Data	Records Post Scraping	With Synthetic Data
Career Mgmt and Training	77	42	42
Child Labor, Forced Labor, and Human Trafficking	7	4	4
Diversity, Equity, and Inclusion	149	99	99
Employee Health and Safety	138	100	100
External Stakeholder Human Rights	14	12	36
Labour Practices and Human Rights	47	29	29
Social Dialogue	52	28	28
Social Discrimination	18	18	42
Working Conditions	201	128	128

Table 5: Number of records for LAB dataset

4. Methodology

Both subtasks involve classification, making encode-only transformer models a suitable choice for fine-tuning. We opted for multi-lingual models for subtask A and English language models for subtask B. Experimentation with various models, some of which had a maximum sequence length of 512, was conducted. However, a significant portion of the input news articles exceeded the 512-length limit, as indicated in Table 2 and Table 3. Thus, we explored methods for handling such documents, including the use of models like Longformer with higher sequence lengths and the employment of data transformation through chunking and summarization.

We used full finetuning of the transformer-based model with Pytorch library with appropriate cross-entropy losses for both the subtasks. We used three models, DeBERTa, MDeBERTa and Longformer for fine-tuning.

Data augmentation, transformation, and experimentation with model architecture were key components in our search for optimal solutions. We designed numerous experiments by exploring various combinations of these steps. The specifics of the data and model architecture-related variations are outlined below

4.1. Summarization

Abstractive summarization was utilized to generate summaries for the articles. This approach enabled us to reduce the length of the articles to fit within the BERT model's context length limit. The articles were segmented into paragraphs, ensuring they did not exceed the maximum context size of the summarization model. Each paragraph was then summarized to a specified number of words, depending on the number of chunks, to ensure the summarized article fit within the context length limit of the downstream classification model. We utilized a T5-based (Raffel et al., 2023) summarization model called "Falconsai/text_summarization", (See Appendix 8.5) which is hosted on Hugging Face (Wolf et al., 2020).

4.2. Chunking

To address the challenge posed by excessively long articles in our dataset, we attempted to divide the dataset into multiple sequences or chunks. Each chunk was created not to surpass the maximum context length limit of the model utilized during the experiment. In subtask A, we could allocate each chunk to the same category and utilize this data to train the model.

4.2.1. Mean Probability based Prediction.

This approach involved leveraging probabilities from each chunk to calculate the overall average probability for each class. Subsequently, we utilized this mean probability to make predictions.

4.2.2. Max Voting based Prediction

This method was solely employed for subtask A. Each chunk predicted a class based on its predicted probabilities. These chunk-wise predictions were aggregated as votes for each class, and the predicted class was determined by the maximum number of chunks voting for it. In case of a tie, we used the maximum probability of each class to make the final prediction.

4.3. Variable Selection Network (VSN)

Given the longer length of the texts, we conjectured that the CLS token might not always encapsulate the complete meaning and context of the input. This prompted us to explore methods to provide selected additional information to the classification layer for improved predictions. VSN (Lim et al., 2020) is a neural network architecture designed for feature selection in high-dimensional datasets. It utilizes a gating mechanism to learn the most relevant features for a given task, focusing on important features while disregarding irrelevant ones. The architecture of VSN is illustrated in Figure 1.
We employed VSN to generate an embedding representing all token embeddings in the BERT-based model. This embedding was concatenated with the CLS token and passed for fine-tuning. (We refer to Figure 6 which illustrates the VSN architecture.)

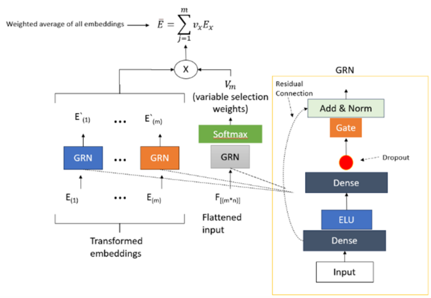


Figure 1: Variable Selection Network

4.4. Synthetic Data Generation

We utilized the GPT-4 (OpenAI et al., 2024) model by OpenAI to generate artificial samples of minority classes. This was achieved using prompts specifically tailored for generating synthetic text articles. The prompts utilized for generating data for selected classes in both subtasks can be found in appendix 8.4.

4.5. Title Concatenation

As the news articles provided by the organizers were in URL format, we had access to rich metadata from these news webpages. After careful analysis and observation of available common metadata fields, we opted to incorporate page titles along with article text. Page titles are crafted to succinctly represent page content and offer high-quality condensed information.
In certain experiments, we concatenated the title with the original text of the news article, separated by the SEP token of the model. Different token type IDs were assigned to both the title and text to ensure that the model could attend to both sequences.

4.6. Translation

The data provided by the organizers for subtask A was in English, while the test set was announced to be in French and Chinese, in addition to English. Given the challenge of identifying, scraping, and tagging relevant French and Chinese data, we decided to translate the provided English articles into French and Chinese for training. We ensured consistency by maintaining the same records in the validation set across all languages. We utilized pretrained models from Hugging Face for translation, specifically "Helsinki-NLP/opus-mt-tc-big-en-fr" (See Appendix 8.5) for French translation and "Helsinki-NLP/opus-mt-en-zh" (See Appendix 8.5) for simplified Chinese translation. (Tiedemann and Santhosh, 2020).

4.7. Dynamic Weighted Loss

As the dataset was imbalanced, we decided to assign weights to each class in the loss function. We started with equal weights for each class in the initial epoch, then at the start of the next epoch, we assigned weights in such a way that class with minimum f1 should get maximum weight. Here is the formula:
Weight for class i,

$$W_i = \frac{1-F1_i}{\sum_{j=0}^{n}(1-F1_j)}$$

Where n is the number of classes

5. Experiments

After implementing the methods described in the previous section, we conducted numerous experiments. In this section, we present our best systems for both subtask A and subtask B. For details on other experiments, please refer to the Appendix. (refer 8.1, 8.2, 8.3)

5.1. Subtask A

As outlined in section 3.2, subtask A poses a multilingual multi-class classification challenge. Through a combination of methodologies discussed in section 4, we discovered that utilizing article titles indeed improves performance. In addition, we observed that using variable selection network along with title concatenation further enhanced the performance of the system.

ENV F1	LAB F1	FBP F1	SUP F1	ACC	language
96.3	96.5	95.4	71.4	94.9	English
96.8	96.6	94.5	79.1	95.3	French
92.4	92.6	86.6	66.6	90.0	Chinese

Table 6: Subtask A validation set results per language

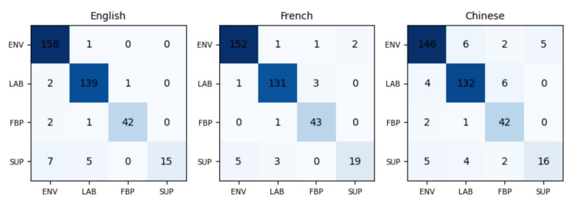


Figure 2: : Subtask A confusion matrix per language

Air pollution	Biodiversity	Customers Health and Safety	Energy Consumption and GHGs	Environmental Services and Advocacy	Materials, Chemicals and Waste	Product End of Life	Product Use	Water	micro avg
71.42	66.66	85.71	89.69	60.93	59.45	68.75	37.50	72	73.20

Table 7: Subtask B, ENV, F1-Scores

Career Mgmt and Training	Child Labor, Forced Labor, and Human Trafficking	Diversity, Equity, and Inclusion	Employee Health and Safety	External Stakeholder Human Rights	Labour Practices and Human Rights	Social Dialogue	Social Discrimination	Working Conditions	micro$_{avg}$
69.56	66.66	72.22	81.63	40	22.22	94.11	57.14	80.48	71.48

Table 8: Subtask B, LAB, F1-Scores

We evaluated our models based on their F1 scores for each class and language. Our best-performing model utilized the DeBERTa-based multi-lingual model, specifically the "microsoft/mdeberta-v3-base (8.5)" model hosted on Hugging Face and pretrained on CC100 multilingual data. This best system incorporated artificial data for the SUP class, as discussed in section 4.4, and utilized translated datasets, as described in section 4.6. Additionally, it involved title concatenation, as outlined in section 4.5, and employed a variable selection network, as discussed in section 4.3. In terms of inference, we utilized chunking with mean probability, as detailed in section 4.2.1 . During model training, we included original SUP data in the validation set and artificial SUP data in the training set. For classes other than SUP, original data were used in training and validation set. The results of the best-performing model on validation set are provided in Table 6. In Figure 2, i-th row and j-th column entry indicates the number of samples with true label being i-th class and predicted label being j-th class. The results of best performing model on test set are given in the Table 12.

5.2. Subtask B

In subtask B, a multi-class, multi-label problem was tackled using fine-tuning of transformer-based models, employing the binary-crossentropy loss for each class. This approach suits multi-label classification due to its capability to handle instances with multiple labels and provide a probabilistic interpretation of predictions, facilitating overall loss reduction.

In the ENV dataset, shorter articles were predominant compared to the LAB dataset (see Table 3). Initially using DeBERTa, the baseline model was shifted to Longformer due to a large portion of data exceeding 512 tokens, the maximum context size for DeBERTa. Title concatenation notably enhanced performance across experiments with Longformer. Further experiments included VSN exploration, adjusting context-length limits with Longformer, and prediction with chunked articles, though no improvement over the best model was observed. The optimal model for the ENV dataset was Longformer with a 1500 context-length limit and title concatenation.

Similarly, in the LAB dataset, Longformer outperformed DeBERTa after summarization. Title concatenation, however, led to performance deterioration. The same experiments as in the ENV dataset were conducted, with the best model again being Longformer with a 1500 context-length limit. Best results for ENV and LAB subtask on validation set can be found Table 7 and Table 8 in respectively. Summarized results of Task B on test set can be found in Table 13.

6. Conclusion

In this paper, we introduce systems developed by our team for the Cross-Lingual Classification of Corporate Social Responsibility (CSR) Themes and Topics tasks. This task aims at automatically classifying news articles into CSR themes and identifying fine-grained topics within them. To achieve this, we finetuned transformer models along with a variable selection network to classify articles into suitable CSR themes. By experimenting with title along with article text we uncovered that using metadata effectively along with the article text can help immensely in improving the accuracy of the classification.

7. Bibliographical References

Charu C. Aggarwal and ChengXiang Zhai. 2012. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer.

Bongsug (Kevin) Chae and Eunhye (Olivia) Park. 2018. Corporate social responsibility (csr): A survey of topics and trends using twitter data and topic modeling. *Sustainability*, 10(7).

Alaeddin Gürel and Emre Emin. 2021. ALEM at CASE 2021 task 1: Multilingual text classification on news articles. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 147–151, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. 2020. Temporal fusion transformers for interpretable multi-horizon time series forecasting.

Yola Nayekoo, Sophia Katrenko, Véronique Hoste, Aaron Maladry, and Els Lefever. 2024. Shared task for cross-lingual classification of corporate social responsibility (csr) themes and topics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Lucas Ou-Yang. 2018. Newspaper3k: Article scraping curation. Technical report.

Subhash Chandra Pujari, Annemarie Friedrich, and Jannik Strötgen. 2021. A multi-task approach to neural multi-label hierarchical patent classification using transformers. In *Advances in Information Retrieval*, pages 513–528, Cham. Springer International Publishing.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Camilla Salvatore, Silvia Biffignandi, and Annamaria Bianchi. 2022. Corporate social responsibility activities through twitter: From topic model analysis to indexes measuring communication characteristics. *Social Indicators Research*.

Jörg Tiedemann and Thottingal Santhosh. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Ziyun Wang, Xuan Liu, Peiji Yang, Shixing Liu, and Zhisheng Wang. 2021. Cross-lingual text classification with heterogeneous graph neural network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 612–620, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

8. Appendix

8.1. Table A: Experiments and results on Subtask A

Experiment No.	1	2	3	4	5	6
MODEL_NAME	mDeBERTa	mBERT	Longformer	mDeBERTa	mDeBERTa	mDeBERTa
Methods_used				4.2, 4.2.2	4.2, 4.2.1	4.1
English						
ENV F1	97.79	94.67	92.77	94.26	95.95	93.04
LAB F1	96.48	94.04	93.19	94.55	95.74	94.48
FBP F1	91.67	89.16	93.18	86.67	92.47	82.35
SUP F1	0.00	35.29	0.00	60.00	60.00	40.00
ACCURACY	95.45	92.33	92.07	92.92	94.90	91.22
French						
ENV F1	91.67	89.76	90.21	87.36	90.96	89.47
LAB F1	92.26	85.42	90.04	87.45	88.30	89.11
FBP F1	86.36	61.11	81.32	72.94	80.00	77.11
SUP F1	0.00	16.67	0.00	20.00	22.22	37.50
ACCURACY	90.34	83.81	88.10	84.70	87.54	86.69
Chinese						
ENV F1	91.86	91.19	90.96	93.25	92.26	93.46
LAB F1	90.91	90.58	86.86	88.97	91.37	90.68
FBP F1	86.02	77.65	77.42	74.73	81.63	82.93
SUP F1	0.00	28.57	0.00	0.00	0.00	53.33
ACCURACY	89.77	88.07	86.69	87.54	89.52	90.27

Experiment No.	7	8	9	10	11	12	**13**
MODEL_NAME	mDeBERTa	mDeBERTa	mDeBERTa	mDeBERTa	mDeBERTa	mDeBERTa	**mDeBERTa**
Methods_used	4.4	4.2.1, 4.4	4.7	4.2.1, 4.4, 4.6	4.2.1, 4.4, 4.5, 4.6	4.2.1, 4.3, 4.4, 4.6	**4.2.1, 4.3, 4.4, 4.5, 4.6**
English							
ENV F1	93.05	93.05	94.05	94.41	95.71	93.46	**96.34**
LAB F1	95.47	95.50	95.31	95.95	95.92	95.04	**96.53**
FBP F1	94.12	87.50	89.16	93.02	89.16	90.53	**95.45**
SUP F1	65.12	69.57	0.00	70.37	69.77	70.00	**71.43**
ACCURACY	92.40	91.90	92.80	93.13	93.50	91.80	**94.90**
French							
ENV F1	89.51	93.01	89.22	92.95	95.60	92.60	**96.82**
LAB F1	86.93	93.53	88.65	95.59	94.70	94.38	**96.68**
FBP F1	73.97	86.08	79.49	90.32	85.71	87.23	**94.51**
SUP F1	60.47	57.89	0.00	68.09	78.26	73.08	**79.17**
ACCURACY	85.25	90.60	86.60	91.90	93.09	91.10	**95.30**
Chinese							
ENV F1	89.41	88.24	91.99	88.52	92.11	91.03	**92.41**
LAB F1	91.16	92.65	91.64	93.43	94.24	92.78	**92.63**
FBP F1	85.00	87.18	86.75	84.21	88.37	84.21	**86.60**
SUP F1	31.25	25.00	22.22	50.00	66.67	69.23	**66.67**
ACC	87.13	86.90	90.30	87.20	90.80	89.19	**90.00**

Table 9: All experiments results for Subtask A

8.2. Table B : Subtask B, Experiments on ENV dataset

Experiment	1	2	3	4	**5**	6	7
Model-Name	DeBERTa	Longformer	DeBERTa	DeBERTa	**Longformer**	DeBERTa	DeBERTa
Context Length	512	1500	512	512	**1500**	512	512
Methods Used			4.1	4.5	**4.5**	4.3	4.3, 4.5
Air pollution f1-score	50.00	50.00	50.00	33.33	**71.43**	52.63	50.00
Biodiversity f1-score	12.50	50.00	40.00	33.33	**66.67**	55.56	62.07
Customers Health and Safety f1-score	58.33	66.67	58.33	63.64	**85.71**	76.92	80.00
Energy Consumption & GHGs f1-score	86.08	86.08	84.81	87.65	**89.70**	87.18	87.43
Environmental Services & Advocacy f1-score	42.55	44.68	46.81	56.86	**60.94**	58.06	59.63
Materials, Chemicals & Waste f1-score	48.15	51.85	44.44	40.00	**59.46**	56.76	53.57
Product End of Life f1-score	56.25	43.75	64.52	62.50	**68.75**	68.57	68.57
Product Use f1-score	0.00	42.86	14.29	40.00	**37.50**	16.67	34.04
Water f1-score	64.52	81.25	56.25	64.00	**72.00**	76.47	66.67
micro_avg	58.98	64.41	61.54	64.72	**73.20**	68.05	66.20

Experiment	8	9	10	11	12
Model Name	Longformer	Longformer	Longformer	DeBERTa	Longformer
Context Length	1500	1500	2000	512	1200
Methods used	4.3	4.3, 4.5	4.5	4.2.1, 4.5	4.5
Air pollution f1-score	66.67	70.00	57.14	61.54	71.43
Biodiversity f1-score	77.78	71.43	66.67	25.00	63.64
Customers Health and Safety f1-score	83.33	84.21	80.00	77.78	73.68
Energy Consumption & GHGs f1-score	88.89	85.53	91.12	85.71	89.17
Environmental Services & Advocacy f1-score	57.69	60.00	59.35	60.00	59.35
Materials, Chemicals & Waste f1-score	54.29	51.16	57.63	43.24	50.00
Product End of Life f1-score	71.43	74.29	65.00	70.59	70.27
Product Use f1-score	29.41	24.00	27.59	35.29	28.57
Water f1-score	88.24	66.67	81.82	55.17	80.00
micro_avg	69.17	68.56	69.96	64.92	67.30

Table 10: All experiments results, Subtask B, ENV Dataset

8.3. Table B : Subtask B, Experiments on LAB dataset

Experiment	**1**	2	3	4	5	6
Model Name	**Longformer**	DeBERTa	Longformer	DeBERTa	DeBERTa	DeBERTa
Context Length	**1500**	512	1500	512	512	512
Methods Used		4.5	4.5	4.1	4.3	4.3, 4.5
Career Mgmt & Training f1-score	**69.56**	60.86	60.86	60.00	76.19	63.15
Child Labor, Forced Labor, and Human Trafficking f1-score	**66.66**	0	13.33	3.33	14.28	2.77
Diversity, Equity, and Inclusion f1-score	**72.22**	66.67	66.67	68.75	66.67	76.47
Employee Health & Safety f1-score	**81.63**	81.82	72.34	69.57	72.34	72.73
External Stakeholder Human Rights f1-score	**40.00**	14.29	50.00	6.82	12.24	23.08
Labour Practices and Human Rights f1-score	**22.22**	37.50	32.43	20.00	23.53	34.78
Social Dialogue f1-score	**94.12**	80.00	80.00	80.00	80.00	72.73
Social Discrimination f1-score	**57.14**	38.10	25.00	54.55	50.00	57.14
Working Conditions f1-score	**80.49**	78.95	85.37	81.72	81.93	77.50
micro_avg	**71.49**	48.56	63.94	44.50	58.98	51.38

Experiment	7	8	9	10
Model Name	Longformer	Longformer	Longformer	Longformer
Context Length	1500	1500	1500	1200
Methods Used	4.3	4.3, 4.5	4.4	
Career Mgmt & Training f1-score	81.81	63.15	66.66	70
Child Labor, Forced Labor, and Human Trafficking f1-score	100	12.50	100	66.66
Diversity, Equity, and Inclusion f1-score	68.42	70.00	70.59	71.79
Employee Health & Safety f1-score	76.92	82.61	80.85	72.34
External Stakeholder Human Rights f1 score	66.67	25.00	40.00	24.00
Labour Practices and Human Rights f1-score	17.54	37.04	17.78	24.24
Social Dialogue f1-score	87.50	88.89	87.50	94.12
Social Discrimination f1-score	75.00	42.11	57.14	53.33
Working Conditions f1-score	78.16	86.42	79.55	81.08
microavg	65.28	64.77	66.42	64.47

Table 11: All experiments results, Subtask B, LAB Dataset

8.4. PROMPTS

As mentioned in the section 4.4 we generated synthetic data using gpt4 . The prompt that we used for generating those synthetic samples is given below:

Sample Artificial Data related to Sustainable Procurement:
Greenway Partners Join Forces with Financing Institutions to Promote Sustainable Production In a bold move towards fostering a greener future, tech giant Greenway Partners has teamed up with several financing entities to reward their suppliers who work towards inclusivity and sustainability. Suppliers would have to prove their commitment towards environmentally friendly practices to get special financing rates. The program's evaluation basis includes an independently made roadmap and classification outline designed in tandem with GOSE, an international charity organization focused on global environmental disclosure. The initiative is expected to incentivize suppliers to cut down their carbon emission, helping Greenway Partners meet its emissions targets. Greenway's collaboration is part of its broader effort to assist clients in achieving their own eco-friendly aspirations. We're thrilled to be a part of Greenway's vision in realizing a sustainable future for all. We firmly believe in achieving net zero emission and we're more than happy to assist Greenway in their significant emissions reduction's strategies says Katy Peterson, director of sustainability programs at one of the financing institutions. Senior corporate banking executive Josh Crawford adds, 'This venture further solidifies our long-term association with Greenway that spans over years and across multiple countries.'

Instruction: You are given an article related to sustainable procurement. You have to generate 2 artificial articles related to sustainable procurement similar to given article. Make sure article should not contain information from below given article.

Article:

{article}

format instructions: {format_instructions}

Figure 3: : Prompt for Sustainable Procurement data

Sample Artificial Data related to External Stakeholder human rights:
In the context of a globally intertwined business environment, the rights and interests of players that extend beyond immediate business circles hold undeniable relevance. These rights often encapsulate diverse aspects that organically pertain to the ethos of international social justice and ethical operations. For example, a company operating across international territories must ensure that the safety and welfare of its workforce is upheld irrespective of its operational decisions. This signifies the company's commitment to creating an inclusive work culture that is driven by diversity and empowerment. On similar lines, corporations have a responsibility towards limiting any adverse environmental ramifications stemming from business operations. This commitment, though presents its challenges, resonates with a higher ethos of global health and sustainability. These concepts shaping the dynamics between businesses and international societal players pave the way towards an equitable environment that drives companies towards innovative and sustainable practices yielding mutual benefits. Therefore, it is the shared responsibility of companies to acknowledge and prioritize these diverse interests of players outside immediate business relationships for the growth of inclusive societies and robust businesses.

Instruction: You are given an article related to external stakeholder human rights. You have to generate 2 artificial articles related to external stakeholder human rights. Make sure article should not contain information from below given article and avoid mentioning directly human rights or external stakeholder.

Article:

{article}

format instructions: {format_instructions}

Figure 4: Prompt for External Stakeholder Human Rights

Sample Artificial Data related to Social Discrimination:

Employment violations related to unfair expectations have surfaced in a recent case involving EnviroLibrium, a leading environmental solutions company headquartered in Houston, Texas. The company is facing accusations from multiple employees who claim it employs unfair practices when addressing employees with certain health conditions. These employees leveled allegations stating that their employer imposed disconnected, counterproductive, and invasive requirements on them, including forced meetings with medical professionals, regardless of their consent or personal medical treatment plans. These behaviors directly oppose fair labor principles and may infringe on employee rights to receive reasonable accommodation for their medical conditions. This case represents the need for employers to comprehend the complexities surrounding employee health conditions and the provision of adaptive support in compliance with their rights. EnviroLibrium has agreed to provide comprehensive training on Americans with Disabilities Act (ADA) to educate its workforce, to avoid such potential missteps in the future. The company, however, denies any allegations of discriminatory practices, asserting their commitment to a diverse and inclusive work environment. This case further emphasizes the necessity of abiding by ADA guidelines, not just towards safeguarding employee rights, but also towards cultivating an inclusive work culture.

Instruction: You are given an article related to social discrimination. You have to generate 2 artificial articles related to social discrimination. Make sure article should not contain information from below given article and avoid mentioning directly social discrimination.

Article:

{article}

format instructions: {format_instructions}

Figure 5: Prompt for Social Discrimination

8.5. Huggingface Model References

'microsoft/deberta-v3-base' :https://huggingface.co/microsoft/deberta-v3-base
'microsoft/mdeberta-v3-base' :https://huggingface.co/microsoft/mdeberta-v3-base
'allenai/longformer-base-4096' :https://huggingface.co/allenai/longformer-base-4096
'Falconsai/text_summarization' :https://huggingface.co/Falconsai/text_summarization
'Helsinki-NLP/opus-mt-en-fr' :https://huggingface.co/Helsinki-NLP/opus-mt-en-fr
'Helsinki-NLP/opus-mt-en-zh' :https://huggingface.co/Helsinki-NLP/opus-mt-en-zh

8.6. Test Set Results

acc.	prec.	rec.	f1-m	f1 -w	language
0.95	0.97	0.95	0.77	0.96	English
0.78	0.87	0.78	0.61	0.81	Chinese
0.94	0.95	0.94	0.87	0.94	French

Table 12: Subtask A test set results per language

task	accuracy	precision	recall	f1 macro	f1 weighted	hamming
ENV	87.53	88.40	87.53	72.48	87.67	87.53
LAB	87.54	90.12	87.54	68.24	87.93	87.54
overall	87.54	89.26	87.54	70.36	87.80	87.54

Table 13: Subtask B, test set results

8.7. Illustrations

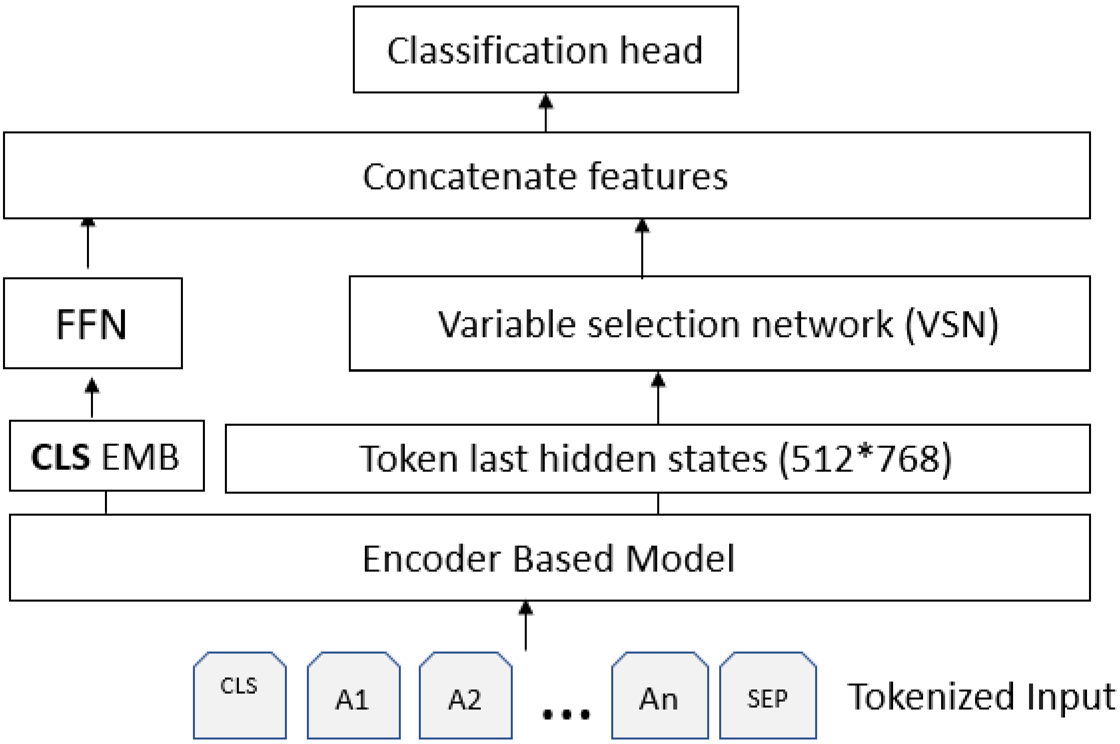


Figure 6: : CLS and VSN Architecture

Author Index