

FinNLP 2023

**The Sixth Workshop on Financial Technology and Natural
Language Processing**

Proceedings of the Workshop

November 1, 2023

©2023 The Asian Federation of Natural Language Processing and The Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-024-0

Message from the Organizers

Welcome to FinNLP, a platform dedicated to promoting international collaboration and the exchange of knowledge in the application of Natural Language Processing (NLP) within the rapidly evolving world of FinTech. Recognizing that some participants might not be able to join us in person, FinNLP maintains a hybrid mode of participation. Whether you join us in person or remotely, we hope that every participant finds value and gains new insights during FinNLP-2023.

Over the recent year, FinNLP has centered its attention on leveraging NLP in finance for social impact. We've introduced several shared tasks. Initiatives like FinSim4-ESG, ML-ESG-1, and ML-ESG-2 have been focused on deriving ESG insights and scoring. The ERAI shared task was created to foster discussions on investor education through opinion scoring. We express our profound gratitude to the organizers, namely Juyeon Kang, Ismail El Maarouf, Yu-Min Tseng, Anais Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and teams from 3DS Outscale and National Taipei University for their pivotal role in curating these datasets. Our aspiration is that these shared tasks shed a unique perspective on FinNLP and stimulate conversations centered on social impact and societal benefit.

This workshop's fruition owes its success to countless individuals, and we offer our deepest thanks to each one. We're immensely grateful to the program committee members who dedicated their time and expertise in reviewing submissions and steering the selection for FinNLP-2023. This includes: Paulo Alves, Emmanuele Chersoni, Ismail El Maarouf, Jinhang Jiang, Juyeon Kang, Chuan-Ju Wang, Yung-Chun Chang, Shih-Hung Wu, Chit-Kwan Lin, Chenyang Lyu, Jinhua Du, Haithem Afli, Hilal Pataci, Min-Yuh Day, Nelson Correa, and Ke Tian. Furthermore, I'd like to convey my appreciation to every attendee. The continued growth and influence of this workshop since its inception in 2019 wouldn't have been possible without your engagement.

In conclusion, our sincere gratitude goes out to Project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO). Their invaluable financial backing has been pivotal, enabling us to achieve the objectives of FinNLP and propel research in this exciting field.

FinNLP-2023 Organizers

Chung-Chi Chen, Hen-Hsen Huang, Hiroya Takamura, Hsin-Hsi Chen, Hiroki Sakaji, Kiyoshi Izumi

Organizing Committee

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

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

Hiroki Sakaji, School of Engineering, The University of Tokyo, Japan

Kiyoshi Izumi, School of Engineering, The University of Tokyo, Japan

Table of Contents

<i>Large Language Model Adaptation for Financial Sentiment Analysis</i> Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut and Jingshu Liu	1
<i>From Numbers to Words: Multi-Modal Bankruptcy Prediction Using the ECL Dataset</i> Henri Arno, Klaas Mulier, Joke Baeck and Thomas Demeester	11
<i>Headline Generation for Stock Price Fluctuation Articles</i> Shunsuke Nishida, Yuki Zenimoto, Xiaotian Wang, Takuya Tamura and Takehito Utsuro	22
<i>Audit Report Coverage Assessment using Sentence Classification</i> Sushodhan Vaishampayan, Nitin Ramrakhiyani, Sachin Pawar, Aditi Pawde, Manoj Apte and Girish Palshikar	31
<i>GPT-FinRE: In-context Learning for Financial Relation Extraction using Large Language Models</i> Pawan Rajpoot and Ankur Parikh	42
<i>Multi-Lingual ESG Impact Type Identification</i> Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu and Hsin-Hsi Chen	46
<i>Identifying ESG Impact with Key Information</i> Le QIU, Bo PENG, Jinghang GU, Yu-Yin HSU and Emmanuele CHERSONI	51
<i>A low resource framework for Multi-lingual ESG Impact Type Identification</i> Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru and Sudip Naskar	57
<i>GPT-based Solution for ESG Impact Type Identification</i> Anna Polyanskaya and Lucas Fernández Brillet	62
<i>The Risk and Opportunity of Data Augmentation and Translation for ESG News Impact Identification with Language Models</i> Yosef Ardhito Winatmoko and Ali Septiandri	66
<i>ESG Impact Type Classification: Leveraging Strategic Prompt Engineering and LLM Fine-Tuning</i> Soumya Mishra	72
<i>Exploring Knowledge Composition for ESG Impact Type Determination</i> Fabian Billert and Stefan Conrad	79
<i>Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models</i> Hariram Veeramani, Surendrabikram Thapa and Usman Naseem	84

Conference Program

Large Language Model Adaptation for Financial Sentiment Analysis

Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut and Jingshu Liu

From Numbers to Words: Multi-Modal Bankruptcy Prediction Using the ECL Dataset

Henri Arno, Klaas Mulier, Joke Baeck and Thomas Demeester

Headline Generation for Stock Price Fluctuation Articles

Shunsuke Nishida, Yuki Zenimoto, Xiaotian Wang, Takuya Tamura and Takehito Utsuro

Audit Report Coverage Assessment using Sentence Classification

Sushodhan Vaishampayan, Nitin Ramrakhiani, Sachin Pawar, Aditi Pawde, Manoj Apte and Girish Palshikar

GPT-FinRE: In-context Learning for Financial Relation Extraction using Large Language Models

Pawan Rajpoot and Ankur Parikh

Multi-Lingual ESG Impact Type Identification

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Yohei Seki, Min-Yuh Day, Teng-Tsai Tu and Hsin-Hsi Chen

Identifying ESG Impact with Key Information

Le QIU, Bo PENG, Jinghang GU, Yu-Yin HSU and Emmanuele CHERSONI

A low resource framework for Multi-lingual ESG Impact Type Identification

Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru and Sudip Naskar

GPT-based Solution for ESG Impact Type Identification

Anna Polyanskaya and Lucas Fernández Brillet

The Risk and Opportunity of Data Augmentation and Translation for ESG News Impact Identification with Language Models

Yosef Ardhito Winatmoko and Ali Septiandri

ESG Impact Type Classification: Leveraging Strategic Prompt Engineering and LLM Fine-Tuning

Soumya Mishra

Exploring Knowledge Composition for ESG Impact Type Determination

Fabian Billert and Stefan Conrad

No Day Set (continued)

Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models

Hariram Veeramani, Surendrabikram Thapa and Usman Naseem

Large Language Model Adaptation for Financial Sentiment Analysis

Pau Rodriguez Inserte¹, Mariam Nakhle^{1,2}, Raheel Qader^{1*}, Gaëtan Caillaut¹, Jingshu Liu¹

¹Lingua Custodia, Paris, France

firstname.lastname@linguacustodia.com

²Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

Abstract

Natural language processing (NLP) has recently gained relevance within financial institutions by providing highly valuable insights into companies and markets' financial documents. However, the landscape of the financial domain presents extra challenges for NLP, due to the complexity of the texts and the use of specific terminology. Generalist language models tend to fall short in tasks specifically tailored for finance, even when using large language models (LLMs) with great natural language understanding and generative capabilities. This paper presents a study on LLM adaptation methods targeted at the financial domain and with high emphasis on financial sentiment analysis. To this purpose, two foundation models with less than 1.5B parameters have been adapted using a wide range of strategies. We show that through careful fine-tuning on both financial documents and instructions, these foundation models can be adapted to the target domain. Moreover, we observe that small LLMs have comparable performance to larger scale models, while being more efficient in terms of parameters and data. In addition to the models, we show how to generate artificial instructions through LLMs to augment the number of samples of the instruction dataset.

1 Introduction

Natural Language Processing (NLP) has become an increasingly important field in the financial industry, with applications ranging from sentiment analysis and named entity recognition to question answering. Information retrieved using machine learning from financial reports, news or posts in social media can be used as indicators of companies' performance or as insights of a market. Many industry actors are interested in extracting this information to use it as a resource that can provide them

with a competitive advantage, such as firms forecasting internal future benefits and losses, investors extracting differential information for trading purposes or any practitioner interested in tracking financial assets. Nevertheless, some characteristics of financial text make these tasks especially challenging for models that have been trained on general domain data. The use of specific terminology along with the high complexity of the documents, leads these generalist language models to underperform on financial tasks, which suggests that domain adaptation might be required to improve accuracy of interpretation and analysis.

Furthermore, the rapid evolution of large language models (LLMs) and their proven capabilities for NLP tasks has made them stand out and become an interesting option to study. Due to the fact that even the best general language models fall short for some financial tasks, some proposals have been recently presented for a financial domain adaptation of LLMs. These models tailored for finance, such as BloombergGPT (Wu et al., 2023), have been introduced as multitasking generative models specifically designed for financial text understanding and generation. However, these fine-tuned models still show room for improvement, both in performance and in the efficiency of the proposed training strategies.

This paper tackles various aspects of adapting LLMs to the financial domain. In particular, we explore diverse strategies of domain adaptation and fine-tuning of LLMs for financial sentiment analysis, and conduct a series of experiments over two different foundation models. The study focuses particularly on smaller manageable models, up to 1.5B parameters, in order to explore the possibilities of models that can be accessible with relatively low hardware requirements. Although the adapted models are smaller than the current state-of-the-art ones, results show that they achieve similar or higher performance. In addition, a curated data collection

*Corresponding author.

with two main datasets is also presented. One constructed with financial documents and reports, and the other a set of instructions for financial tasks. We show, step by step, the process of creating these datasets and particularly focus on the use of more powerful LLMs to generate synthetic instructions to fine-tune smaller LLMs. Finally, apart from the main focus of the study which is on financial sentiment analysis, other tasks have also been evaluated to analyze the multitasking capabilities of our models.

2 Related work

Sentiment analysis is one of the most common use cases of NLP. In this task, a model classifies a text according to the sentiment detected, usually between *positive*, *negative* and *neutral*. However, while any sentiment analysis model would be capable of undertaking financial sentiment analysis, an adaptation is required. In this section, the evolution of sentiment analysis in the financial domain is studied using models based on Transformers (Vaswani et al., 2023).

FinBERT¹ (Araci, 2019) is based on the idea of training a BERT (Devlin et al., 2018) model in two steps to adapt it to the financial domain and the sentiment analysis task. The first step consists of further pre-training the model on financial documents, as this strategy has already been proven to be effective (Howard and Ruder, 2018) for domain adaptation. This step aims at helping the model to understand financial terminologies better than the base model. Authors used a subset of Reuters' TRC24 dataset², a collection of news articles published by Reuters that was filtered with keywords related to finance, to fine-tune the model. In the second step, the model is prepared for the sentiment analysis task by adding a dense layer to the last hidden state of the classification token CLS of the encoder-based architecture, a recommended practice for classification with BERT. This task is fine-tuned using the Financial PhraseBank (FPB) (Malo et al., 2014), a financial sentiment analysis dataset. FinBERT presents remarkable results on financial sentiment analysis, outperforming the state-of-the-art. Nevertheless, the model is strongly limited to sentiment analysis and underperforms greatly on other tasks.

¹Several models under the name of FinBERT exist, however in this work we only discuss the first model.

²<https://trec.nist.gov/data/reuters/reuters.html>

2.1 Base large language models

Recent advances in the field of large language models (LLMs) have shown that these models can achieve remarkable capabilities in understanding complex natural language. They are also capable of performing zero-shot and few-shot learning, in which they can generate accurate responses for tasks that they have not seen during training (Radford et al., 2019). This makes LLMs a great choice in multitask settings where one model is expected to perform several tasks. Most of today's LLMs are based on Transformer models (Vaswani et al., 2023), typically set in decoder-only architectures.

Training of LLMs is typically split in two stages. The first part of the training is the most computationally expensive since the model is trained using large amounts of text. For this reason, conducting the training of a LLM from scratch requires high computational resources. Nevertheless, many research groups and companies are releasing these models to the public to be used as base or foundation for other models, to enable research to move forward. Using these pre-trained models is highly beneficial for researchers with fewer data or hardware resources, as they can be used as a starting point for fine-tuning on specific tasks, such as chatting, following instructions or giving outputs in a specific style or format. Although most LLMs are trained on general domain data, there have been a few works recently to adapt LLMs to the financial domain. In the next subsections two such work are reviewed.

2.2 Financial large language models

BloombergGPT. One of the first decoder-only LLMs trained specifically for finance is BloombergGPT (Wu et al., 2023), a model of 50B parameters based on BLOOM's architecture (Scao et al., 2023). The corpus collected for the training of this LLM consisted in the combination of 363 billion tokens from financial documents with 345 billion tokens from general purpose datasets. The model was trained from scratch, without using any foundation model as a base, with the objective of predicting the next token of the documents, and without fine-tuning on instructions. However, the results presented by BloombergGPT are far from the ones achieved by other models, some of them of a much smaller scale. In addition, the results reported did not outperform other generalist LLMs, as we will show later in this paper.

FinMA. The open model FinMA from PIXIU’s framework (Xie et al., 2023) introduced by Chance-Focus reported better scores on several financial tasks than larger generalist LLMs, such as GPT-4 (OpenAI, 2023), and BloombergGPT. They used LLaMA (Touvron et al., 2023a) as the pre-trained model and fine-tuned it with instructions tailored for financial multitasking. The instruction dataset consists of texts formed by an instruction, an input and an answer. The dataset includes a data augmentation strategy in which the inputs of those tasks with few samples were used with 10 different instructions. This augmentation strategy, while increasing the number of samples in the dataset, did not increase its diversity as the same set of 10 instructions were always repeated.

In the same paper in which FinMA was presented, the PIXIU framework also included FLARE, a financial evaluation benchmark. This benchmark has been used to evaluate the experiments carried out in this project.

2.3 Financial benchmark

For the evaluation of large language models, the FLARE benchmark³ from PIXIU framework has been used. The tasks of this benchmark which are relevant to our work are presented below.

Financial Sentiment Analysis. Financial sentiment analysis task over two different benchmarks, the Financial Phrase Bank (FPB) (Malo et al., 2014) and FIQA-SA (Maia et al., 2018).

News Headline Classification. Headlines task contains 9 different subtasks, each one associated with 9 different gold questions, in which the expected answers are “yes” or “no”. The inputs analyzed are gold news from the Gold dataset (Sinha and Khandait, 2020).

Named Entity Recognition. NER task is based on detecting financial named-entities in U.S. public agreements in the (Salinas Alvarado et al., 2015) dataset. The tagged entities correspond to people, organizations and locations.

3 Methodology

In this section, we describe the methods designed to conduct the experiments. First, we list the foundation models that are used as a part of this project. Then, two new dataset collections are introduced, one with data based on documents and the second with instructions. We also give details of designing

a data augmentation strategy for the instructions as well as the description of the training process carried out to fine-tune the foundation models.

3.1 Foundation models

As stated earlier, the focus in this work is on smaller sized models that can be adapted to achieve performance of larger models. The two models that we use are listed below:

OPT. Meta AI’s large language models suite OPT (Open Pre-trained Transformers) (Zhang et al., 2022) were presented as a collection of 9 models ranging from 125M to 175B parameters, being one of the first publicly available LLMs.

Pythia. EleutherAI presented Pythia (Biderman et al., 2023), a suite of decoder-only language models with sizes ranging from 70M to 12B parameters. These models are trained on the Pile dataset (Gao et al., 2020), a curated collection of English texts from a wide variety of sources.

3.2 Datasets

In order to train LLMs, two main different approaches can be taken with respect to data. When a model is trained from scratch, the data used are collections of documents, for which the model has the objective of predicting the next token. This is usually the training carried out to obtain foundation LLMs. However, these models can be further pre-trained for domain adaptation, in the same way that FinBERT was trained. This approach is based on the idea of continuing the training of the model with financial documents to shift from a general to a financial language model. Moreover, it has been proven that large language models can improve their performances, especially on unseen (or zero-shot) tasks by fine-tuning them to follow instructions. For this fine-tuning method, the training objective is the same, predicting the next token of the text, with the only difference being that the format of this data relies on an “instruction”, “input”, “answer” format. For this project, one dataset was collected for each of these two training strategies. In addition, the instruction-based dataset was augmented artificially with samples generated from another LLM (LLaMA 2 13B (Touvron et al., 2023b)).

Document dataset. The collection of documents used to further pre-train the base LLMs is a combination of general and financial documents from different sources. The purpose of this mixture is to add diversity to the training data, with finance

³<https://github.com/chancefocus/PIXIU>

being the most represented domain. Having general data in the training set prevents the model to completely drift the domain and result in a model that is unable to understand general language. The data sources of these documents are described below:

- **EDGAR Files** (Financial). EDGAR is the Electronic Data Gathering, Analysis, and Retrieval system online platform operated by the SEC⁴ (United States Securities and Exchange Commission). It is used by companies to electronically file registration statements, periodic reports, and other forms required by the SEC. The database of these documents is open to everyone, allowing the retrieval of high-quality financial text.
- **Reuters News** (Financial). Reuters is a news agency specializing in business and finance that released Reuters Corpora, a collection of financial news made available for use in NLP research. The collection used in this dataset is TRC2 (Thomson Reuters Text Research Collection), that contains more than 1.8 million news.
- **In-house Dataset** (Financial). As a part of this project, a diverse collection of in-house financial text has been obtained. The text in this dataset is mostly at sentence level, as they were originally used for machine translation. This is the only private set used for the project.
- **The Pile** (General). The Pile (Gao et al., 2020), from EleutherAI, is a dataset that comprises 22 diverse high-quality subsets, several of which originate from academic or professional sources. The idea behind this dataset’s construction is that diversity enhances general cross-domain knowledge and downstream generalization capability of large language models. It includes data from general news to scientific articles, code, etc. . . The proportions of these subsets are kept as-is in the sub-sample used for this project.

The lengths of the documents of this dataset had to be adapted to the models’ context length, which corresponds to the longest sequence of tokens that the model can support. In this project the context was limited to 2048 tokens. The pre-processing of

⁴<https://www.sec.gov/edgar/searchedgar/companysearch>

the dataset consisted in concatenation of all documents from the same source, using a special token (<endoftext>) to separate them. The long concatenated text is then sliced in blocks of 2048 tokens, which are mixed and shuffled with all the other blocks of the dataset. Since some datasets used in this project are extremely large, we decided to take a smaller proportion from each one. The summary of the ratio used for each partition is shown in Table 1.

Subset	Domain	# Tokens	%
EDGAR	Finance	100k	25.7
Reuters	Finance	36k	9.3
In-house	Finance	38k	9.7
The Pile	General	215k	55.3
Total		389k	100

Table 1: Proportion and absolute number of tokens taken from each dataset.

Instruction-based dataset. Instruction fine-tuning is a strategy used to improve LLMs’ performance for specific tasks by teaching them to follow specific format of questions and answers. LLMs learn by being trained on this specific format of text, while keeping the same training objective, predicting the next sequence of tokens. Fine-tuning on instructions is the most common technique to adapt foundation models to specific use-cases, mainly because this method not only improves performance on the trained tasks, but also augments zero-shot and few-shot capabilities. Models trained on instructions are usually consistent in the format in which data is presented. In Table 2 the format used for our dataset is displayed.

Template
Instruction: Description of the task
Input: Input to analyze
Answer: Answer to predict

Table 2: Template for instructions.

To create the instruction dataset, we used the instructions dataset published by PIXIU that targeted the FLARE benchmark. However, this dataset has poorly curated prompts and includes a suboptimal data augmentation strategy. For instance, certain parts of the dataset have been up-sampled by using ten different instructions over the same input. Despite having more samples, the up-sampled version

Subset	Instr	PIXIU	Augm
FPB	4,838	48,380	6,633
FiQA-SA	1,173	11,730	2,825
NER	609	6,090	2,609
Headline	102,708	102,708	102,708
FinQA	8,242	8,242	8,242
ConvFinQA	3,892	3,892	3,892
BigData22	0	7,164	0
ACL18	0	27,053	0
CIKM18	0	4,967	0
TOTAL	121,462	220,226	126,909

Table 3: Comparison of the instructions used for each task in the original instructions dataset (*Instr*) with one input-answer by instruction, the up-sampled version (*PIXIU*), and the dataset augmented by LLM inference (*Augm*).

of dataset lacks diversity which may lead to poor performance as discussed by Zhou et al. (2023). For this reason, the dataset proposed in this project has been designed from scratch, only reusing the unique *input - answer* pairs from a down-sampled version of PIXIU’s dataset that includes a single instruction for each input.

Instruction data augmentation. The main idea behind instruction data augmentation is to bring new inputs to the dataset, so the model has more diverse examples to learn from. Two different methods have been defined for generating these instructions dependent on the target task. For sentiment analysis, the model has to generate an input for a given sentiment given an example with that label. This strategy has been used to augment both FPB and FIQA-SA subsets. In Table 7 of Appendix A the template used for this task is presented.

For NER, since it is not a sequence classification task, the inference method is different. The first solution proposed was based on letting the model generates both the new sentence and its NER tags at the same time, only guiding the model by including a few examples in the prompt. However, the variety of the sentences generated by the model was too short and the tags were incorrect, indicating that the task was too hard for the model. Our solution was to use existing unlabeled sentences, which reduced the generative task to a tagging process. The sentences used for this augmentation were in-house financial sentences. Moreover, in this case the example given to the model is fixed in order to make sure all types of entities are present in the prompt. The format of the tagging was chosen using prompt

engineering. In Table 8 of Appendix A, the template used for NER data augmentation is shown. The Headline task was not augmented since it had enough samples, even when considering that there are 9 subtasks in the benchmark.

Using the above-mentioned two templates, new inputs are inferred to be added to the instructions dataset. The model used for the generation of these new samples is LLaMA-2-13B, quantized in 4-bits to reduce the GPU memory required. Table 3 shows a comparison of the number of samples targeting each task before and after the augmentation. The decision on the number of synthetic samples generated was taken considering the number of original samples. The reason behind not generating even a larger number of instructions is that despite their high-quality, artificial samples could introduce some noise to the dataset by generating sentences too different from the original distribution, introducing erratic *input-answer* pairs or NER tags or to duplicate some inputs after several iterations. In Table 3 there is a comparison between PIXIU’s dataset before and after down-sampling as well as after the data augmentation.

3.3 Training method

As stated earlier, in this work, we use two pre-trained foundation models, namely Pythia-1.4B and OPT-1.3B, and fine-tune them in two stages as detailed below:

- **Further pre-training.** The models are fine-tuned to predict the next token of the text in the document-based dataset, following the same idea as in FinBERT and without being fine-tuned on a specific task. The idea here is to tilt the models to become more familiar with the financial domain. Both models are trained on a total of 389,000 tokens introduced in context blocks of 2,048 tokens. The models are trained for two epochs, saving 4 checkpoints at every epoch. The best checkpoint is selected for each model.
- **Instruction fine-tuning.** The model is instructed to perform financial tasks using the instructions dataset. Since the length of these instructions is generally shorter than on the document-based dataset, the context length is reduced to 1,000 tokens to speed up the training. Sequences shorter than this length are padded with a padding token. Instructions

longer than 1,000 tokens are cut off. For instruction fine-tuning, models are trained for 1 epoch.

For both set-ups, training is performed with AdamW optimizer (Loshchilov and Hutter, 2019), a batch size of 32, and applying gradient accumulation of 4 for training efficiency. The initial learning rate is set to 1e-4, while the weight decay is adjusted to 0.1. These values remained the same for all the conducted experiments. The training of these models was carried out on a H100 GPU.

4 Results

4.1 Classical algorithms versus LLMs for financial sentiment analysis

Prior to conducting an evaluation of our fine-tuned LLMs for financial sentiment analysis, we study the performance of the current state-of-the-art models and classical machine learning algorithms. Compared to LLMs, classical algorithms do not require a lot of computation, they could be easily trained and tested. For the sake of simplicity, the evaluation has been carried out only on the FPB. Based on the results in Table 4, the lowest score, unsurprisingly, is obtained by the lexicon approach. Classical machine learning algorithms on the other hand are able to obtain results considerably higher than lexicon, and even match or pass LLM scores in some cases. Overall conclusions that can be depicted from the results can be summarized as follows:

- The domain adaptation and training of FinBERT on this specific task, gives the model an advantage over general models. Comparing FinMA-30B with GPT-4, it can be seen that a smaller model fine-tuned for finance has better performance than a generalist one.
- BloombergGPT was a good starting point for financial LLMs. However, its performance on tasks like sentiment analysis is poor. One likely reason is that this model has not been fine-tuned on instructions.
- FinMA-30B proves the relevance of fine-tuning on instructions to improve performance on financial tasks. Nevertheless, as mentioned before, the train dataset might be not sufficiently diverse, which may impact the model’s capability in real-world scenario.

Algorithm and features	Accuracy
Lexicon approach	
Loughran-McDonald dictionary	0.59
Classical ML algorithms	
SVM	0.77
Naive Bayes	0.73
XGB	0.80
Transformers approach	
FinBERT	0.85
GPT-4	0.71
BloombergGPT	-
FinMA-30B	0.87

Table 4: Performance of financial sentiment analysis. A comparison between traditional approaches and modern transformer based models.

4.2 Financial domain adaptation

In this section we show the impacts of the two stage fine-tuning as well as improvements brought by the artificially augmented instruction dataset. Models are evaluated using a subset of the tasks proposed in the FLARE benchmark: FPB, FIQA-SA, Headlines and NER. For the classification tasks (FPB, FIQA-SA and Headlines), the predictions are obtained by forcing the model to generate one of the expected class label. For example, in FPB, this means to choose the next token only amongst the ones needed to generate the labels (*positive*, *negative* or *neutral*), and sticking with the most probable ones (the highest logits). When evaluating on NER, the generation is not constrained.

The first experiment that we carry out is to see the effects of fine-tuning on documents versus instructions. Based on the results of Table 5, it is clear that performance of both Pythia and OPT models show similar behaviors and that fine-tuning brings significant improvements over the base models. Particularly, instruction fine-tuning improvement is much higher than just further pre-training on documents. This conclusion seems to be aligned with what we observe in the literature of instruction tuning of other domains.

Next, in order to evaluate the effect of augmenting the number of instructions using the strategy designed for this dataset, the models are compared after fine-tuning with the base instructions dataset and with the augmented instructions dataset. The results of using data augmentation are not straightforward. In Table 5, it can be seen that the performance of the models augmented instructions is improved for the sentiment analysis tasks, but the scores goes down for the other two. In the

		F1 scores			
Fine-tuning data		FPB	FIQA-SA	Headlines	NER
Pythia-1.4B	<i>(base)</i>	0.20	0.29	0.16	0
	<i>Docs</i>	0.41	0.16	0.57	0.30
	<i>Instr</i>	0.82	0.73	0.93	0.59
	<i>Augmented instr</i>	0.82	0.79	0.90	0.56
	<i>Docs + Augmented instr</i>	0.84	0.83	0.97	0.69
OPT-1.3B	<i>(base)</i>	0.19	0.48	0.29	0
	<i>Docs</i>	0.13	0.58	0.39	0
	<i>Instr</i>	0.84	0.77	0.93	0.53
	<i>Augmented instr</i>	0.86	0.79	0.97	0.29
	<i>Docs + Augmented instr</i>	0.86	0.81	0.96	0.34

Table 5: Comparison of Pythia-1.4B and OPT-1.3B fine-tuned with different strategies. The results reported correspond to the base models without fine-tuning (*base*), models with document further pre-training (*Docs*), models fine-tuned on instructions (*Instr*), models fine-tuned on augmented instructions dataset (*Augmented instr*), fine-tuning first with documents and then with augmented instructions (*Docs + Augmented instr*).

case of Headlines, this effect can be caused by the fact that this task is the most represented in the dataset and, by introducing new samples, the model is less focused on this task. For NER, the issue can be explained by the difference between the text of the synthetic samples and the original test set. As explained in previous sections, NER is augmented using in-house data, and even though the chosen sentences were also in the financial domain, the sources are different and that might have introduced errors in the predictions.

Finally, we can test the implications of instruction fine-tuning after further pre-training the model with the financial documents. This simply means that the model is fine-tuned two times. Since the augmented instructions proved to be better than the original instructions, this experiment is conducted on the earlier instruction dataset. As shown in the last row of Table 5, this approach seems to lead to a higher score in every task. Therefore, the domain adaptation method inspired by FinBERT’s training strategy, proves to be effective for decoder-only LLMs and not only for financial sentiment analysis, but for multiple financial NLP tasks.

4.3 Comparison with other Financial LLMs

In this section, the results of the best models obtained through the previous experiments (*Docs + Augmented instr*) are compared against the state-of-the-art LLMs for finance. As can be seen in Table 6, both fine-tuned Pythia-1.4B and OPT-1.3B over perform GPT-4 in classification tasks, which includes financial sentiment analysis. This is made possible because of the domain adaptation conducted for these two base models. For NER, which

is a generative task, GPT-4 is still the LLM with the highest score. When the models of these projects are compared to BloombergGPT, the biggest current LLM tailored for finance, it can be observed that the scores obtained are much higher for classification tasks, specially for sentiment analysis, and that Pythia also obtains better score for NER. In terms of efficiency, these results are achieved with models that have approximately 97% fewer training parameters than BloombergGPT⁵. In the comparative with the collection of FinMA models, the PIXIU LLMs still outperform the models fine-tuned with our domain adaptation strategy in some tasks, specially when compared to FinMA-30B. However, when FinMA-7B, the model with the closest size to the models presented in this project, is evaluated in financial sentiment analysis and Headlines, it can be observed that the scores are almost equivalent to the fine-tuned Pythia-1.4B and OPT-1.3B. In this case, however, the biggest improvement with respect to FinMA-7B is in terms of efficiency. Pythia-1.4B and OPT-1.3B have approximately 78% fewer training parameters than FinMA-7B, and the number of instructions used goes from 220,226 down to 126,909, which is only a 57% of the number of samples used for PIXIU models.

Therefore, from the general comparison it can be seen that the models fine-tuned in this project over perform most LLMs in financial tasks, with the only exception of FinMA models. In addition, the size of the models and the training strategy

⁵BloombergGPT has 50B trainable parameters. Pythia-1.4B and OPT-1.3B have approximately 1.5B parameters. The amount of data used is not comparable since BloombergGPT was trained from scratch.

Models	F1 scores			
	FPB	FIQA-SA	Headlines	NER
BloombergGPT	0.51	0.75	0.82	0.61
GPT-4	0.78	-	0.86	0.83
FinMA-7B	0.86	0.84	0.98	0.75
FinMA-30B	0.88	0.87	0.97	0.62
Pythia-1.4B	0.84	0.83	0.97	0.69
OPT-1.3B	0.86	0.81	0.96	0.34

Table 6: Comparison of state-of-the-art with Pythia-1.4B and OPT-1.3B fine-tuned on documents and the augmented dataset. Performance of models retrieved from BloombergGPT and PIXIU papers. BloombergGPT is not a publicly available model, so it is not possible to evaluate it under the same conditions as the other models. Thus, ChatGPT, GPT-4 and FinMA-30B are evaluated on zero-shot, BloombergGPT is only reported in a five-shot setting and its accuracy was not published.

have been proven to be more efficient than the ones proposed for other models.

5 Conclusion

This project has covered a wide range of aspects of financial LLMs. Through a series of experiments, using Pythia-1.4B and OPT-1.3B as base models, we studied the adaptation of relatively small LLMs for finance. The experiments we conducted first show that LLMs adapted to the financial domain through further pre-training followed by instruction fine-tuning perform better than some of the best current generalist LLMs (such as GPT-4) on financial tasks. Second, it validates our training strategy since our LLMs obtain higher or similar scores than other financial LLMs that were trained with much more parameters and larger datasets.

Lowering the requirements to fine-tune LLMs for this specific industry can be key for the future of several companies, since it can enable smaller organizations to host their own LLMs or, at least, to make them more accessible. Furthermore, it is worth mentioning that the models used for this project as well as most datasets, except for the in-house subset (only 9.7% of the documents dataset), are open and publicly available. In addition to the findings related to domain adaptation of LLMs for financial tasks and the models presented, a strategy for the generation of samples for the instructions dataset is introduced. Moreover, the two datasets used for the project are described with enough details to be reproduced by other researchers. Finally, the paper also presented a comprehensive study that delves into the state-of-the-art and the evolution of approaches for financial sentiment analysis, ranging from traditional dictionary-based methods to the more recent advancements in LLMs.

Despite the fact that the results showed great performance of the small-sized models, in further research these fine-tuning strategies could be applied to larger models and study their impact on different scales and domains. An interesting option to study in the future are Low-Rank Adapters or LoRA (Hu et al., 2021), a method that reduces the number of trainable parameters by freezing the foundation model weights and injecting trainable rank decomposition matrices into each layer of the LLM.

Limitations

The limitations of this work can be summarized as following:

- **Generative capabilities:** The final fine-tuned model seems to perform very well on classification tasks such as sentiment analysis, while still lagging behind in generative ones.
- **Unseen tasks:** our work concentrates on certain tasks that have been studied in previous similar work, but for a full understanding of its limitations, one needs to test it on unseen tasks.
- **Large models:** we believe that testing the same strategy of multiple fine-tuning stages would yield even better results with larger models such as LLaMA-2-7B or even larger models.

References

Dogu Araci. 2019. [Finbert: Financial sentiment analysis with pre-trained language models.](#)

- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. [Pythia: A suite for analyzing large language models across training and scaling](#).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#).
- Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#).
- 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](#).
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#).
- 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 Walenius, 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](#).
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Julio Cesar Salinas Alvarado, Karin Verspoor, and Timothy Baldwin. 2015. Domain adaption of named entity recognition to support credit risk assessment. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 84–90, Parramatta, Australia.
- Teven Le Scao, Angela Fan, Christopher Akiki, Elie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and et al. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#).
- Ankur Sinha and Tanmay Khandait. 2020. [Impact of news on the commodity market: Dataset and results](#).
- 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. 2023a. [Llama: Open and efficient foundation language models](#).
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu 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. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. [Attention is all you need](#).
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. 2023. [Bloomberggpt: A large language model for finance](#).
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. [Pixiu: A large language model, instruction data and evaluation benchmark for finance](#).
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. [Opt: Open pre-trained transformer language models](#).
- Chunting Zhou, Pengfei Liu, Puxin Xu, Srinu Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. [Lima: Less is more for alignment](#).

A Instruction data augmentation examples

Template

Write a sentence with a $\{y_i\}$ financial sentiment. Use the format `<stc> sentence </stc>`. Reuse terms from the example. Example: '`<stc> $\{x_i\}$ </stc>`'

Example

Write a sentence with a **positive** financial sentiment. Use the format `<stc> sentence </stc>`. Reuse terms from the example. Example: '`<stc> Shares of Standard Chartered (STAN) rose 1.2 % in the FTSE 100 , while Royal Bank of Scotland (RBS) shares rose 2 % and Barclays shares (BARC) (BCS) were up 1.7 % . </stc>`'

Table 7: Template for sentiment analysis input generation with financial sentiment fixed and dynamic shot. Example given for positive input inference. $\{x_i, y_i\}$ is an *input-answer* pair sampled from one of the two subsets.

Template

Identify the named entities that represent a person ('PER'), an organization ('ORG'), or a location ('LOC') in a financial context. Use the format 'Entities: entity name, entity type'.

Sentence: 'The Bank gave money to the Borrower to open a business in New York.'; Entities: 'Bank, ORG | Borrower, PER | New York, LOC'

Do the same with this sentence, identifying 'PER', 'ORG', 'LOC' entities.

Sentence: $\{x_i\}$; Entities:

Example

Identify the named entities that represent a person ('PER'), an organization ('ORG'), or a location ('LOC') in a financial context. Use the format 'Entities: entity name, entity type'.

Sentence: 'The Bank gave money to the Borrower to open a business in New York.'; Entities: 'Bank, ORG | Borrower, PER | New York, LOC'

Do the same with this sentence, identifying 'PER', 'ORG', 'LOC' entities.

Sentence: '**350 , Wellesley , Massachusetts 02481 doing business as " Silicon Valley East " and AKAMAI TECHNOLOGIES , INC . (" Borrower ") , whose address is 201 Broadway , 4th Floor , Cambridge , Massachusetts 02139 provides the terms on which Bank will lend to Borrower and Borrower will repay Bank**'; Entities:

Table 8: Template for NER tags generation given a sentence of the financial domain.

From Numbers to Words: Multi-Modal Bankruptcy Prediction Using the ECL Dataset

Henri Arno^{1*} and Klaas Mulier¹ and Joke Baeck¹ and Thomas Demeester²

¹Ghent University

²Ghent University - imec

Henri.Arno@UGent.be

Abstract

In this paper, we present ECL, a novel multi-modal dataset containing the textual and numerical data from corporate 10K filings and associated binary bankruptcy labels. Furthermore, we develop and critically evaluate several classical and neural bankruptcy prediction models using this dataset. Our findings suggest that the information contained in each data modality is complementary for bankruptcy prediction. We also see that the binary bankruptcy prediction target does not enable our models to distinguish next year bankruptcy from an unhealthy financial situation resulting in bankruptcy in later years. Finally, we explore the use of LLMs in the context of our task. We show how GPT-based models can be used to extract meaningful summaries from the textual data but zero-shot bankruptcy prediction results are poor. All resources required to access and update the dataset or replicate our experiments are available on github.com/henriarnoUG/ECL.

1 Introduction

Bankruptcy has far-reaching consequences that extend beyond the business owners, affecting various stakeholders such as employees, suppliers and creditors. On an economy-wide scale, bankruptcy risk plays a structural role in propagating recession (Bernanke, 1981). Predicting the occurrence and timing of this corporate event precisely is challenging, due to the external factors and complex financial dynamics at play. Yet, certain warning signals, such as decreasing revenues and rising debt, can serve as an indication of imminent bankruptcy. Therefore, several researchers have directed their efforts towards the development of sound bankruptcy prediction models in the past decades (Beaver, 1966; Ohlson, 1980). Increasingly advanced prediction models (Odom and Sharda, 1990; Kim and Kang, 2010), combined with well-chosen, informative features (Mai et al.,

2019), have led to increased predictive performance in the field.

In this paper, we contribute to the literature in two ways. First, we present ECL, a new dataset that contains the textual and numerical data from corporate 10K filings (cf. Section 2) and associated binary bankruptcy labels. It is a unique compilation of three existing data sources: the EDGAR-corpus (Loukas et al., 2021), CompuStat¹ and the LoPucki Bankruptcy Research Database.¹ Second, we present baseline bankruptcy prediction models on each data modality, as well as on the combination, and critically evaluate their performance. Based on our findings, we identify and formulate interesting avenues for future research.

In recent work (Arno et al., 2022), we argue that contributions in the field of bankruptcy prediction are difficult to compare since (1) the considered evaluation scenarios vary strongly (which is related to the temporal nature of the data and the class imbalance) (2) there is no consensus on key evaluation metrics and (3) there is a lack of benchmark datasets. We introduced a carefully designed evaluation strategy, applied to a text-only benchmark dataset. In this paper, we adopt this evaluation setup, report the suggested evaluation metrics for our baseline models on ECL, and share our code and dataset to encourage reproducible future research (see github.com/henriarnoUG/ECL).

Our findings suggest that the textual and numerical content from a 10K contain complementary information for bankruptcy prediction. In some cases, the management of a company explicitly state that they consider filing for bankruptcy, making the prediction task based on text trivial.

¹URLs accessed 2023-10-05:

[https://www.marketplace.spglobal.com/en/datasets/computat-financials-\(8\)](https://www.marketplace.spglobal.com/en/datasets/computat-financials-(8)) and <https://lopucki.law.ufl.edu>

If this is not mentioned, the accounting figures are more informative for bankruptcy prediction. Furthermore, the results show that our models trained on binary labels cannot distinguish 10K records filed in the year before bankruptcy from those records filed by financially unhealthy companies that did not file for bankruptcy just yet. Based on this finding we argue that modelling the financial health of a company with a more gradual label is an interesting direction for future research. Finally, we explore the potential of LLMs in the context of our task. We show that GPT-generated summaries from the text contained in the 10K filings are useful for bankruptcy prediction. Despite this promising result, we find that the zero-shot prediction results of GPT-3.5 are significantly worse than the results of a simple keyword-based TF-IDF model.

The structure of this paper is as follows. In Section 2 we present our dataset and discuss the prediction task. Section 3 contains an overview of our experimental setup. The results, along with an in-depth qualitative analysis, are presented in Section 4. The potential of LLMs for our task is explored in Section 5 while Section 6 concludes.

2 The ECL Dataset

Large companies operating in the U.S. are required to submit a variety of filings with the Securities and Exchange Commission (SEC) throughout the year. Potential investors and other stakeholders use these filings to gain insight into the financial performance, business operations, risks and other aspects of the company of interest. Notably, the most widely consulted SEC filing is the Form 10K, which is reported annually and contains detailed information on a company’s past fiscal year. A 10K filing is composed of 15 different items including a description of the business (item 1), the management discussion and analysis (item 7) and a section on executive compensation (item 11), among others. Item 8 of a Form 10K contains the consolidated financial statements such as the balance sheet, the income statement and the cashflow statement. We carefully compiled the EDGAR-CompuStat-LoPucki dataset, further referred to as ECL, containing data in two modalities (i.e., textual and numerical) from such 10K records that companies filed with the SEC in the past. We present the dataset in the context of our current

work on bankruptcy prediction, but are convinced that the multi-modal dataset has other possible uses, in terms of analysis or predictive modelling of a companies’ financial and business situation.

2.1 Data Sources

Most SEC filings, including the Form 10K, are publicly available through the Electronic Data Gathering, Analysis and Retrieval (EDGAR) website as a text file or as an XBRL file (an HTML based document type). The same 10K data can be accessed through a variety of other sources. ECL is a unique compilation of three existing data sources: the textual data is collected from (1) the EDGAR-corpus (Loukas et al., 2021), the numerical financial data is gathered from (2) CompuStat² while (3) the LoPucki BRD provides the labels for the bankruptcy prediction task.

2.2 Dataset Construction and Labelling

Some firms are required to file a 10K every year, such as companies whose stock is traded on a U.S. stock exchange, while others voluntarily submit 10K filings. Using the EDGAR-crawler tool,³ we have collected the textual data (and corresponding metadata) from all 10K filings on the EDGAR website from 1993⁴ onwards.

Next, we add the structured, financial information, reported in item 8 of a 10K, to the dataset by linking the collected 10K records from the previous step to CompuStat records. We use the CompuStat Fundamentals North-America table and filter out the records that originate from sources other than the Form 10K (some records are collected from the prospectus, the annual letter to the shareholders, Form 20-F, ... etc.). We merge the collected 10K records and the filtered CompuStat records based on two conditions. First, matching records must have the same company name or company identifier (the Central Index Key) and second, the fiscal year end (the date) of the records must lie within 7 days of each other.⁵ Remaining 10K records or CompuStat records left unmatched are discarded.

²In order to use ECL, access to CompuStat is required. For details, we refer to our GitHub repository.

³Available at: github.com/nlpaueb/edgar-crawler

⁴This is the starting point of the EDGAR-corpus as well.

⁵Our analysis revealed that, due to data quality issues, the fiscal year end can be a couple of days off in CompuStat.

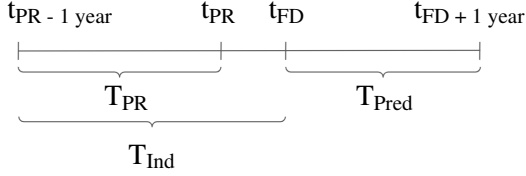


Figure 1: The labelling strategy for the 10K records in our dataset for the next year bankruptcy prediction task.

Finally, we assign the labels for the next year bankruptcy prediction task following our proposed labelling strategy (Arno et al., 2022). We collect the bankruptcy data from the LoPucki Bankruptcy Research Database (BRD). This dataset contains the exact date on which companies filed for bankruptcy under chapter 7 or chapter 11 of the U.S. bankruptcy code. Only firms that (1) submit 10K filings with the SEC and (2) have a total asset value exceeding 1,000,000, measured in 1980 dollars, qualify for inclusion in the LoPucki BRD. Data is available for all bankruptcies between 1979 and the end of 2022. Before we assign labels, we tag the 10K filings in ECL based on these criteria. The total asset value reported in the 10K must exceed the (inflation-corrected) threshold and the filing date must lie within the correct time frame.

Each qualified 10K record in the dataset is assigned a binary label. A 10K filing covers a fiscal year (T_{PR}), which concludes on the fiscal year end (t_{PR}), and is released to the public on the filing date (t_{FD}), after the filing period. The bankruptcy label is true if the company filed for bankruptcy in the year following the filing date (i.e. during T_{Pred}) and false otherwise. The task is to predict whether a company will file for bankruptcy in the next year, given the multi-modal data contained in the 10K filing (covering the period T_{Ind}). This labelling strategy is graphically depicted in Figure 1.

2.3 ECL Statistics

ECL consists of 170,139 Form 10K filings for which numerical and textual data is available. From the 277,940 collected 10K filings and the 241,825 filtered CompuStat records, 107,801 and 71,686 records remain unmatched, respectively. The vast majority (over 56%) of these unmatched records come from companies with a standard industrial classification (SIC) code in the finance, insurance or real estate division. Some examples include investment offices (24.2% of the unmatched records) or companies issuing asset-backed securities (12.5% of the unmatched

records). The distribution of the 10K records in our dataset over the different industries (SIC divisions) can be found in Table 4 in the Appendix.

The 10K filings in our dataset come from 18,582 unique companies for which we have 9.16 years of data on average. These companies are relatively large with an average total asset value of 1.39 billion dollars⁶ and are well distributed across the United States as can be seen in Figure 3 in the Appendix. The state where most companies have their headquarters is California, followed by Texas, New York and Florida. The 10K filings in the dataset are relatively long, consisting on average of 29,247 words. The longest items in the 10K filings are item 7: the management discussion and analysis or the MD&A (6,810 words on average), item 1: the business description (6,123 words on average) and item 15: the exhibits (4,799 words on average).

2.4 Data Splits for Bankruptcy Prediction

From the 170,139 records in our dataset, 84,652 qualify for inclusion in the LoPucki BRD (cf. Section 2.2) and were assigned a binary label. Among these 10K records, 662 were filed in the year preceding bankruptcy (i.e. the positives) while 83,990 were not. This implies a strong class imbalance with about 1 positive sample for every 127 negative samples. The labelled 10K records filed prior to 2012 are used to train the models while the records filed between 2012 and 2015 are assigned to the validation set, which is used for hyperparameter optimisation and model selection. The remaining 10K records, filed after 2015, make up the test set and are used to evaluate the final models (which are retrained on all 10K’s in the train and validation set). The train, validation and test sets consists of 54,039; 12,324 and 18,289 filings respectively with 481, 59 and 122 positive cases each. For an overview of the splits, see Table 6 in the Appendix.

3 Experimental Setup

As discussed above, a 10K record consists of various items and contains different data modalities. First, we separately explore the predictive value of (1) the numerical financial data of the 10K’s and (2) the text in the reports, specifically from item 7: the management discussion and analysis. Afterwards, we build a predictive model that uses both data types jointly. In this section we cover the design of the models and briefly discuss the training details.

⁶After removal of outliers exceeding the 95% quantile.

3.1 Numerical Models

The consolidated financial statements are reported in item 8 of a 10K and contain a large number of financial figures. For our prediction models, we employ the most informative accounting figures in line with previous work (Mai et al., 2019).⁷ In Table 5 in the Appendix we give an overview of the variables that serve as an input for our classifiers. As a baseline, we train a logistic regression classifier with an L2-regularisation penalty. Furthermore, we include a multi-layer perceptron and an XGBoost classifier (Chen and Guestrin, 2016) as more advanced alternatives. For the logistic regression model, we only tune the regularisation strength. The dimensions of the hidden layer(s), the learning rate and the regularisation strength are the hyperparameters of the MLP. For the XGBoost model, we optimise the number of trees, the shrinkage factor, the proportion of the data to sample at each split and the maximum depth of the trees.

Due to the infrequent occurrence of bankruptcy, our dataset is heavily imbalanced. As we want our models to be able to discriminate between bankrupt and non-bankrupt firms, we need a strategy to deal with the small number of positive samples (i.e. the 10K records filed in the year preceding bankruptcy). Therefore, we randomly oversample the minority instances in our training data and treat the ratio of positive over negative samples as a hyperparameter as well. Furthermore, we impute missing values (except for the XGBoost model that can handle missing data), centre the variables around the mean and scale them to unit variance. For each model, we set the values of the hyperparameters that maximise the area under the receiver operating curve (ROC-AUC)⁸ (cf. Mai et al. (2019) and Arno et al. (2022)) (see Section 4.1 for more details on this performance metric).

3.2 Textual Models

A Form 10K is an extensive document. On average, a filing in our dataset has 29,247 words. However, much of this content is not relevant for our prediction task (such as the description of the business or the exhibits). The most informative

⁷We discard the market-based predictors (e.g. stock market returns) used by Mai et al. (2019) and only use those features that can be computed from the 10K.

⁸We do not report the results of the models when tuned on average precision (AP) instead of ROC-AUC as there was no substantial difference.

part of the 10K can be found in item 7: the management discussion and analysis. In this section, the management of the company gives its view on the past fiscal year, discussing the risks that the company faced, special circumstances that had an effect on the company and many other interesting aspects that may have had an impact on the results. Consistent with prior literature (Cecchini et al., 2010; Mayew et al., 2015; Mai et al., 2019; Arno et al., 2022), we use the text from this part of the 10K in our prediction models.

First, we train an L1-regularised logistic regression classifier that uses *Term Frequency - Inverse Document Frequency* (TF-IDF) features as input. This keyword-based document representation technique has achieved good performance in information retrieval and document classification tasks (Manning et al., 2008) and serves as our baseline. The regularisation strength and the size of the n-grams are treated as hyperparameters. Second, we finetune a pretrained RoBERTa-large model on our classification task (Zhuang et al., 2021). We only pass the first 512 tokens to the model, which corresponds to its maximum sequence length. In the first epoch, we train only the classification head and freeze the parameters of the encoder. For the second and last epoch, we adjust the learning rate downwards and train the entire model. We use a batch size of 320 instances. In order to handle the class imbalance, we weigh the samples inversely proportional to the class frequencies during training of each textual model.

3.3 Combined Numerical and Textual Model

To leverage the combined predictive power of both the numerical and textual data, we employ an ensemble model. By combining the outputs of the best uni-modal classifiers, we aim to achieve the best overall predictive performance for the bankruptcy prediction task. In our ensemble approach, we retrain the best uni-modal models on the train set and have them score the instances in the validation set. Similarly to stacked generalisation (Wolpert, 1992), the normalised scores are then used to train a meta-classifier that makes the final prediction. Finally, we can use the base classifiers to make uni-modal predictions on the test set, which are used by the meta-classifier to generate a final prediction, taking both data modalities into account.

Data Modality	Numeric			Textual		Combined
Model	LogReg	MLP	XGBoost	TF-IDF	RoBERTa	XGBoost + TF-IDF
ROC-AUC	0.915	0.925	0.936	0.886	0.778	0.948
AP	0.115	0.162	0.156	0.239	0.060	0.264
Recall@100	0.148	0.197	0.189	0.287	0.090	0.287
CAP ratio	0.830	0.851	0.873	0.771	0.554	0.896

Table 1: The results of the numerical, textual and combined models, tuned on ROC-AUC, evaluated on the test set. For each data modality, the best result is shown in bold.

4 Bankruptcy Predictions Result and Analysis

In this section we motivate the choice of performance metrics, report the results of the models on the test set, trained with optimal hyperparameter values, and discuss our most interesting findings.

4.1 Performance Evaluation

We report the area under the receiver operating curve (ROC-AUC), the average precision (AP), the cumulative accuracy profile ratio (CAP ratio) and the recall@100 for each classifier. The **ROC-AUC** summarises the ROC curve, which shows the true positive rate and false positive rate at each classification threshold. The ROC-AUC can be interpreted as the probability that a randomly chosen positive instance (i.e. a 10K record filed in the year before bankruptcy) is scored higher than a randomly chosen negative one by the classifier (Fernández et al., 2018). The **AP** is a metric that summarises the precision-recall (PR) curve and reflects the performance of the model on the minority class. The PR curve graphically depicts the trade-off between precision and recall at each classification threshold. The average precision metric (AP) is particularly valuable when dealing with highly skewed data distributions where the ROC-AUC can be overly optimistic (Davis and Goadrich, 2006). The **recall@100** gives the proportion of positives, retained in the 100 highest ranked instances by the classifier, out of all positives. In our application, it reflects the ability of a model to detect 10K records filed in the year preceding bankruptcy given a fixed budget (i.e., when only 100 filings can be retrieved). Finally, we report the **CAP ratio**, a metric that summarises the cumulative accuracy profile curve (Mai et al., 2019). This curve shows the recall at varying percentages of observations when sorted according to the classifiers’ scores. Furthermore, we also show the PR, ROC and CAP curves for the best numerical, textual and combined models.

4.2 Bankruptcy Classification Performance

From the results in Table 1, we conclude that the MLP and XGBoost classifiers achieve the best performance among the models trained on numerical predictors. The MLP classifier has the best average precision (AP) and recall@100 while the XGBoost model has the highest ROC-AUC and CAP ratio. Furthermore, within the class of models trained solely on text, the keyword-based TF-IDF model attains the best results on all performance metrics. It is worth noting that this is the only model capable of processing the entire documents, unlike RoBERTa, which has a maximum sequence length. As expected, the best results overall are attained by the ensemble model, which leverages both data modalities and combines the predictions from the XGBoost and TF-IDF classifiers.

The PR curve shown in Figure 2a provides additional insights into the performance of the best numerical (XGBoost), textual (TF-IDF) and combined (ensemble) model. We can see that, at low classification thresholds (i.e., when bankruptcy predictions for 10K filings are infrequent), the textual and combined models are comparable while the numerical model lags behind. As the threshold increases and more instances are classified as bankrupt, the precision of the textual models drops quickly and the performance of the numerical model becomes on par with the combined model. The ROC and CAP curves, shown in Figure 2b and Figure 5 in the Appendix, support our previous results and consistently display the best performance for the combined model. The numerical model follows closely while the textual model comes in last.

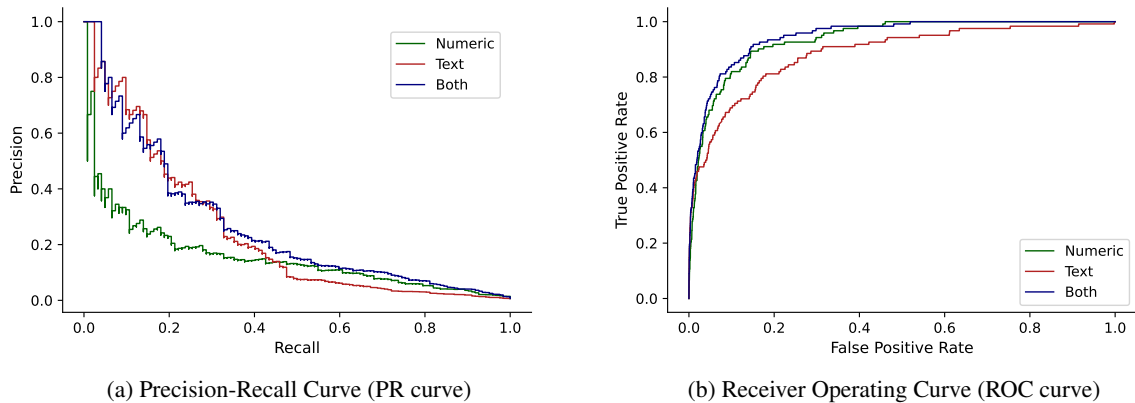


Figure 2: Precision-Recall Curve and Receiver Operating Curve for the best textual (TF-IDF), numerical (XGBoost) and combined (ensemble) models evaluated on the test set.

Snippet from item 7: management discussion and analysis

"...we also may conclude that it is necessary to initiate proceedings under Chapter 11 of the United States Bankruptcy Code..."

"...it may be necessary for us to seek protection from creditors under Chapter 11 of the U.S. Bankruptcy Code..."

"...it may be necessary to seek a private restructuring or protection from creditors under Chapter 11 of the United States Bankruptcy Code..."

Table 2: Snippets from the MD&A of the top ranked instances in our test set by the TF-IDF model.

4.3 Qualitative Analysis of the Results

Complementary Information in the Textual and Numerical Data: A first interesting result is that the TF-IDF model has a better recall@100 but a worse CAP ratio than any classifier trained on numerical data. This suggests that some specific 10K filings are more easily classified when using text from the MD&A as input compared to using the accounting figures. A qualitative inspection of the 10K filings that were ranked highest by the TF-IDF model, revealed that the management of the company sometimes explicitly states that they consider filing for bankruptcy in the coming year. The snippets in table 2 show this behaviour for the 3 highest ranked 10K filings, by the TF-IDF model, in our test set. This information cannot be directly quantified in any of the parameters of the numerical models, and in that respect, we can conclude that the information contained in both data modalities has some complementary value.

As an alternative way of investigating the importance of the textual vs. numerical data, we analyse the combined numerical and textual model introduced in Section 3.3. In particular, we trained a logistic regression classifier with three parameters on the normalised scores from the XGBoost and the TF-IDF classifiers. The weights (β_1 and β_2 in the equation below) represent the relative importance of the information contained in each data modality for the bankruptcy predictions by our ensemble, the best performing model overall.

$$\begin{aligned}
 &P(\text{Next year bankruptcy}) \\
 &= \sigma(\beta_0 + \beta_1 \text{Score}_{XGBoost} + \beta_2 \text{Score}_{TF-IDF})
 \end{aligned}$$

The β_1 and β_2 coefficients are 1.30 and 0.321 respectively, indicating that the numerical data is most informative for the task at hand. This finding supports our previous result, that the textual data is mainly useful for the classification of those few 10K filings where the consideration for bankruptcy is clearly stated. Remaining 10K records are better classified using the accounting figures.

Ranking Performance of the Models: As seen from Table 1, our models attain high values on ROC-AUC and the CAP ratio. This can be directly linked to the imbalanced nature of the problem, since only a very small portion of 10K records are filed in the year before bankruptcy. For the ROC-AUC metric, this means that, for a sufficiently large k , the top k highest ranked instances, by any considered model, would contain a large fraction of all records filed in the year before bankruptcy (leading to a high recall or true positive rate), and a small fraction of all records not filed in the year before bankruptcy (i.e., a low false positive rate or FPR). This is illustrated

by the ROC curves in Figure 2b. In that same set of k highest scoring instances, the number of records filed in the year before bankruptcy is however relatively small compared to the number of records not filed in the year before bankruptcy, leading in turn to a low precision at k , and by extension, a low AP metric. This is illustrated in the precision-recall curve in Figure 2a. When considering the ensemble model and $k = 100$, 35 positives and 65 negatives are retained from the total of 122 positives and 18,167 negatives in the test set (corresponding to a recall@100 of 28.7%, and a FPR of 0.3%).

This result can be further nuanced in light of casting the problem as a binary classification task, which is common in bankruptcy prediction literature. Our hypothesis was, that from these 65 false positives (again considering the top 100 highest ranked 10K records by our best model), some 10K's were filed by companies worthy of further investigation due to their unhealthy financial situation, although they just did not quite file for bankruptcy in the following year yet. Indeed, the 65 false positive 10K records were filed by 53 different companies, of which an additional 23 turned out to have filed for bankruptcy by 2023. Modelling financial health with a more gradual label can therefore be expected to lead to a higher consistency, although the concept of financial health itself is less straightforward to quantify unambiguously. We consider this a highly useful direction for future research.

5 The Potential of LLMs for Text-Based Bankruptcy Prediction

GPT Prompting Strategy Recently, large language models (LLMs) have shown to be tremendously successful on a variety of tasks, including financial text classification (Loukas et al., 2023) and zero-shot text summarisation (Goyal et al., 2022). In this section, we explore how such models can be used in the context of bankruptcy prediction. More specifically, we will use GPT-3.5 Turbo (Ouyang et al., 2022) with a context window of 16,000 tokens to (1) summarise the text from the MD&A section of the 10K filings into a single paragraph and (2) for zero-shot bankruptcy prediction. Due to the associated costs, we do not perform the GPT-based experiments on the entire dataset. Instead, we sample a balanced train set and a random test set from ECL of 1000 instances each.

Using a single prompt (shown in Figure 4 in the Appendix), we ask the model to summarise the MD&A, with a particular focus on the elements that are indicative for the financial health of the company, and to assign a score, ranging from 1 to 10, that indicates how likely it is that the company will file for bankruptcy in the next year. The extracted summaries are then used to re-train the TF-IDF baseline and the RoBERTa model, which is now able to use a compact version of the entire document instead of only the first 512 tokens. We use the same training details as before with two exceptions. Since our sampled training set is balanced, we no longer use a weighted loss function and reduce the batch size to 16. For the zero-shot bankruptcy prediction task, we extract the scores that GPT-3.5 assigned to each document in the test set,⁹ rank the test set accordingly and calculate the performance metrics. Since many instances are assigned the same score, we repeat this process 50 times and randomly shuffle documents with the same score in the ranked test set, to quantify the level of variation in the metrics due to the discrete nature of the GPT-assigned scores. The results of the models trained on the summaries, the original models and GPT-3.5 zero-shot scores on the sampled test of 1000 instances are reported in Table 3.

Summarisation and Zero-Shot Bankruptcy Prediction Performance From the results in Table 3, we can see that the TF-IDF model, trained on the complete text of the MD&A, is still the best textual model overall. GPT-3.5 (zero-shot), the only other model capable of processing entire documents, does significantly worse. An inspection of the top ranked instances by the TF-IDF model, from the sampled test set, showed once again that the good performance of the model can be attributed to its ability to detect 10K filings where the management of the company states that they consider to file for bankruptcy in the next year (cf. Table 2).

The results also show that the summaries extracted by GPT-3.5 are informative for the bankruptcy prediction task. The performance of RoBERTa increased tremendously when trained on these summaries instead of the first 512 tokens of

⁹We were able to extract a score for over 83% of the instances in the test set.

Data Modality	Textual: GPT summaries		Textual: Full MD&A		
	TF-IDF	RoBERTa	TF-IDF	RoBERTa	GPT-3.5 (zero-shot)
ROC-AUC	0.893	0.902	0.912	0.592	0.667 (± 0.022)
AP	0.089	0.202	0.294	0.021	0.019 (± 0.001)
Recall@100	0.600	0.500	0.700	0.200	0.148 (± 0.050)
CAP ratio	0.791	0.804	0.824	0.184	0.335 (± 0.044)

Table 3: The results on the randomly sampled test (of 1000 instances) of the textual models trained on the extracted summaries, the original textual models and GPT-3.5 (zero-shot). Due to the different sizes of this sampled test set and the original test set, the values in this table and Table 1 are not directly comparable.

the MD&A. The performance of the TF-IDF model decreased slightly. This is not surprising, since the summaries contain less information than the complete MD&A and might not capture the sentences where the management states that they consider filing for bankruptcy in the next year. Also, the models trained on the summaries saw only a fraction of the number of training instances compared to the models that saw all of the full-text training instances. Notice how RoBERTa achieves even better performance than the TF-IDF model when both are trained on the summaries, showcasing the strength of the model on a limited context.

In conclusion, the summaries extracted by GPT contained useful information for bankruptcy prediction but the model performed poorly in the zero-shot setting. Additionally, we acknowledge that the quality of the summaries varied and that we encountered some samples where the model suffered from hallucination. In some rare cases, the MD&A is not part of the 10K filing but it is included in another document (such as the annual letter to the shareholders) and item 7 of the 10K contains only a single sentence referencing this document. The GPT-generated summaries in these cases were a paragraph long and contained only imaginary facts. We believe that the performance of LLMs, in terms of summarisation and zero-shot bankruptcy prediction, can be further increased with some additional effort, but that lies outside the scope of this paper.

6 Conclusion

In this paper, we present ECL, a multi-modal dataset of textual and numerical data from corporate 10K filings and associated binary bankruptcy labels. We also present several classical and neural bankruptcy prediction models and provide an in-depth qualitative analysis of the results.

First, our findings highlight the complementarity of the information contained in both data modalities. In the text, the management of the company sometimes explicitly states that they consider filing for bankruptcy in the coming year, making the prediction task trivial for a keyword-based TF-IDF model. If this is not mentioned, the financial numerical features are better predictors for bankruptcy. The best results are attained when we combine the predictions of the textual and numerical models in an ensemble.

Second, we argue that our models achieve acceptable prediction levels that may prove useful in actual applications such as the automated screening of companies’ financial status, although there clearly is room for further research on ECL. We did observe that our models, trained on binary bankruptcy labels, cannot distinguish between 10K records filed in the year preceding bankruptcy and records filed by financially unhealthy companies that are close to bankruptcy but not within one year. This indicates that modelling the financial health of a company using more fine-grained prediction targets, is an interesting avenue for future research as well.

Finally, we study the potential of LLMs in the context of bankruptcy prediction. We observe that the zero-shot bankruptcy prediction results of the GPT-3.5 model are poor. Nonetheless, owing to the large context window of the model, we demonstrate its value by using the LLM to extract meaningful summaries of the text in the 10K’s for the bankruptcy prediction task.

Acknowledgements

This research was made possible through the financial support provided by the Research Foundation Flanders (FWO) under grant number G006421N.

References

- Henri Arno, Klaas Mulier, Joke Baeck, and Thomas De-meester. 2022. [Next-year bankruptcy prediction from textual data: Benchmark and baselines](#). In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 187–195, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- William H Beaver. 1966. Financial ratios as predictors of failure. *Journal of accounting research*, pages 71–111.
- Ben S Bernanke. 1981. Bankruptcy, liquidity, and recession. *The American Economic Review*, 71(2):155–159.
- Mark Cecchini, Haldun Aytug, Gary J Koehler, and Praveen Pathak. 2010. Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, 50(1):164–175.
- 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.
- Jesse Davis and Mark Goadrich. 2006. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine Learning*, pages 233–240.
- Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from imbalanced data sets*, volume 10. Springer.
- Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.
- Myoung-Jong Kim and Dae-Ki Kang. 2010. [Ensemble with neural networks for bankruptcy prediction](#). *Expert Systems with Applications*, 37(4):3373–3379.
- Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. [EDGAR-CORPUS: Billions of tokens make the world go round](#). In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 13–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakasiotis, and Stavros Vassos. 2023. Breaking the bank with chatgpt: Few-shot text classification for finance. *arXiv preprint arXiv:2308.14634*.
- Feng Mai, Shaonan Tian, Chihoon Lee, and Ling Ma. 2019. Deep learning models for bankruptcy prediction using textual disclosures. *European journal of operational research*, 274(2):743–758.
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, USA.
- William J Mayew, Mani Sethuraman, and Mohan Venkatachalam. 2015. Md&a disclosure and the firm’s ability to continue as a going concern. *The Accounting Review*, 90(4):1621–1651.
- Marcus D Odom and Ramesh Sharda. 1990. A neural network model for bankruptcy prediction. In *1990 IJCNN International Joint Conference on neural networks*, pages 163–168. IEEE.
- James A Ohlson. 1980. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, pages 109–131.
- 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.
- David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.
- 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.

Appendix

In the Appendix we show a map with the distribution of the headquarters for the companies in ECL in Figure 3. Table 4 presents the distribution of the industries for the companies in ECL. The prompt given to GPT-3.5 for summarisation and zero-shot bankruptcy prediction is shown in Figure 4. The CAP curve for the best numerical, textual and combined models is shown in Figure 5. Table 5 contains a description of the numerical variables and Table 6 gives an overview of ECL and the dataset splits.

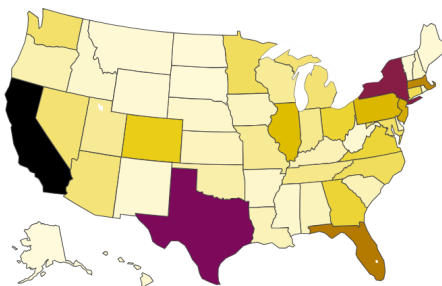


Figure 3: Distribution of the headquarters for the companies in ECL across the United States. A darker colour indicates that more firms are located in this state.

System	You are a financial analyst, specialized in assessing companies' financial health and communicating with clients.
User	I have the management discussion and analysis from a company's 10k report, and would like to know the elements that could indicate its financial health.
Assistant	Show me the 10k report. I will summarize the management discussion and analysis section into a 20-line paragraph with a focus on the company's financial health.
User	Thank you. After the summary, give a conclusion, starting with 'Conclusion: ', where you assign a score from 1 to 10, indicating how likely it is that the company will file for bankruptcy in the next year, with 1 being 'next-year bankruptcy extremely unlikely' and 10 'next-year bankruptcy extremely likely'. The report was filed on + “date” + here it is: + “text”
Assistant	Summary: ...

Figure 4: The prompt given to GPT-3.5 Turbo (with a context window of 16,000 tokens) for (1) summarisation of the MD&A section of the 10K’s and (2) zero-shot bankruptcy prediction.

Table 4: Distribution of the industries (SIC divisions) for the companies in ECL. Most companies are active in the manufacturing industry (shown in bold), followed by finance, insurance and real estate.

SIC Division	Proportion of Data
Agriculture, Forestry, Fishing	0.35%
Mining	4.61%
Construction	1.02%
Manufacturing	37.14%
Transportation & Public Utilities	9.21%
Wholesale Trade	3.04%
Retail Trade	5.09%
Finance, Insurance, Real Estate	20.95%
Services	17.01%
Public Administration	1.58%

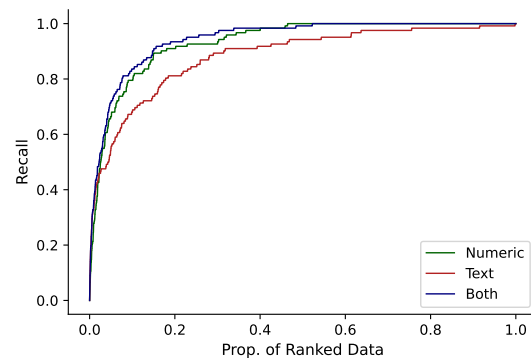


Figure 5: Cumulative Accuracy Profile (CAP Curve) for the best textual (TF-IDF), numerical (XGBoost) and combined (ensemble) models evaluated on the test set. Note that this curve is similar to the ROC curve due to the class imbalance.

Variable	CompuStat	Variable	CompuStat
Current Assets / Current Liabilities	ACT / LCT	Current Liabilities / Sales	LCT / SALE
Accounts Payable / Sales	AP / SALE	Total Liabilities / Total Assets	LT / AT
Cash and Short Term Investments / Total Assets	CHE / AT	Log (Total Assets)	Log (AT)
Cash / Total Assets	CH / AT	Log (Sales)	Log (SALE)
Cash / Current Liabilities	CH / LCT	Net Income / Total Assets	NI / AT
(EBIT + Depreciations and Amortisations) / Total Assets	(EBIT + DP) / AT	Net Income / Sales	NI / SALE
EBIT / Total Assets	EBIT / AT	Operating Income After Depreciations / Total Assets	OIADP / AT
EBIT / Sales	EBIT / SALE	Operating Income After Depreciations / Sales	OIADP / SALE
[Total Debt in Current Liabilities + (0.5)*Total Long Term Debt] / Total Assets	(DLC + 0.5*DLTT) / AT	(Current Assets - Inventory) / Total Current Liabilities	(ACT - INVT) / SALE
Inventory Decrease / Inventory	INVCH / INVT	Retained Earnings / Total Assets	RE / AT
Inventory / Sales	INVT / SALE	Retained Earnings / Current Liabilities	RE / LCT
Current Liabilities - Cash / Total Assets	(LCT - CH) / AT	Sales / Total Assets	SALE / AT
Current Liabilities / Total Assets	LCT / AT	Total Equity / Total Assets	SEQ / AT
Current Liabilities / Total Liabilities	LCT / LT	Working Capital / Total Assets	WCAP / AT

Table 5: This table presents the numerical variables used by our classifiers and the corresponding formulas in CompuStat. We derived the variables from the work of [Mai et al. \(2019\)](#) but only include those that can be computed from the 10K and discard the variables that require market information (e.g. stock market returns).

Dataset	Number of 10K Filings	Period (Filing Year)	Average Asset Value (Billion \$)	Number of Positives	Proportion of Positives	Negatives per Positive
ECL Complete	170,139	1993 - 2023	1.387	-	-	-
ECL Labelled	84,652	1993 - 2021	3.435	662	0.78%	127
Full Training Set	66,363	1993 - 2015	2.851	540	0.81%	122
Training Set	54,039	1993 - 2011	2.518	481	0.89%	112
Validation Set	12,324	2012 - 2015	4.547	59	0.48%	208
Testing Set	18,289	2016 - 2021	5.995	122	0.67%	149

Table 6: This table gives an overview of the ECL dataset and the training, validation and test sets that were used for the next year bankruptcy prediction task. A positive sample refers to a 10K record filed in the year before bankruptcy. The average asset value (in billion \$) is not corrected for inflation and computed after removal of the outliers exceeding the 95% quantile. Note that we do not include statistics on the label distribution for the next year bankruptcy prediction task for the complete ECL dataset. Some samples in this dataset cannot be assigned a label since they do not qualify for inclusion in the LoPucki BRD.

Headline Generation for Stock Price Fluctuation Articles

Shunsuke Nishida Yuki Zenimoto Xiaotian Wang

Takuya Tamura Takehito Utsuro

Degree Programs in Systems and Information Engineering,
Graduate School of Science and Technology, University of Tsukuba
{s2320778, s2220753, s2320811, s2120744}_@_u.tsukuba.ac.jp
utsuro_@_iit.tsukuba.ac.jp

Abstract

The purpose of this paper is to construct a model for the generation of sophisticated headlines pertaining to stock price fluctuation articles, derived from the articles' content. With respect to this headline generation objective, this paper solves three distinct tasks: in addition to the task of generating article headlines, two other tasks of extracting security names, and ascertaining the trajectory of stock prices, whether they are rising or declining. Regarding the headline generation task, we also revise the task as the model utilizes the outcomes of the security name extraction and rise/decline determination tasks, thereby for the purpose of preventing the inclusion of erroneous security names. We employed state-of-the-art pre-trained models from the field of natural language processing, fine-tuning these models for each task to enhance their precision. The dataset utilized for fine-tuning comprises a collection of articles delineating the rise and decline of stock prices. Consequently, we achieved remarkably high accuracy in the dual tasks of security name extraction and stock price rise or decline determination. For the headline generation task, a significant portion of the test data yielded fitting headlines.

1 Introduction

For individuals engaged in stock trading, acquiring up-to-date information regarding stock price fluctuations is highly important. Knowledge of not only whether stock prices have risen or declined, but also the underlying causes such as product launches or sociopolitical conditions, can inform future investment strategies. While news articles on stock trading serve as a primary source of information, manually creating articles for a diverse array of securities¹ is considered challenging due to time constraints.

¹The term "security" is used for expressing the company of the stock.

Hence, it is desirable to construct a system capable of automatically generating stock price fluctuation articles using quantitative stock information and related textual data. Articles typically consist of a body and a headline, with the latter expected to succinctly include, at a minimum, the security name of the fluctuating stock and a term indicating whether the stock price has risen or declined.

This paper assumes that the body of a stock price fluctuation article has already been generated automatically and aims to develop a headline generation model that produces the headline based on the article's content (Figure 1)².

With respect to the headline generation model, which takes the article's content as input, we solve three distinct tasks: in addition to the task of generating article headlines, two other tasks of extracting the relevant security names, and determining stock price rise or decline. Regarding the headline generation task, we also revise the task as the model utilizes the outcomes of the security name extraction and rise/decline determination tasks, thereby for the purpose of preventing the inclusion of erroneous security names.

In each task, we employ pre-trained models that have demonstrated high performance in the field of natural language processing. By fine-tuning³ these pre-trained models for each respective task, we aim to enhance the models' accuracy. The dataset used for fine-tuning consists of the stock price fluctuation article dataset, which is com-

²Although we assume that the headline generation model developed in this paper is to be applied to the body of automatically generated stock price fluctuation articles, in the evaluation of this paper, we report the results of applying the headline generation model to the body of manually written articles. Note that automatically generated articles may have some bias which may not be the case for manually written articles, where future works include studying issues arising from this difference of automatically generated and manually written articles.

³Using the weights of each layer in the pre-trained model as initial values, additional training is conducted with fine-tuning datasets to make subtle adjustments to the weights.

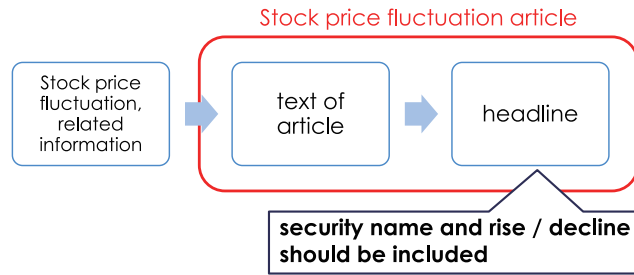


Figure 1: Overall Purpose of the Paper: Headline Generation for Stock Price Fluctuation Articles

posed of articles regarding the rise and decline of stock prices. In this paper, we report the results of the three tasks above as well as the preliminary evaluation results of the revised task of the headline generation task.

Our contributions are as follows:

1. We fine-tuned the pre-trained models XLM-RoBERTa (Conneau et al., 2020) and mT5 (Xue et al., 2021) using the stock price fluctuation article dataset to perform three tasks: generating article headlines, extracting target security names, and determining stock price rise and decline.
2. We developed the dataset for fine-tuning those three models, where the dataset comprises a collection of articles delineating the rise and decline of stock prices.
3. We were able to achieve quite high accuracy in the tasks of security name extraction and determining price rise and decline. For the headline generation task, appropriate headlines were generated for many test data, comparable to the actual article headlines. However, some generated headlines contained incorrect information, such as the target security names.
4. We revised the headline generation task to prevent headlines containing incorrect security names from being generated. Although we were able to reduce the number of headlines generated with incorrect security names while maintaining similar ROUGE (recall) scores as before the revision, the improvement was not significant.

2 Related Work

In the realm of research related to generating headlines from news article content, there exist preced-

ing related work (Rush et al., 2015; Chopra et al., 2016; Nallapati et al., 2016; Kikuchi et al., 2016; Takase and Okazaki, 2019; Hitomi et al., 2019). Among them, Rush et al. (2015) developed the first approach to neural abstractive summarization. After that, Chopra et al. (2016) used the encoder-decoder framework, and Nallapati et al. (2016) incorporated additional features such as parts-of-speech tags and named entities. There also exist attempts to control the output length in neural abstractive summarization (Kikuchi et al., 2016; Takase and Okazaki, 2019; Hitomi et al., 2019). For example, in the works of Hitomi et al. (2019), they propose a corpus for evaluating headline generation models that take output length into consideration. In this paper, on the other hand, we focus on stock-specific terminology characteristic of stock-related news articles, striving to create headlines more suitable for stock price fluctuation articles.

Additionally, in the context of studies on news article headlines and stock prices, there exist several prior investigation. The proposed methods for stock price prediction using news headlines vary across different approaches. In one approach, it is suggested to combine news headlines with technical indicators to predict stock prices (Kalshani et al., 2020). Another approach is also proposed, which predicts the short-term movement of stock prices after financial news events using only the headlines of the news (Chen, 2021). In a third approach, they discussed the failure of the Efficient Market Hypothesis and proposed a project on stock trend prediction using news (Kalyani et al., 2016). Two other approaches evaluate different machine learning and deep learning methods, such as Support Vector Machines (SVM) and Long Short-term Memory (LSTM), to predict stock price movement using financial news (Liu et al., 2018; Gong et al., 2021).

3 Dataset

3.1 Stock Price News of “MINKABU”

In this paper, we utilized the web-based media platform minkabu.jp⁴, which delivers news articles on finance, as the source for collecting news articles on stock price fluctuations to create our dataset. Within minkabu.jp, the distribution source “MINKABU” contains a substantial number of stock price fluctuation news articles. We collected 23,989 news articles⁵ with “MINKABU” as the distribution source. Approximately 290,000 articles⁶ are published with minkabu.jp as the distribution source, and it is estimated that around 81,200 of these articles pertain to stock price rise and decline (Tsutsumi and Utsuro, 2022). Therefore, it can be asserted that the scale is sufficiently ample for collecting stock price fluctuation articles.

3.2 The Procedure of Dataset Development

The method for creating the stock price fluctuation article dataset utilized in this paper is outlined below. In the creation of the stock price rise article dataset, we selected “individual words representing rise in stock prices”, which are vocabulary indicating a stock price increase. We also selected “individual words representing cause of stock price rise”, which are words used when explaining the reasons for a stock price increase. Total number of those words is 76 which include more than ten stock price domain specific words (Tsutsumi and Utsuro, 2022) such as “反発 (correction)”, “続伸 (continued to rise)”, “高値 (high price)”, “カイ気配 (bid price)”, “大幅高 (large rise)”, “上昇 (rise)”, “ストップ高 (hit limit high)”, “急伸 (rise rapidly)”, “連騰 (winning streak)”, “堅調 (increase steadily)”, “急騰 (sharp rise)”, and “好感 (favorable)”. Next, from 3,300 articles⁷ by the distribution source “MINKABU” we extracted 2,734 articles containing either “individual words representing rise in stock prices” or “individual words representing cause of stock price rise” in the article body text. Furthermore, from these 2,734 articles, we excluded articles in which the company code did not appear in the

text⁸ and articles whose headlines had a low probability of being stock price rise news⁹, resulting in the extraction of 1,185 articles with a high likelihood of being stock price rise news. After filtering, we manually created a dataset of 617 stock price rise articles out of the 1,185 articles.

For stock price decline articles, we selected “individual words representing decline in stock prices”, which are vocabulary indicating a stock price decrease, and “individual words representing cause of stock price decline”, which are words used when explaining the reasons for a stock price decrease. Total number of those words is 103 which also include more than ten stock price domain specific words (Tsutsumi and Utsuro, 2022) such as “嫌気 (unfavorable)”, “反落 (reactionary fall)”, “続落 (continued to decline)”, “赤字 (deficit)”, “急落 (fall rapidly)”, “減益 (decrease in profit)”, “出尽くし感 (material exhaustion)”, “転落 (fall)”, “下落 (decline)”, “下振れ (downside)”, “大幅安 (large decline)”, and “引き下げ (reduction)”. Next, from the 23,989 articles by the distribution source “MINKABU”, we extracted 12,887 articles containing either “individual words representing decline in stock prices” or “individual words representing cause of stock price decline” in the article body text¹⁰. Furthermore, from these 12,887 articles, we excluded articles in which the company code did not appear in the text and articles whose headlines had a low probability of being stock price decline news, resulting in the extraction of 7,986 articles with a high likelihood of being stock price decline news. After filtering, we manually created a dataset of 777 stock price rise articles out of the 7,986 articles.

Those 76 stock price rise related words and 103 stock price decline related words have a certain overlap such as “発表 (announcement)”, “影響 (influence)”, “要因 (cause)”, and “見通し (estimation)”. There could be cases where articles may include both stock price rise and decline related words, or may include those overlapping words. Even in such cases, however, most articles usually report only either stock price rise or decline, but do not discuss the trend in change of the stock price

⁴<https://minkabu.jp/>

⁵15,300 articles distributed from June 30, 2020, to December 3, 2020, and 8,689 articles distributed from March 5, 2021, to June 1, 2021.

⁶As of November 2021.

⁷Articles distributed from October 30, 2020, to December 3, 2020.

⁸Articles where “.T>” does not appear in the text.

⁹Articles whose title begins with “<”. This is used in headlines for articles on foreign exchange, bonds, and individual investor trends.

¹⁰For both stock price rise and decline articles, those excluded articles mostly do not report stock price fluctuation.

over longer duration. We also manually exclude those case of exceptional article types.

4 Headline Generation Model

4.1 Overall Procedure

In this paper, with respect to the headline generation model, using the article’s main text as input, we solve three distinct tasks: in addition to the task of generating article headlines, two other tasks of extracting the relevant security names, and determining stock price rise or decline.

For the stock price rise and decline judgment task and security name extraction task, we used the XLM-RoBERTa (base-sized model) (Conneau et al., 2020)¹¹ as the pre-trained model¹². XLM-RoBERTa is a multilingual model pre-trained on Common-Crawl data¹³ containing 100 languages¹⁴.

For the headline generation task, we used the mT5 (small-sized model) (Xue et al., 2021)¹⁵ as the pre-trained model¹⁶. mT5 is a multilingual model pre-trained on the mC4 corpus¹⁷, which includes 101 languages¹⁸.

For fine-tuning each pre-trained model, we used the stock price fluctuation article dataset mentioned in section 3. As a preprocessing step for this dataset, all security name codes¹⁹ appearing in the dataset’s context were removed to prevent easy identification of security name positions. After ensuring that the training, validation and test data did not contain articles about the same security name, we randomly divided the whole dataset into three parts of 200 validation examples, 200 test examples and the remaining 994 examples for training. Those divided datasets were used for all tasks (Table 1).

4.2 Rise and Decline Detection

For fine-tuning the pre-trained model for the stock price rise and decline judgment task, we used the

¹¹<https://huggingface.co/xlm-roberta-base>

¹²XLM-RoBERTa is designed to be easy to apply to classification tasks and sequence labeling tasks.

¹³<https://commoncrawl.org>

¹⁴Including the Japanese language.

¹⁵<https://huggingface.co/google/mt5-small>

¹⁶mT5 is designed to be easy to apply to text to text conversion tasks including summarization and generation tasks.

¹⁷<https://www.tensorflow.org/datasets/catalog/c4>

¹⁸Including the Japanese language.

¹⁹Assigned to all listed companies, composed of “unique name code with 4 digits + 1 reserve code digit.”

Huggingface Transformers Text classification library²⁰. For model fine-tuning²¹, we used 994 training examples from the stock price fluctuation article dataset.

After fine-tuning, we inputted the 200 test examples from the stock price fluctuation article dataset to the model and tested its performance. We used accuracy as the performance evaluation metric²². The accuracy was 1.0, indicating that the model correctly determined the rise and decline for all 200 test examples.

4.3 Extraction of Security Names

For fine-tuning the pre-trained model for the security name extraction task, we used the Huggingface Transformers Question Answering library²³. For model fine-tuning²⁴, we used 994 training examples from the stock price fluctuation article dataset. We set the question text for each example to be blank and the answer to be the security name covered in the article, and inputted the pair of answers and article body text to the model.

After fine-tuning, we inputted the 200 test examples from the stock price fluctuation article dataset to the model and tested its performance. We used exact match rate as the performance evaluation metric²⁵. The exact match rate was 99.5%, indicating that the model could correctly extract security names for almost all of the 200 test examples.

4.4 Headline Generation

For fine-tuning the pre-trained model for the headline generation task, we used the Huggingface Transformers Summarization library²⁶. For fine-tuning²⁷, we used 994 training examples from the stock price fluctuation article dataset. We treated

²⁰<https://github.com/huggingface/transformers/tree/v4.21.1/examples/pytorch/text-classification>

²¹The learning rate was set to 0.00002, the batch size was set to 8, and the number of epochs is 10.

²²Accuracy is given by (number of correctly classified data)/(number of data).

²³<https://github.com/huggingface/transformers/tree/v4.21.1/examples/pytorch/question-answering>

²⁴The learning rate was set to 0.00003, the batch size was set to 8, and the number of epochs is 2.

²⁵The proportion of cases where the model’s output string exactly matched the reference security name string.

²⁶<https://github.com/huggingface/transformers/tree/v4.21.1/examples/pytorch/summarization>

²⁷The learning rate was set to 0.00003, and the batch size was set to 8. The model for the minimum validation loss is selected.

	stock price rising articles	stock price declining articles	Total
training data	432	562	994
validation data	94	106	200
test data	91	109	200
total	617	777	1,394

Table 1: Number of Articles of Each Type

the article headlines as summaries for each data and inputted the pair of summary and article body text to the model.

After fine-tuning, we inputted the 200 test examples from the stock price fluctuation article dataset to the model and tested its performance. We used ROUGE (recall) as the performance evaluation metric. ROUGE (recall) measures the degree of match between the model-generated summary and the reference summary, where it is measured as the rate of the intersection of the model-generated summary and the reference summary over the reference summary. ROUGE-1 (recall) measures the match at the 1-gram (word) level, ROUGE-2 (recall) at the 2-gram (bi-gram) level, and ROUGE-L (recall) measures the match of the longest common sequence. As shown in the row of “before task revision” in Table 4, ROUGE-1 (recall) was 53.03, ROUGE-2 (recall) was 35.73, and ROUGE-L (recall) was 52.77²⁸.

Next, we conducted a manual evaluation of 100 out of the 200 headline generation results for the test data. The column of “before task revision” in Table 2 shows examples of model-generated headlines that were manually evaluated. We examined whether the information in the model-generated headlines corresponded to the information in the input context and found that 80 out of the 100 test examples contained relevant information. Of these 80 examples, 38 (47.5%) had a perfect string match, and 65 (81.3%) were appropriate as headlines.

Additionally, while all the 100 test examples contained correct information about stock price rise or decline, there were six cases where the security name was incorrect.

²⁸Without fine-tuning, the accuracy of rise and decline detection was 0, and the exact match rate of extraction of security names was 45.5%, which was pretty low compared with after fine-tuning. ROUGE (recall) of headline generation was also pretty low, where ROUGE-1 (recall) was 2.56, ROUGE-2 (recall) was 0.29, and ROUGE-L (recall) was 2.60.

5 Revised Task of Headline Generation and its Preliminary Evaluation Results

In the headline generation model described in the previous section, high accuracy was achieved for the tasks of determining whether the stock price rose or declined and extracting the security name. However, for the headline generation task, the model generated titles containing incorrect information not found in the input context for 20 out of 100 test examples. Here, we focus on the fact that six of these examples had incorrect security names and aim to improve the headline generation task by inserting the security name and a tag (“r” or “d”) representing whether the stock price rose or declined in the input context^{29,30} to prevent headlines with incorrect security names from being generated.

For fine-tuning the pre-trained model for the revised headline generation task, we used the Huggingface Transformers Summarization library, as in the case of the headline generation task of section 4.4. The training data of the stock price fluctuation article dataset and the various parameters during training were also set to the same values as the previous headline generation task. We inputted the pair of the article headline and the body text with the security name and rise/decline tags added to the model.

After fine-tuning, we inputted the 200 test examples from the stock price fluctuation article dataset to the model and performed the same performance test as in the previous headline generation task. As shown in the row of “after task revision” in Table 4, ROUGE-1 (recall) was 57.22%,

²⁹For example, “company-name-A is surging . . .” → “r company-name-A company-name-A is surging . . .”

³⁰Both in the training and in the test, we insert the reference tag representing the correct information on whether rise or decline, as well as the reference security name. This is simply because the test performance presented in section 4.2 and section 4.3 is almost perfect. We also confirmed that, when we inserted a randomly chosen security name in this revised task, the generated headline was with an incorrect security name for 13% cases.

content of the article (partially omitted)	reference headline	model-generated headline
Sグループは続伸している。25日の取引終了後、株主優待制度の内容を変更すると発表しており、これが好感されているようだ。創業30周年記念優待制度の内容を継続し、21年12月末以降も100株以上保有者を対象に、保有株数と継続保有期間に応じて1000円から1万円分のオリジナルクオカードを贈呈するという。(= S Group continues to rise. After the close of trading on the 25th, they announced a change in their shareholder benefits program, which seems to be well-received. They will continue the 30th anniversary commemorative benefits program, and after December 2021, they will give original QUO cards worth 1,000 to 10,000 yen to shareholders who hold 100 or more shares, depending on the number of shares held and the continuous holding period.)	SGは続伸、創業30周年記念優待制度の内容を継続へ(= SG Continues to Rise, Commemorative Benefits Program)	SGは続伸、株主優待制度の変更を好感(= SG Continues to Rise, Favorably Received Changes to Shareholder Benefits Program) (Information corresponding to the content of the article is included in the model-generated headline, which is comparable to the reference headline.)
M社が大幅続伸している。きょう付の日本経済新聞朝刊で、「1千キロメートル離れた場所から複数のドローンをまとめて操作できるシステムを2021年度にも実用化する」と報じられており、これが好材料視されているようだ。... (= M Company has made significant gains. According to today's morning edition of the Nikkei Shimbun, it was reported that they "will commercialize a system that can control multiple drones from a distance of 1,000 kilometers by the end of fiscal 2021," which seems to be viewed as good news. ...)	M社が大幅続伸、1000キロ先のドローン操作する技術を21年度にも実用化と報じられる(= M Company Surges, Reported to Commercialize 1000km Distant Drone Control Technology by FY 2021)	M社が大幅続伸、1千キロメートル離れた場所から複数のドローンをまとめて操作(= M Company Surges, Control Multiple Drones from a Distance of 1,000 Kilometers) (Information corresponding to the content of the article is included in the model-generated headline, while the model-generated headline is not sufficient)
K社が一時ストップ高まで買われた。同社はきょう、画像処理検査エンジンの販売強化などを目的に12月1日から社長直轄のプロジェクトチームでの活動を開始すると発表しており、今後の展開などが期待されているようだ。同社はこれまでに培った技術やノウハウを生かし、さまざまな顧客ニーズに応えられる画像処理外観検査用のエンジンを開発し、従来の液晶向け以外の分野にも進出・拡販することで、収益の拡大と安定によって、収益性の高い事業体制を確立している。(= K Company's stock temporarily hit the daily limit high. The company announced today that it will start a project team under the direct control of the president from December 1st, aiming to strengthen the sales of image processing inspection engines, among other goals. It seems that future developments are expected. The company plans to establish a highly profitable business structure by expanding and stabilizing revenues, by developing image processing exterior inspection engines that can respond to various customer needs using the technology and know-how they have cultivated so far, and expanding into fields other than traditional liquid crystal displays.)	K社は一時S高、画像処理検査エンジンの販売強化に向けた活動開始へ(= K Company Temporarily Hits S High, Activities to Strengthen Sales of Image Processing Inspection Engines Begin)	K社が一時S高、画像処理検査エンジンの販売強化などを目的に事業展開(= K Company Temporarily Hits S High, Business Expansion Aimed at Strengthening Sales of Image Processing Inspection Engines) (Information corresponding to the content of the article is included in the model-generated headline, while the model-generated headline is not sufficient. Especially, the financial terminology "daily limit high" is abbreviated as "S high".)
J社が急反落。23日付の日本経済新聞朝刊で、「世界の飛行機の3分の1、8600機は休暇シーズンの8月に入っても地上に留め置かれている。需要拡大を前提に投資してきたリース産業や金融商品もリスクにさらされ、新造機はキャンセルが相次ぐ」と報じられており、航空機リースを主力とする同社への警戒感が強まっているようだ。... (= J Company's stock is plummeting. According to the morning edition of the Nihon Keizai Shimbun on the 23rd, "One-third of the world's airplanes, or 8,600 aircraft, are still grounded even in August, the vacation season. The leasing industry and financial products that have invested based on demand expansion are exposed to risks, and new aircraft are being canceled one after another." It seems that vigilance is increasing towards the company, which mainly deals with aircraft leasing. ...)	J社が急反落、眠る飛行機8600機と報じられ業績への影響を警戒(= J Company's Stock Plummet, 8,600 Grounded Aircraft Reported, Increasing Caution on Earnings Impact)	J社が急反落、国内有力証券が急反落、国内有力機はキャンセルと報じられる(= J Company's Stock Plummet, Domestic Leading Securities Plunge, Domestic Leading Aircraft Reported to be Canceled) (Information irrelevant to the content of the article is included in the model-generated headline, such as "domestic leading securities".)

Table 2: Examples of Stock Price Fluctuation Article Content, Headlines, and Model-generated Headlines in the Test Data (Japanese text with English translation)

content of the article (partially omitted)	reference headline	model-generated headline	
		before task revision	after task revision
Uホールディングスが続急伸している。12日の取引終了後、第2四半期累計(49月)連結決算を発表しており、売上高20億9300万円(前年同期比8.8%増)、営業利益3億9300万円(同22.0%増)、純利益2億5400万円(同7.7%増)となり、営業利益、純利益とも第2四半期における過去最高益更新したことが好感されている。... (= U Holdings continues to surge rapidly. After the close of trading on the 12th, the company announced its second quarter consolidated financial results (from April to September), with net sales of 2.093 billion yen (an increase of 8.8% year-on-year), operating profit of 393 million yen (an increase of 22.0% year-on-year), and net profit of 254 million yen (an increase of 7.7% year-on-year). The record-breaking operating profit and net profit in the second quarter are well-received...)	Uが続急伸、高収益サブスクモデルの早期確立で第2四半期は過去最高益更新(= U continues to surge, setting record-high earnings in the second quarter with early establishment of high-profit subscription model)	BBHDが続急伸、第2四半期営業利益22%増(= BBHD continues to surge, 22% increase in second quarter operating profit)	Uが続急伸、第2四半期営業利益は過去最高益更新(= U continues to surge, updating record-high operating profit in the second quarter) (The security name was incorrect before task modification, while it is generated correctly after task modification.)
T社が6連騰と上げ足を強め、25日移動平均線を大きく上に放れてきた。同社は樹脂封鎖装置を主力とする半導体製造装置メーカーで、コンプレッション型を中心に収益を伸ばしている。9日取引終了後に21年3月期業績予想の修正を発表、トップラインは計画ラインを減額したものの、営業利益は従来予想の20億円から23億1000万円(前期比2.8倍)に大幅増額しており、これを手掛かり材料に上値を見込んだ買いが継続している。... (= T Corp has strengthened its upward momentum with a six-day winning streak, breaking significantly above its 25-day moving average. The company, a semiconductor manufacturing equipment manufacturer specializing in resin sealing devices, has been expanding its revenues mainly through compression-type devices. After the close of trading on the 9th, the company announced a revision to its earnings forecast for the fiscal year ending March 2021. Although the top line was reduced from the initial plan, operating profit was significantly increased from the previous forecast of 2 billion yen to 2.31 billion yen (2.8 times the previous fiscal year), and this has served as a catalyst for continued buying with an upward outlook...)	T社が6連騰と上げ足加速、21年3月期営業大幅増額で前期比2.8倍に(= T Corp's 6-day winning streak accelerates, with operating profit for the fiscal year ending March 2021 significantly increased to 2.8 times the previous fiscal year)	T社が6連騰、半導体製造装置の受注拡大で21年3月期営業利益予想を大幅増額(= T Corp on a 6-day winning streak, significantly raising its operating profit forecast for the fiscal year ending March 2021 due to expanded orders for semiconductor manufacturing equipment)	T社 T社が25日線を大きく上放れ、半導体製造装置の受注拡大で21年3月期営業利益予想を大幅増額(= T Corp T Corp surges significantly above the 25-day line, significantly raising its operating profit forecast for the fiscal year ending March 2021 due to expanded orders for semiconductor manufacturing equipment) (Before task modification, the security name was correctly identified, while after inserting the tag through revision, extra security name words "T Corp" were incorrectly generated.)

Table 3: Examples of Stock Price Fluctuation Article Content, Headlines, and Model-generated Headlines in the Test Data (before and after task revision, Japanese text with English translation)

ROUGE-2 (recall) was 36.94%, and ROUGE-L (recall) was 52.27%. Table 4 also compares the ROUGE values before and after task revision, where, although the ROUGE-1 (recall) value improved slightly after the revision, the ROUGE-2 (recall) and ROUGE-L (recall) values hardly changed.

Next, we manually evaluated the 100 headlines generated for the test data, which were previously evaluated in the headline generation task of section 4.4. Table 3 shows examples of differences in the model-generated headlines before and after the task revision based on the manual evaluation.

Upon examining whether the information in the

generated headlines corresponded to the information in the input context, we found that 82 out of 100 examples contained relevant information. Out of these 82 examples, 30 (36.6%) had exact matches in terms of character strings, and 62 (75.6%) were considered appropriate headlines. In addition, while all the 100 examples contained correct information about stock price rise and decline, 5 cases had errors in the security names. Table 5 shows the results of the manual evaluation before and after the revision.

From the above, it can be seen that, by the task revision, the number of headlines generated with incorrect security names is reduced, while main-

	ROUGE-1	ROUGE-2	ROUGE-L
before task revision	53.03	35.73	52.77
after task revision	57.22	36.94	52.27

Table 4: ROUGE (recall) Scores Before and After Task Revision

evaluation point		before task revision	after task revision
(i)	Relevant information exists in the model-generated headline.	80	82
(ii)	(Out of (i)) The model-generated headline is appropriate as a headline.	65	62
(iii)	(Out of (ii)) The model-generated and reference headlines matched exactly.	38	30
(iv)	The model-generated headline contains correct information on stock price rise/decline.	100	100
(v)	The model-generated security name is correct (including appropriate abbreviations).	94	95

Table 5: Results of Manual Evaluation for 100 Test Examples (before and after task revision)

taining the same level of accuracy as before. However, it can be said that the improvement was not significant.

6 Conclusion

In this paper, we fine-tuned the pre-trained models XLM-RoBERTa and mT5 using the stock price fluctuation article dataset to perform three tasks: generating article headlines, extracting target security names, and determining stock price rise and decline. We constructed a stock price fluctuation article headline generation model.

We were able to achieve quite high accuracy in the tasks of security name extraction and determining stock price rise and decline. For the headline generation task, appropriate headlines were generated for many test data, comparable to the actual article headlines. However, some generated headlines contained incorrect information, such as the target security names. We revised the headline generation task to prevent headlines containing incorrect security names from being generated. Although we were able to reduce the number of headlines generated with incorrect security names while maintaining similar ROUGE (recall) scores as before the revision, the improvement was not significant.

As future work, we would like to investigate methods such as setting loss functions for each of the tasks of security name extraction, stock

price rise and decline determination, and headline generation, and performing multi-task learning to ensure that the security name information in the headlines is correct. This is to clarify whether the multi-task learning approach can improve the performance of the method employed in this paper, which is quite specific to the financial domain compared with the general language models studied in Radford et al. (2019). Another future work is to apply Large Language Models (LLMs) such as ChatGPT (OpenAI, 2023) and LLaMA (Touvron et al., 2023) to the task studied in this paper.

References

- Qinkai Chen. 2021. [Stock movement prediction with financial news using contextualized embedding from BERT](http://arxiv.org/abs/2107.08721). <http://arxiv.org/abs/2107.08721>.
- Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. [Abstractive sentence summarization with attentive recurrent neural networks](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.
- 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 Asso-*

- ciation for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Jiaying Gong, Bradley Paye, Gregory Kadlec, and Hoda Eldardiry. 2021. Predicting stock price movement using financial news sentiment. In *Proceedings of the 22nd Engineering Applications of Neural Networks Conference*, pages 503–517, Cham. Springer International Publishing.
- Yuta Hitomi, Yuya Taguchi, Hideaki Tamori, Ko Kikuta, Jiro Nishitoba, Naoaki Okazaki, Kentaro Inui, and Manabu Okumura. 2019. [A large-scale multi-length headline corpus for analyzing length-constrained headline generation model evaluation](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 333–343, Tokyo, Japan. Association for Computational Linguistics.
- Ali Hassanzadeh Kalshani, Ahmad Razavi, and Reza Asadi. 2020. [Stock market prediction using daily news headlines](#). <https://ssrn.com/abstract=3685530>.
- Joshi Kalyani, H. N. Bharathi, and Rao Jyothi. 2016. [Stock trend prediction using news sentiment analysis](#). <http://arxiv.org/abs/1607.01958>.
- Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. [Controlling output length in neural encoder-decoders](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.
- Yang Liu, Qingguo Zeng, Huanrui Yang, and Adrian Carrio. 2018. Stock price movement prediction from financial news with deep learning and knowledge graph embedding. In *Knowledge Management and Acquisition for Intelligent Systems*, pages 102–113, Cham. Springer International Publishing.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gu?lçehre, and Bing Xiang. 2016. [Abstractive text summarization using sequence-to-sequence RNNs and beyond](#). In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- OpenAI. 2023. [GPT-4 technical report](#). <http://arxiv.org/abs/2303.08774>.
- Alec Radford, Jeffrey 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.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. [A neural attention model for abstractive sentence summarization](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.
- Sho Takase and Naoaki Okazaki. 2019. [Positional encoding to control output sequence length](#). 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 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothee Lacroix, Baptiste Roziere, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [LLaMA: Open and efficient foundation language models](#). <http://arxiv.org/abs/2302.13971>.
- Gakuto Tsutsumi and Takehito Utsuro. 2022. [Detecting causes of stock price rise and decline by machine reading comprehension with BERT](#). In *Proceedings of the 4th Financial Narrative Processing Workshop @ LREC2022*, pages 27–35, Marseille, France. European Language Resources Association.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Audit Report Coverage Assessment using Sentence Classification

Sushodhan Vaishampayan, Nitin Ramrakhiani, Sachin Pawar, Aditi Pawde*
Manoj Apte, Girish Keshav Palshikar

TCS Research, Tata Consultancy Services Limited, India.

{sushodhan.sv, nitin.ramrakhiani, sachin7.p, manoj.apte, gk.palshikar}@tcs.com
aditi.pawde@walchandsangli.ac.in

Abstract

Audit reports are a window to the financial health of a company and hence gauging coverage of various audit aspects in them is important. In this paper, we aim at determining an audit report’s coverage through classification of its sentences into multiple domain specific classes. In a weakly supervised setting, we employ a rule-based approach to automatically create training data for a BERT-based multi-label classifier. We then devise an ensemble to combine both the rule based and classifier approaches. Further, we employ two novel ways to improve the ensemble’s generalization: (i) through an active learning based approach and, (ii) through a LLM based review. We demonstrate that our proposed approaches outperform several baselines. We show utility of the proposed approaches to measure audit coverage on a large dataset of 2.8K audit reports.

1 Introduction

Financial audit is a complex process used by organizations to assure the stakeholders about the quality and trustworthiness of the governance (Whittington and Pany, 2021), (Arens and Loebbecke, 1999). Auditors examine data, documents, systems and processes, physical assets to ensure that they comply with the required standards, guidelines, laws and regulations and also to ensure that the reported financial information is fair and accurate. Outside the organization, stakeholders use audited financial statements (FS) - such as balance sheet, income statement, cash-flow statement etc.- for making important decisions such as investments, loans, taxation and so forth. One important outcome of an audit is the *audit report* prepared by the auditors, wherein the auditor declares the FS are free from material misstatement, are fair and accurate and are presented in accordance with the relevant accounting standards. If not, the auditor

identifies several types of issues, makes suggestions for improvement, and identifies instances of non-conformance, misinformation, irregularities, inconsistencies, errors, inaccuracies, frauds, lapses, non-compliance, violations etc.

Given the crucial importance of audits, and the high demands on the knowledge, experience and efforts of the auditing team, it is important to measure the *quality* of an audit in order to ensure that it was carried out efficiently and effectively. Poor quality audits, whether intentional or not, can have disastrous consequences, such as frauds, loss of earnings, loss of goodwill, litigations, inability of the company to function as a going concern and even bankruptcy; e.g., see (Lennox and Li, 2019). There are many reasons why an audit can be of poor quality: lack of expertise in the auditing team (Reichelt and Wang, 2010), compromised auditor independence (Tepalagul and Lin, 2015), biases, conservatism (*recognize bad news rather than good news*) (Basu, 1997) and risk-averse attitudes of auditors, non-cooperation from management, insufficient time/efforts spent in auditing etc. The Sarbanes-Oxley Act 2002 in the US is explicitly aimed at improving auditing and public information disclosure, in the light of persisting scandals fueled by auditing failures such as Enron (Beasley et al., 1999) and Satyam (Bhasin, 2013).

A good comprehensive audit report is an important indicator of a good audit. Audit monitoring bodies such as The Chartered Accountants (CA) Society of India have issued guidelines on the contents of audit reports wherein they describe a set of audit aspects which the auditor should touch upon and describe. In this paper, we focus on measuring the coverage of the audit report based on such statutory requirements, as one of the initial steps to gauge audit quality. We pose the problem of gauging coverage of the audit aspects in an audit report through classification of sentences in the report into one or more of these aspects. Given the

*Work done while working at TCS Research

Class	Description	Example Sentence
<i>approval of managerial remuneration</i>	Compliance as per applicable act and payment for managerial remuneration.	We draw attention to Note 42 to the financial statements relating to managerial remuneration paid which is in excess of the limits approved by the Central Government to the extent of Rs. 214.45 lakhs ...
<i>fraud reporting</i>	Fraud by the company or officers or employees, if any, is mentioned and whether any whistleblower complaints were received	According to the information and explanations given to us, a fraud on or by the company has not been noticed or reported during the year.
<i>nidhi company</i>	Remarks on type of company: nidhi, chit fund, etc.	In our opinion, the nature of activities of the Company does not attract any special statute applicable to chit fund and nidhi / mutual benefit funds / societies.
<i>non-cash transactions</i>	Remarks on compliance applicable for non-cash transactions with directors and related persons	Cash flows are reported using the indirect method, whereby profit before tax is adjusted for the effects of transactions of non - cash nature ...
<i>private placement or preferential issues</i>	Remarks on whether company has made preferential allotment or private placement of shares	The Company had invested Rs. 1000 million in 8.75% Cumulative Preference Shares of M/S. ITI Limited during the year 2001 - 02.
<i>utilization of ipo and other public offers</i>	Remarks on money raised through IPOs or other public offers	The Company has not granted any loans and advances on the basis of security by way of pledge of shares, debentures and other securities.
Complex Classes		
<i>cost records</i>	A remark about maintenance of cost records.	However, we have not made a detailed examination of the cost records with a view to determine whether they are accurate or complete.
<i>fixed assets</i>	Remarks on purchase of fixed assets, holding of benami property, physical verification of property, plant and equipment by the management at reasonable intervals.	The company has maintained proper records showing full particulars, including quantitative details and situation of fixed assets.
<i>human resources, payroll processing</i>	Remarks on employee wages, leaves, bonus, pension, full and final settlement and mentions of policies for leave, gratuity and pension.	Also Defined benefits obligations in nature of Gratuity and Leave encashment are to be accounted on accrual basis.
<i>internal control system</i>	Remarks on evaluation of internal control procedures with respect to the size and the nature of the company.	During the course of our audit, no major weakness has been noticed in the internal control system in respect of these areas.
<i>inventory</i>	Remarks on possession and purchase of inventory, its physical verification at timely intervals and record keeping	On the basis of the records of inventory, we are of the opinion that the Company is maintaining proper records of inventory and no material discrepancies were noticed on physical verification.
<i>investments</i>	Remarks on investments by the company and compliance to respective Acts	The company has a strategic long term investments in Equity Shares of certain companies, the cost of acquisition of those investments is Rs. 722.50 lacs.
<i>litigations</i>	Remarks about ongoing litigations on the company	Contempt Petition filed against Excise Department at Allahabad High Court against our refund of Rs. 17,25,392/- against the order of Supreme Court in our favor.
<i>material uncertainty</i>	Remarks on material uncertainties for the company such as net worth, accumulated losses and going concern	The Company 's accumulated losses at the end of the financial year are less than fifty per cent of its net worth.
<i>operational and administrative expenses</i>	Remarks on company's operational expenses	The Company has Capitalized expenses to the tune of Rs. 25.40 Crores in Pulp Mill Unit till the date of last balance sheet...
<i>payables</i>	Remarks on details of amount/money to be paid by the company such as repayment of loans	The repayment of loan is on demand, there is no overdue amount remain outstanding.
<i>purchase and procurement</i>	Remarks on purchases and procurement of any kind	The activities of the Company do not involve purchase of inventory and the sale of goods.
<i>receivables</i>	Remarks on details of amount/money to be received by the company such as loans given	The net amount recoverable of Rs. 23640.05 million is subject to reconciliation and confirmation.
<i>sales, services and revenue</i>	Remarks on sales, services and revenue	The Company is a service company, primarily rendering software services.
<i>statutory dues</i>	Remarks on payment of statutory dues and related disputes	The Company is regular in depositing with appropriate authorities undisputed statutory dues including provident fund, employees ' state insurance ...
<i>working capital</i>	Remarks on working capital and cash/bank balance	No long terms funds have been used to finance short - term except permanent working capital.

Table 1: List of classes in the annotated audit reports with their description and examples

large number of these aspects and domain expertise required to create labelled training data, the text classification problem becomes highly challenging. In this regard, we propose a weakly supervised text classification algorithm based on regular expression based patterns and a multi-label BERT based classifier. To supplement the approach for increasing its recall, we explore two directions - (i) using active learning requiring manual labelling effort and, (ii) using support from LLMs, requiring effort on prompt creation. We present our experimentation and analysis on a dataset of audit reports of companies based in India discussing their audits for the year 2014. To demonstrate the impact of the learning from this work, we present a brief statistical analysis on the dataset.

2 Problem Definition

An audit report consists of various sections mentioning details about a company being audited, responsibility of management and auditor followed by remarks or comments by the auditors pertaining to company’s business operations. Generally, auditors adhere to standard *audit checklist* that includes scope of the audit, evidence collection, audit tests, result analysis and conclusions to be drawn from audit. Moreover, auditors also have to comply with any legislation by local regulatory bodies. Since, the goal in this paper is to determine the *audit coverage* and data under consideration is of Indian companies, the coverage is checked with respect to a standard auditing checklist (ICAI, 2017) and Companies (Auditor’s Report) Order, 2020 (CARO) (ICAI, 2020). Accordingly, the union of classes from both these sources is considered as given in table 1. A sentence in audit report can belong to 0, 1 or more labels from this list. Thus, this is a multi-class multi-label classification problem with number of class labels $m = 21$. Sentences that do not belong to any class label, are considered to be *Not applicable* or *NA*. In Table 1, we list the classes with a brief description and an example sentence from an audit report for each class.

3 Proposed Approach

We propose a *sieved* approach which combines the power of multiple techniques such as Rules, a Standard BERT based classifier, Active Learning and Large Language Models. We explain the contribution of each of the techniques individually and then how they are combined in an ensemble for the final

prediction on test data.

3.1 Rules - Regular Expression based Patterns

As can be observed in Table 1, sentences belonging to certain classes are clearly amenable for rule based labelling. For e.g., sentences in classes such as *Nidhi Company* and *non-cash transactions* typically mention the class names in exact and very rarely in a different format. This exactness is by virtue of how auditors are trained to mention their findings about these aspects/classes. Hence, this facet prompts us to use rules in the form of regular expression patterns for a precise identification of these specific classes.

We devise regular expression based patterns which are constructed by tokens indicative of the respective class. In Table 2, we show some of the regular expression patterns for a subset of classes. Consider for example, the regular expressions for the class *Fixed Assets*. As can be seen tokens such as *intangible* or *immovable* followed by tokens such as *assets* or *properties* would be indicative of the *Fixed Assets* class. For certain classes, the rule may be built of more than one component patterns and all pattern components must match in the sentence, though in any order, for the class to get predicted. An example is seen for the class *Litigations*, where the first component searches for words such as *cases* or *appeals* and the second component searches for words such as *courts* and *tribunals*. The regular expressions also involve negative look-aheads such as the second pattern for the class *Material Uncertainty* in Table 2. It ensures that a phrase such as *no uncertainty* or *not uncertain when* is observed, labelling to the class *material uncertainty* is avoided. We also develop patterns for indicating sentences which are template sentences that auditors include as part of the report and should be marked with a NA label. The rule based classifier labels them with the *NAconfirm* label which is treated as *NA* during evaluation.

As the classification problem is multi-label in nature and the rules may predict multiple labels for a sentence leading to no conflicts. This makes it little different from Snorkel (Ratner et al., 2017) like data programming paradigms.

3.2 Multi-label Sentence Classifier

We also observe that there are sentences wherein the belongingness to the corresponding class is not lexically closed and hence classification using only

Class	Regular Expression Pattern
fixed assets	<code>\b((fixed intangible immovable)(assets)? propert(y ies)))\b</code>
litigations	<code>\b(litigations? cases? arbitrations? appeals? matters? disputes?)\b AND \b(courts? tribunals? judges? nclt)\b</code>
material uncertainty	<code>\b(financial debts?)re\W?structur(e[d]? ling)\b \bre\W?structur(e[d]? ling)\b.*\b(debts?)\b</code>
material uncertainty	<code>(^(?!\\bnot?\\b).*?)\\buncertain(y ies)\b \\b(nolnot)(\\w+)? (certaint(y ies) ascertain(ed ing)? ascertainable))\b(statements?)\b</code>
statutory dues	<code>\b(tax(es)? provident funds? (customs? excise)duty duty of (customs? excise))\b AND \\bdues?\\b</code>

Table 2: Example Regular Expression Patterns

lexical patterns may not be sufficient. For e.g., sentences in classes such as *payables* and *fixed assets* mention about the payables and assets in various ways apart from few standard ways which rules can capture. To classify such sentences, a more general understanding of the class’ sentence is required. Hence, we propose the use of a BERT based multi-label multi-head attention sentence classifier.

3.2.1 Network Architecture

The classifier works on contextual embeddings of the input tokens obtained from a transformer’s encoder such as BERT. Any other encoder such as RoBERTa (Liu et al., 2019) can be employed. These encoder architectures emit the input sentence’s representation for the CLS token and embeddings for each of the tokens. Additionally, a domain specific feature extraction module processes the input sentence to emit a k-hot representation denoting presence of audit report specific phrases. This module is currently devised to simply recognize audit domain specific phrases and emit a 1 in the slot for the phrase in the representation. This k-hot representation is then passed to a linear layer to emit a dense representation and its weights are learnt during training. These phrases have been collected upon observation of multiple audit reports and the k-hot representation size is equivalent to the number of these phrases. It is important to note that the domain specific feature extractor is a generic component and can be generalized in ways suitable to the classification problem.

Following this input processing, the architecture consists of multiple class-specific classification heads formed of a combo of an attention layer, a hidden layer and softmax layer. Having such classification heads for each class is necessary as the problem is a multi-label classification and this provides the necessary one-vs-all arrangement. We hypothesize that the class specific attention heads should learn about specific tokens which are indicative of the class and get tuned, while training, to signal for the class, while inference. The attention

layer would then emit a sentence representation re-weighting the token embeddings giving more importance to tokens highly indicative of the class. The consequent hidden layer takes as input a concatenation of the CLS representation, the attention layer emitted representation and the domain specific feature based representation. The softmax layer post this performs the class vs not_class classification. During inference, whichever classification head emits a confidence of 0.5 or greater, the respective class is added to the list of predicted classes for the input sentence. For a detailed network diagram, refer to Appendix D.

3.2.2 Training Data

It is important to note however, that creating annotated data is effort and time consuming and requires domain expertise. With unavailability of annotated data for training the classifier, we resort to weak supervision wherein we label a large set of audit report sentences automatically using the rules devised earlier. We consider a large number of audit reports (See Section 4.1) and run the rule based classification to collect about 16K sentences. After removal of near duplicates, we arrive at about 4.9K sentences which we consider for training the classifier. Additionally, we train the classifier for only 15 out of the above 21 classes (classes marked *Complex Classes* in Table 1), given the understanding that the rest of the 6 classes are easily recognizable through the pattern based rules.

3.3 Ensemble with Rules based classification

To combine the power of both classification approaches and also to check how much generalization the classifier has been able to achieve, we combine them in an ensemble. We allow the rules to first predict the set of possible classes P for an audit report sentence S . Now if the classifier predicts a label for S with confidence greater than 0.5, which is not already in P , the label is added to P , allowed as per the multi-label setting of the classification. This new label prediction may happen if the clas-

sifier has observed certain class indicative aspects of the sentence which the rules have failed to exploit. Only in cases when the rule based approach has predicted the *NAconfirm* class for the sentence, we refrain from predicting using the classifier and predict only the *NA* label.

3.4 Boosting Generalization of the approach

We hypothesized that the classifier, trained on the data labelled by the rules, may not generalize well on sentences which are not labelled by the rules and hence the approach may require more support in terms of generalization in understanding the classes. We explore two ways to boost the generalization of the approach.

3.4.1 Active Learning

One way to achieve the necessary generalization is to add a set of sentences which are not getting classified by the rules and classifier to the training data. To perform this addition in a methodical and an effective manner, we take help of the Active Learning paradigm.

Active learning is a strategy to select some instances from the dataset which are hardest to be correctly classified by the trained classifier. These selected sentences are then added to the training data and the classifier is trained using this supplemented dataset. For finding the sentences which are the most difficult to classify, we developed a strategy called Closest-To-Local-Midpoint (CTLTM) which is a modified version of the query synthesis procedure in Wang et al. (2015). For each pair of classes, this strategy selects those sentences which are approximately equidistant from both the classes. The details of the strategy is present in Appendix C. Sentences selected using this approach are added to the training dataset and the classifier is retrained (C_{AL}). We use C_{AL} as part of the ensemble approach and report the results.

3.4.2 LLM Review

In the original ensemble of the rules and the classifier, we add only the labels from the classifier which have a prediction confidence of 0.5 or greater. Another way to increase the generalization capability of the ensemble approach, is to get the classifier’s low confidence predictions (less than 0.5) reconsidered by an independent reviewer. This is to harness those cases where the classifier may have spotted the correct class, but is less confident. We enable this independent review with the help of a LLM by

prompting it to re-confirm or abandon a candidate label. With this we also pseudo-enact a scenario of considering a LLM as a domain expert which understands these aspects of audit coverage.

To enable this LLM review, we first collect all sentences for which the classifier has predicted atleast one class with confidence less than 0.5 and higher than 0.1. We decide this lower bound, empirically. For each sentence and such possible class, we prepare a prompt consisting of the sentence and the class’ description (Table 3). We then probe the LLM using the prompt and find out whether the class being tested is really applicable. The class descriptions used in the prompts are based on the descriptions provided in Table 1 earlier.

The domain expert further commented that certain classes may get a higher benefit when using language models to discern them, given the larger amount of financial audit discourse those classes are discussed in. His suggested classes were: *payables*, *fixed assets*, *litigations* and *inventory*. To confirm his hypothesis, we devised a small recall measurement experiment. We used the dataset of 4.9K sentences labelled by the rules (the same data used as training data for the classifier) and ran the above LLM review on it, to confirm the rule based label. As this data is not manually labelled, we simply measure the recall i.e. number of sentences where the LLM could successfully confirm the rule based label and not miss it. We ordered the classes according to recall and found out that not only did the expert suggested four classes appear in the top 6 (recall ≥ 0.8), but also introduced us to 2 more similar classes: cost records and internal control system. We term these 6 classes as LLM-HR (High Recall) classes. As the final approach - Ensemble (R + C + LLM-HR), for a sentence if the classifier has a low confidence prediction which is one of these high recall classes, we review the label using the LLM. If confirmed, we add the class to the existing set of labels provided by the ensemble.

4 Experiments and Evaluation

4.1 Dataset

We used the web-scraped audit reports made available by authors of (Maka et al., 2020). We consider 3744 reports for the year 2014 from which 932 reports which were too small (less than 35 sentences) or too noisy were removed leading to a final set of 2812 reports.

Sentence: We are of the opinion that in view of Memorandum of Settlement with the workers the company should make a provision of crystallized dues of Rs 40 Crores, irrespective of sale of Mohali Assets	
Classifier Predictions: fixed assets (0.187), sales, services and revenue (0.141)	
Prompt Template: <Sentence>. Does the previous sentence talk about <Class description>? Answer as Yes or No.	
Class	Final Prompt
fixed assets	We are of the opinion that in view of Memorandum of Settlement with the workers the company should make a provision of crystallized dues of Rs 40 Crores, irrespective of sale of Mohali Assets. Does the previous sentence discuss about fixed assets such as equipment, land, building, plant, machinery and their physical verification? Answer as Yes or No.
sales, services and revenue	We are of the opinion that in view of Memorandum of Settlement with the workers the company should make a provision of crystallized dues of Rs 40 Crores, irrespective of sale of Mohali Assets. Does the previous sentence discuss about revenue from sale of goods and services excluding sale of shares and assets? Answer as Yes or No.

Table 3: Example Prompt Creation

Test dataset: We select a set of 10 audit reports which are labelled manually for the 22 classes (including NA) to form the test set. As part of the annotation guidelines, we used the descriptions in Table 1. Two annotators were part of the annotation exercise, one of which was the domain expert. High inter-annotator agreement was observed and any conflicts were resolved through discussion. The test set consists of a total of 1668 annotated sentences (class-wise statistics in Appendix A).

4.2 Baselines

We experiment with a number of standard machine learning classifiers as baseline approaches. We implement these approaches through the classifiers provided in scikit-learn with their default parameters while setting the class weights as “balanced” wherever possible. It is important to note that we use the same rule-based approach annotated data as training for these classifiers and report results on the 15 complex classes as we do for the BERT-based multi-label classifier.

Additionally, we also try using ChatGPT as a baseline and provide it a suitable prompt (details in Appendix B) to make it predict the suitable classes on the input sentence.

4.3 Experimentation Details

We considered a set of documents separate from the test set to tune the rules, tried few best configurations on the test set and selected the best one. We then use the best set of rules to label the data for creating training data for the classifier and the high recall experiment of the LLM. Further, for hyperparameter tuning of the BERT based classifier, we used a 20% validation split of the training data. Certain important hyperparameters to note are: (batch_size: 8 with gradient accumulation of 8 steps, learning rate: 0.00005, epochs: 16). We used

two attention heads in each class’ attention module, to attend to two important words in that sentence indicative of the class. We only allowed the final encoder layer in the BERT model to get fine-tuned while keeping all other layers frozen. Also, for the LLM experiment we employed the Falcon-7B-instruct (Almazrouei et al., 2023) model, which is a resource and license friendly model capable of responding to question like prompts, similar to what we have devised.

4.4 Evaluation and Analysis

Approach	P	R	F1
Rules (R)	0.887	0.557	0.684
Naive Bayes [†]	0.820	0.307	0.446
SVM (Linear) [†]	0.722	0.698	0.710
Logistic Regression [†]	0.670	0.742	0.704
Random Forests [†]	0.844	0.570	0.680
Gradient Boosting [†]	0.853	0.589	0.697
ChatGPT (Zero-shot)	0.487	0.557	0.520
BERT-based Classifier (C) [†]	0.849	0.639	0.729
Ensemble (R + C)	0.843	0.652	0.735
Ensemble (R + C _{AL})	0.835	0.692	0.757
Ensemble (R + C + LLM-HR)	0.823	0.692	0.752

Table 4: Comparative Performance of different baselines and proposed approaches. († indicates evaluation over 15 complex classes)

In Table 4, we report the performance of the baselines and different proposed approaches. We use precision, recall and F1-score micro-averaged over multiple labels due to the multi-label setting. The rule based approach based on concrete and specific rules performs with the best precision as desired. This is an important reason for using the rule based approach for creating annotated data. Further, we see that in terms of F1-score, some of the baselines such as SVM, Logistic Regression and Gradient Boosting, outperform the rule based approach thereby implying that they are able to

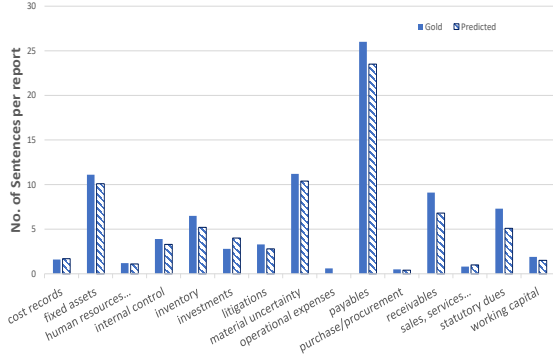


Figure 1: Class-wise coverage in the Test set

generalize even while trained using rules annotated data. The BERT-based classifier outperforms both the rule-based approach and the baseline classifiers. Further the ensemble of rules-based approach and the BERT-based classifier performs slightly better, with increase in recall by 2%.

If we observe the generalization boosting approaches, both the Active Learning (AL) one and the LLM review (LLM-HR) one perform the best and give about a 6% increase in recall, thereby increasing the overall F1 by 2-3%. The AL approach, performs slightly better, but requires 500 labelled sentences to be infused as part of the retraining exercise, which may be difficult to obtain given limited availability of domain expertise. The LLM review approach, though only requires the effort for prompt creation, it does requires domain support to identify the right high recall classes.

In Figure 1, we present how different is the distribution of predicted classes over the test set when compared to the distribution of the classes as per gold labels. The ensemble approach maintains the distribution well (with Jensen-Shannon Divergence of 0.005) and is in conformance to the gold distribution across classes. This makes the ensemble a close approximation of the true underlying distribution of classes and hence worthy for use in analysis on a larger set of reports (Section 5).

On detailed analysis of class-wise results for the best approach: Ensemble (R + C_{AL}), we observe that from the set of complex classes, *cost records*, *internal control system*, *fixed assets*, *working capital*, *human resources*, *utilization of ipo* and *inventory* perform well and have an individual class F1 of 0.8 or greater. Classes which perform moderately well ($0.7 \leq F1 < 0.8$) include *payables*, *investments*, *receivables*, and *material uncertainty*. There lies scope to

improve the performance for these classes. Some of the low performing classes ($F1 < 0.7$) are *litigations*, *statutory dues*, *private placement or preferential issues*, *purchase & procurement* and *sales, services & revenue*. Investigating further for these classes, we find that due to presence of certain indicative phrases in the sentence, which are used in a different semantic context, confuse the approach. For example, presence of the phrase the Company has sold and transferred its branded domestic formulations business, prompts the approach to assign *sales, services and revenue*, which is not valid here as this is not related to sales of products or services, but a business division. Similarly, reference to possible scenarios in the sentence also leads to false positives, such as the following sentence gets labelled as *material uncertainty*, when it is referring to a possible negative implication: Relying on the assertions as detailed in notes no adjustments have been made in the financials towards possible impairment.

A small discussion on why ChatGPT performs on the lower side is also important. Firstly, we observed that in spite of specifying to classify the input sentence into the given class names, ChatGPT started predicting new class names formed of phrases related to the correct class name. Secondly, even on specifying to emit multiple relevant classes, it still sticks to predicting only one class. When forced, it starts predicting lots of irrelevant classes. Thirdly, at times it simply classified some of the input sentences and then generically specified that “other sentences can be classified similarly”. Overall we believe that in a challenging scenario with large number of domain specific classes with complex semantics, ChatGPT output is not usable in deployed applications.

5 Audit Report Coverage Analysis

Audit report coverage refers to the extent and scope of an audit report, detailing what aspects of an organization’s financial statements, instruments and operations have been examined and reported on, by an external auditor.

As iterated earlier, we aim to measure the audit coverage through classification of sentences into audit aspects specified in regulatory checklists. We consider that if a sentence is mapped to a class, it is generally commenting about aspects of that class and in turn achieves the objective of checking that

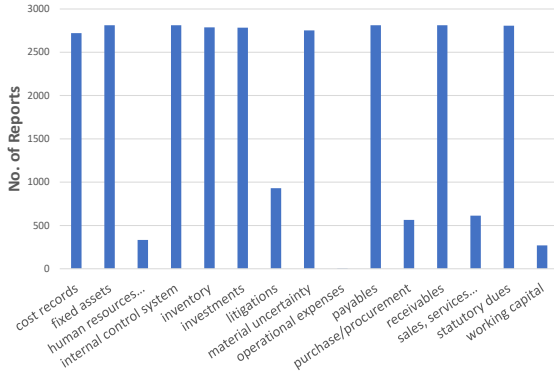


Figure 2: Checklist Coverage

class. At the document level, sentences commenting on the various classes can then be an indicator of which classes were checked and reported upon.

We consider the set of 2812 reports for measuring audit report coverage on the complex classes and report on it from two perspectives:

1. **Checklist Coverage:** In this perspective, if a class appears at least once in the audit report then we can consider that aspect is covered in the audit. This is mainly important to check the compliance requirements where the regulatory bodies expect mandatory coverage of specific areas.

Figure 2 shows the number of reports in which the complex classes were reported on at least once. We can conclude that classes like *human resource and payroll processing*, *litigations*, *operational and administrative expenses*, *purchase and procurement*, and *working capital* show considerably lower coverage than other classes. The lower coverage could be due to (a) absence of litigations or (b) lower importance given by the auditor for that aspect. E.g., most companies may not be involved in litigations or issues relating to human resources, hence those aspects may be skipped.

2. **Weighted Coverage:** Through this perspective, the number of sentences specifying a certain class can be considered as weight/importance devoted to the corresponding aspect. This can be used by the stakeholders for analyzing the weightage given by the audit report for a specific aspect, for e.g. while lending money to a firm the bank can check the focus given on aspects like *payables*, *internal control*, *reevaluation of fixed assets*, etc.

Figure 3 shows the distribution of weightage for the complex classes over the considered reports. We observe that some of the classes such as *cost records* and *working capital* have comparatively

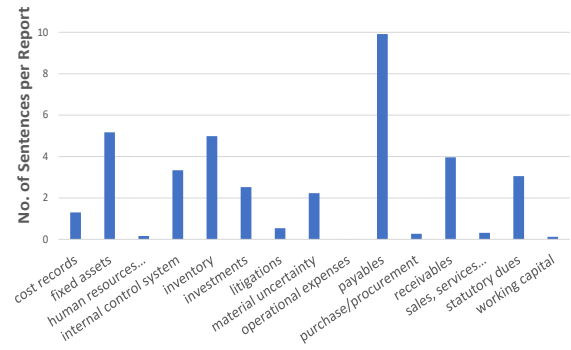


Figure 3: Weighted Coverage

less weightage than classes such as *payables* and *fixed assets*. This helps in understanding the importance auditors place on certain aspects and their implications on the functioning of the company.

6 Conclusion and Future Work

In this paper, our objective was to find whether all the necessary audit aspects are being covered in an audit report. We proposed a set of 21 classes corresponding to these audit aspects. We proposed a weakly supervised approach for automatic multi-class multi-label classification of sentences in an audit report. Due to absence of training data, we use a rule-based technique to automatically create labelled dataset for training a BERT-based sentence classifier. Further, we employ two novel ways to improve the generalization – (i) through an active learning based approach which needs manual annotation efforts and, (ii) through a LLM based review which needs efforts for prompt engineering. Given the complex and domain specific semantics of the classes and unavailability of labelled data, we were still able to achieve the F1-score of more than 75% with our approaches outperforming several baselines. We also showed the utility of the proposed classification approaches to measure audit coverage on a large dataset of 2.8K audit reports.

As part of future work, we would like to explore open source LLMs further for our sentence classification problem. From domain point of view, we plan to extend our techniques for different stakeholders such as regulatory bodies or banks to automatically evaluate the audit reports in an unbiased way. We also plan to apply these techniques for audits in other domains like software quality audits or Environment, Social & Corporate Governance (ESG) audits, using domain knowledge such as audit guidelines from the respective domains.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Hestlow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Alvin A. Arens and James K. Loebbecke. 1999. *Auditing: An Integrated Approach*, 8th edition. Pearson.

S. Basu. 1997. The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics*, 24:3–37.

Mark S Beasley, Joseph V Carcello, Dana R Hermanson, Committee of Sponsoring Organizations of the Treadway Commission, et al. 1999. Fraudulent financial reporting: 1987-1997: an analysis of us public companies.

Madan Lal Bhasin. 2013. Corporate accounting fraud: A case study of satyam computers limited. *Open Journal of Accounting*, 2(2).

ICAI. 2017. Internal audit checklist. <https://kb.icai.org/pdfs/44970iasb34918.pdf>. [Online; accessed 8-September-2023].

ICAI. 2020. ICAI'S GUIDANCE NOTE ON CARO 2020 (CARO). <https://wirc-icai.org/wirc-reference-manual/part2/icai-guidance-note-on-caro-2020.html>. [Online; accessed 8-September-2023].

C. Lennox and B. Li. 2019. When are audit firms sued for financial reporting failures and what are the lawsuit outcomes? *Contemporary Accounting Research*, 37(3):1370–1399.

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*.

Kiran Maka, S. Pazhanirajan, and Sujata Mallapur. 2020. Selection of most significant variables to detect fraud in financial statements. *Materials Today: Proceedings*.

Jeffrey Pennington, Richard Socher, and Christopher D 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.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

K.J. Reichelt and D. Wang. 2010. National and office-specific measures of auditor industry expertise and effects on audit quality. *Journal of Accounting Research*, 48(3):647–686.

Nopmanee Tepalagul and Ling Lin. 2015. Auditor independence and audit quality: A literature review. *Journal of Accounting, Auditing and Finance*, 30(1):101–121.

Liantao Wang, Xuelei Hu, Bo Yuan, and Jianfeng Lu. 2015. Active learning via query synthesis and nearest neighbour search. *Neurocomputing*, 147:426–434.

Ray Whittington and Kurt Pany. 2021. *Principles of Auditing and Other Assurance Services*, 22 edition. McGraw-Hill Education.

A Test Dataset Statistics

The test set consists of a total of 1668 annotated sentences. The class-wise statistics are presented in Table 5.

Class	#Annotated Sentences
NA	733
payables	260
material uncertainty	112
fixed assets	111
receivables	91
statutory dues	73
inventory	65
internal control system	39
litigations	33
investments	28
private placement or preferential issues	23
working capital	19
cost records	16
human resources and payroll processing	12
nidhi company	10
utilization of ipo and other public offers	10
fraud reporting	10
sales, services and revenue	8
operational and administrative expenses	6
purchase and procurement	5
approval of managerial remuneration	3
non-cash transactions	1

Table 5: Class-wise annotations

There were 5 other classes that were defined based on the auditing checklist namely *corporate social responsibility*, *resignation of statutory auditors*, *remarks by auditors of included companies*, *related party transaction* and, *register under rbi act*. As these 5 classes were not present in the labelled data, we decided to include only the ones shown in Table 1, in the current analysis.

B Description of the ChatGPT Prompt

We use ChatGPT's user interface to perform the classification of the sentences in the test set by

prompting it with suitable prompts. The prompt consists of a main instruction, descriptions of the 15 complex classes and finally a set of sentences to classify. The prompt template is shown in Table 6, where text in round brackets is for explanation only. As can be seen, that this is a zero-shot setting of classifying using an LLM. A few shot setting, as part of in-context learning, can also be tried where examples of sentences and their gold class can be provided. However, selection of the classes to give as examples and maintaining the instruction’s context are some important challenges, exploration of which we keep as future work.

(—Main Instruction—)

The task is to classify sentences in a financial audit report into one or more of the following classes. Each line below mentions a class name followed by its description.

(—Class Descriptions—)

1. cost records: About maintenance of cost records.
2. fixed assets: About fixed assets such as equipment, land, building, plant, machinery and their physical verification.
3. human resources and payroll processing: About human resources and payroll processing such as employee wages, leaves, bonus, pension, full and final settlement, policies for leave, gratuity or pension.
4. internal control system: About internal control procedures.
- ...
14. statutory dues: About depositing statutory dues like provident fund, ESI, income tax, sales tax, VAT, service tax, GST, duty of customs, duty of excise.
15. working capital: About working capital, cash credit and bank balance.

(—Input Sentences for Classification—)

What are the applicable classes for the following sentences? Simply print the output as Sentence ID: Class name.

Sentence 1: We have audited the accompanying financial statements of ...

Sentence 2: Management is responsible for the preparation of these financial statements that give a true

...

Sentence 10: We conducted our audit in accordance with the Standards on Auditing issued ...

Table 6: ChatGPT Prompt Template

C Details about the Active Learning strategy

For finding the sentences which are the most difficult to classify, we developed a strategy called Closest-To-Local-Midpoint (CTLM) which is a modified version of the query synthesis procedure in (Wang et al., 2015). In CTLM, we first find the center of the cluster having all the sentences belonging to a class in Euclidean space. To elaborate, let us assume, we have a class C_1 . We know from the

predefined rules (as mentioned in Section 3.1), a set S_{C_1} of sentences belonging to C_1 . We transform each sentence $s \in S_{C_1}$ to a vector $\Delta_s \in \mathbb{R}^{300}$, by taking the average of the Glove embeddings (Pennington et al., 2014) of each word in s . The words from a predefined set of insignificant stop words are omitted while computing Δ_s . Once we have the respective vectors for each of the sentence belonging to C_1 , we find the center $\mu(\Delta_{C_1})$ of C_1 by computing the mean vector of all the transformed sentences. Given a set $C = \{C_1, C_2, \dots, C_m\}$ of m (here 15) such classes, we have a set $\mu(\Delta_C) = \{\mu(\Delta_{C_1}), \mu(\Delta_{C_2}), \dots, \mu(\Delta_{C_m})\}$ of their respective centres, which are the representatives of the respective classes. Now we find the sentences that should be difficult to classify. We find the pairwise mid-points of mean vectors of the classes in 300-dimensional space and select the sentences which are nearest to these midpoints. The intuition is that, *the sentences which are approximately equidistant from the cluster centres of two classes will be classified with lowest confidence of belonging strictly to a single class*. As there are large number of sentences common between most of the audit reports, the sentences closest to midpoints of different pairs could be very similar. We want sentences as dissimilar as possible, so the classifier can learn different aspects. To avoid this we select a large number of sentences (2000 here) per pair in descending order of cosine similarity. From these sentences, we removed the common sentences and sentences which were very similar. After removing these common and similar sentences, we were left with 477 distinct, dissimilar and toughest to classify sentences. We ensured that these sentences are ones where the rules are unable to predict any class. These sentences were then labeled by the annotators and then were added to the training set of the classifier.

D BERT-based classifier Network Diagram

The neural network diagram of the BERT-based classifier is shown in Figure 4.

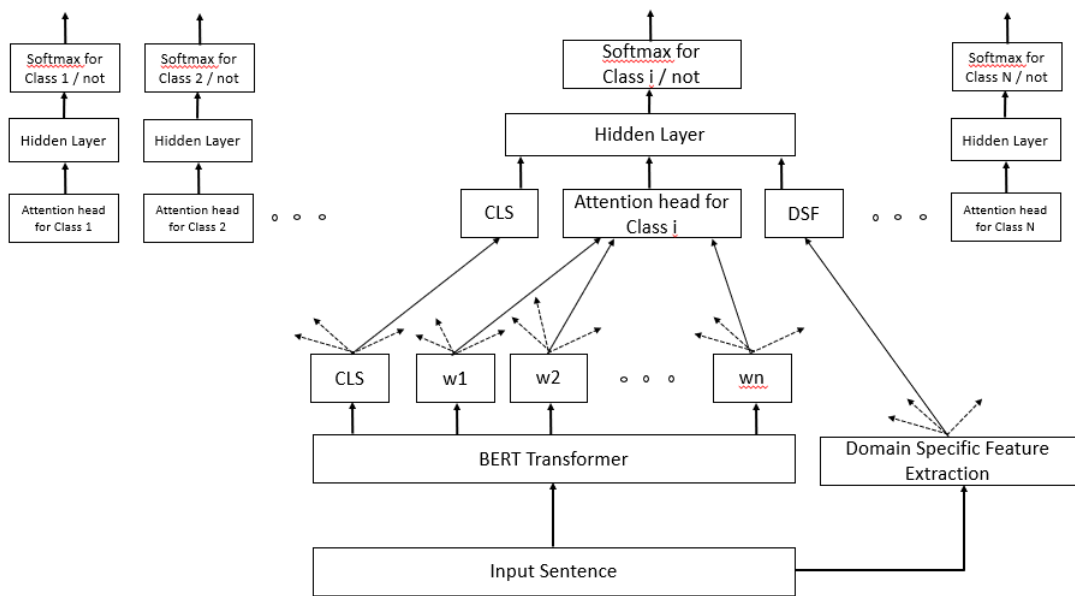


Figure 4: BERT based Multi-label Multi-headed Attention Classifier

GPT-FinRE: In-context Learning for Financial Relation Extraction using Large Language Models

Pawan Kumar Rajpoot**
pawan.rajpoot2411@gmail.com
MUST Research
Bangalore, Karnataka, India

Ankur Parikh
ankur.parikh85@gmail.com
UtilizeAI Research
Bangalore, Karnataka, India

Abstract

Relation extraction (RE) is a crucial task in natural language processing (NLP) that aims to identify and classify relationships between entities mentioned in text. In the financial domain, relation extraction plays a vital role in extracting valuable information from financial documents, such as news articles, earnings reports, and company filings. This paper describes our solution to relation extraction on one such dataset REFinD. The dataset was released along with shared task as a part of the Fourth Workshop on Knowledge Discovery from Unstructured Data in Financial Services, co-located with SIGIR 2023. In this paper, we employed OpenAI models under the framework of in-context learning (ICL). We utilized two retrieval strategies to find top K relevant in-context learning demonstrations / examples from training data for a given test example. The first retrieval mechanism, we employed, is a learning-free dense retriever and the other system is a learning-based retriever. We were able to achieve 3rd rank overall (model performance and report). Our best F1-score is 0.718.

Keywords: relationship extraction, gpt, in context learning, text tagging, REFinD, KDF, SIGIR, CEIL, ICL, In Context Learning, GPT NEO, Finance, Large Language Model

ACM Reference Format:

Pawan Kumar Rajpoot and Ankur Parikh. 2023. GPT-FinRE: In-context Learning for Financial Relation Extraction using Large Language Models. In *Proceedings of The 4th Workshop on Knowledge Discovery from Unstructured Data in Financial Services (KDF @SIGIR '23)*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

The emergence of large language models (LLMs) such as GPT-3 [6][12] represents a significant advancement in natural language processing (NLP). These models have expertise in variety of domains and hence they can be used

*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDF @SIGIR '23, July 27, 2023, Taipei, Taiwan

© 2023 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>


In the third quarter of 2016 , QCR HOLDINGS INC acquired CSB , headquartered in Ankeny , Iowa . 

Figure 1. Relation Extraction example, here both organizations are connected with "acquired by" relation.

as it is in multiple NLP tasks. Traditionally language models use separate pre-training-and fine-tuning pipelines [1] [3] [5] [4] [9] where fine-tune stage follows pre-training. Models are fine-tuned on a task-specific dataset in a fully-supervised manner. More recently a new paradigm known as in-context learning (ICL) [6][14] is being used which formulates an NLP task such that LLMs make predictions by learning from demonstrations. These demonstrations are presented to the LLMs in the context prompt itself.

Under the framework of ICL, LLMs achieve remarkable performance rivaling previous fully-supervised methods even with only a limited number of demonstrations provided in the prompt in various tasks such as solving math problems, commonsense reasoning, text classification, fact retrieval, natural language inference, and semantic parsing [6] [14] [8]. Recently, ICL based approach[16] is utilized for Relation Extraction (RE) task. RE seeks to identify a semantic relationship between a given entity pair mentioned in a sentence, which is the central task for knowledge retrieval requiring a deep understanding of natural language. The approach achieves improvements over not only existing GPT-3 baselines, but also on fully-supervised baselines. Specifically, it achieves SOTA performances on the SemEval and SciERC datasets, and competitive performances on the TACRED and ACE05 datasets.

Retrieval of examples to demonstrate is a key factor in the overall performance on these pipelines. LLMs can relate to the presented "to be predicted" data point more if the contextual examples predicted are similar to it. More relevant examples help us to leverage more out from LLMs both in terms of improvement in performance and less hallucination as examples can demonstrate model not to hallucinate in some cases.

In this paper, we employed GPT-3.5 Turbo and GPT-4 under the framework of ICL for the relation extraction task on REFinD dataset. We utilized two retrieval strategies to find top K relevant in-context learning demonstrations / examples from training data for a given test example. The first mechanism we have employed is a learning-free dense retriever. The other system we have utilized is a learning-based retriever [13].

Entity (subj)	Relation	Entity (obj)
PERSON	title	TITLE
PERSON	employee_of	ORGANIZATION
PERSON	member_of	ORGANIZATION
PERSON	founder_of	ORGANIZATION
PERSON	employee_of	UNIVERSITY
PERSON	member_of	UNIVERSITY
PERSON	attended	UNIVERSITY
PERSON	member_of	GOV. AGENCY
ORGANIZATION	formed_on	DATE
ORGANIZATION	acquired_on	DATE
ORGANIZATION	headquartered_in	GPE
ORGANIZATION	operations_in	GPE
ORGANIZATION	formed_in	GPE
ORGANIZATION	shares_of	ORGANIZATION
ORGANIZATION	subsidiary_of	ORGANIZATION
ORGANIZATION	acquired_by	ORGANIZATION
ORGANIZATION	agreement_with	ORGANIZATION
ORGANIZATION	revenue_of	MONEY
ORGANIZATION	profit_of	MONEY
ORGANIZATION	loss_of	MONEY
ORGANIZATION	cost_of	MONEY
	no/other_relation	

Figure 2. REFinD dataset relation and entity types.

2 Preliminary Background

2.1 Task Definition

As per the challenge "Relation Extraction is the task of automatically identifying and classifying the semantic relationships that exist between different entities in a given text." This shared task is a part of "Knowledge Discovery from Unstructured Data in Financial Services" (KDF) workshop which is collocated with SIGIR 2023.

Let C denote the input context and e_1 in C , e_2 in C denote the pair of entity pairs. Given a set of predefined relation classes R , relation extraction aims to predict the relation y in R between the pair of entities (e_1, e_2) within the context C , or if there is no predefined relation between them, predict y ="no relation".

2.2 Data

The dataset [18] released with this task is the largest relation extraction dataset for financial documents to date. Overall REFinD contains around 29K instances and 22 relations among 8 types of entity pairs. REFinD is created using raw text from various 10-X reports (including 10-K, 10-Q, etc. broadly known as 10-X) of publicly traded companies obtained from US Securities and Exchange Commission. Figure-2 shows different entity types and relations exist between them.

2.3 In Context Learning

In-context learning (ICL) refers to one of the core emergent abilities [17] that infers new tasks from context. We use the terms 'in-weights learning' and 'in-context learning' from prior work on sequence models [6] to distinguish between gradient-based learning with parameter updates and gradient-free learning from context, respectively. Formally, each training instance is first linearized into an input text $x = (x_1 \dots x_n)$ and an output text $y = (y_1 \dots y_n)$, where for all tokens $x_1 \dots x_n, y_1 \dots y_n$ in V and V is the vocabulary set of the

LM. Given a new 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 \sim refers to decoding strategies (e.g., greedy decoding and nuclear sampling [11]), 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 [13] [10] [19].

3 GPT-FinRE

GPT-RE is formalized under the ICL framework, using GPT models as shown in Figure-3.

3.1 Prompt Construction

We construct a prompt for each given test example, which is fed to the GPT models. Each prompt consists of the following components.

Task Description and Predefined Classes : We provide a succinct overview of the RE task description and the subset of predefined classes R , denoted by O . This subset is all possible relations exist between entity types of e_1 and e_2 . The model is explicitly asked to output the relation, which belongs to the O . Otherwise, the model will output "no relation".

K-shot Demonstrations : In the demonstration part, we reformulate each example by first showing the input prompt x -demo = Prompt(C, e_1, e_2) and the relation label y -demo.

Test Input : Similar to the demonstrations, we offer the test input prompt x -test, and GPT models are expected to generate the corresponding relation y -test.

3.2 Retrieval Systems

We have employed two retrieval strategies to find top K relevant in-context learning demonstrations / examples from training data for a given test example.

3.2.1 KNN with OpenAI Embeddings. Since ICL demonstrations closer to the test sample in the embedding space result in more consistent and robust performance [10]. We utilized the KNN to retrieve the most similar examples in the training set as the few-shot demonstrations for each test example. As this learning-free dense retriever relies on the choice of the embedding space, we used OpenAI embeddings (text-embedding-ada-002) to obtain example representations. For similarity search, we used FAISS tool [2].

3.2.2 EPR (Efficient Prompt Retrieval). This learning-based dense retriever is trained to retrieve a better singleton in-context example [13], and Top- K most similar examples are selected in the inference stage. This method for retrieving prompts for in-context learning uses annotated data and a LM. Given an input-output pair, it estimates

Retriever	LLM	F1-Score
KNN with openAI embeddings	GPT 3.5 Turbo (Examples: 5 retrieved + 5 random per possible relation)	0.643
KNN with openAI embeddings	GPT 4 (Examples: 5 retrieved + 5 random per possible relation)	0.697
EPR with GPT-Neo-2.7B	GPT 4 (Examples: 2 retrieved + 3 random per possible relation)	0.703
EPR with GPT-Neo-2.7B	GPT 4 (Examples: 5 retrieved + 4 random per possible relation)	0.718

Table 1. Our performance on test data with different combinations of retriever and LLM

the probability of the output given the input and a candidate training example as the prompt, and labels training examples as positive or negative based on this probability. It then trains an efficient dense retriever from this data, which is used to retrieve training examples as prompts at test time. Due to limited access to OpenAI, we have used the gpt-neo-2.7B model [7] as our choice of LM.

3.2.3 Random Class Examples. Along with KNN / EPR based examples, we also added K examples randomly for each possible class between two entity types to add more variety in the final prompt.

4 Experiments

Due to limited access and cost associated with OpenAI, We performed 4 primary experiments on the test dataset. We tried various rule based heuristics to improve the F1-score, but it didn't work as expected. We used retriever implementations from ¹.

5 Results

The results are shown in Table-1. Our best F1-Score is 0.718. We got 4th position in the shared-task. We find that GPT 4 performs better than GPT 3.5 Turbo. We also find that learning based retriever (EPR) outperforms learning-free retriever (KNN with OpenAI embeddings).

6 Future Work

In future we want to utilize GPT 4 for EPR. We also want to use different retrieval approaches such as Compositional Exemplars for In-context Learning (CEIL)[15].

7 Conclusion

This work explores the potential of GPT + ICL on Financial Relation Extraction (REFinD dataset). We used two retrieval mechanisms to find similar examples: (1) KNN with OpenAI Embeddings (2) EPR. We tried two different GPT models: (1) GPT 3.5 Turbo and GPT 4. The experimental results show that GPT 4 with learning based retriever EPR is giving the best F1-Score of 0.718.

References

- [1] 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [2] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
- [3] 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)*. Association for Computational Linguistics, Hong Kong, China, 3615–3620. <https://doi.org/10.18653/v1/D19-1371>
- [4] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv:1909.11942 [cs.CL]
- [5] 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. arXiv:1910.10683 [cs.LG]
- [6] 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 (2020), 1877–1901.
- [7] 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 (2020).
- [8] Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained Language Models Yield Few-Shot Semantic Parsers. *CoRR* abs/2104.08768 (2021). arXiv:2104.08768 <https://arxiv.org/abs/2104.08768>
- [9] 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*. Chinese Information Processing Society of China, Huhhot, China, 1218–1227. <https://aclanthology.org/2021.ccl-1.108>
- [10] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What Makes Good In-Context Examples for GPT-3? *CoRR* abs/2101.06804 (2021). arXiv:2101.06804 <https://arxiv.org/abs/2101.06804>
- [11] 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 (2022).
- [12] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh

¹<https://github.com/HKUNLP/icl-ceil>

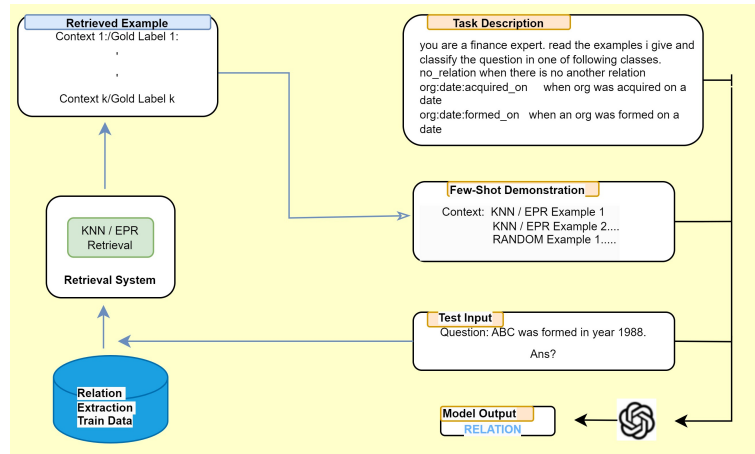


Figure 3. GPT-FinRE pipeline flow

Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. *arXiv:2201.08239* [cs.CL]

- [13] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 2655–2671. <https://doi.org/10.18653/v1/2022.naacl-main.191>
- [14] Sewon Min, Xinxu Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? *arXiv preprint arXiv:2202.12837* (2022).

- [15] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional Exemplars for In-context Learning. (2023). *arXiv:2302.05698* [cs.CL]
- [16] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *arXiv preprint arXiv:2305.02105* (2023).
- [17] 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* (2023).
- [18] Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. REFinD: Relation Extraction Financial Dataset. *arXiv preprint arXiv:2305.18322* (2023).
- [19] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. *arXiv:2212.10375* [cs.CL]

Multi-Lingual ESG Impact Type Identification

Chung-Chi Chen,¹ Yu-Min Tseng,² Juyeon Kang,³ Anaïs Lhuissier,³ Yohei Seki,⁴
Min-Yuh Day,⁵ Teng-Tsai Tu,⁶ Hsin-Hsi Chen⁷

¹AIST, Japan

²Data Science Degree Program, National Taiwan University and Academia Sinica, Taiwan

³3DS Outscale, France, ⁴University of Tsukuba, Japan

⁵Graduate Institute of Information Management, National Taipei University, Taiwan

⁶Graduate Institute of International Business, National Taipei University, Taiwan

⁷Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan

Abstract

Assessing a company’s sustainable development goes beyond just financial metrics; the inclusion of environmental, social, and governance (ESG) factors is becoming increasingly vital. The ML-ESG shared task series seeks to pioneer discussions on news-driven ESG ratings, drawing inspiration from the MSCI ESG rating guidelines. In its second edition, ML-ESG-2 emphasizes impact type identification, offering datasets in four languages: Chinese, English, French, and Japanese. Of the 28 teams registered, 8 participated in the official evaluation. This paper presents a comprehensive overview of ML-ESG-2, detailing the dataset specifics and summarizing the performance outcomes of the participating teams.

1 Introduction

In the rapidly shifting global business milieu, relying solely on traditional financial metrics to gauge companies is no longer adequate. The imperative for companies to make meaningful contributions to society and the environment, underpinned by robust governance, has intensified. These principles, often denoted as Environmental, Social, and Governance (ESG) factors, have surged in significance for stakeholders from investors to consumers. Evaluating a company’s adherence to these principles is crucial not only for its sustainable growth but also for setting the benchmark for corporate responsibility in the modern era.

The increasing need to precisely appraise and contrast ESG ratings across corporations has ignited extensive research and discourse. The influence of news and contemporaneous events on these ratings offers a captivating subject for study. This dynamic environment necessitates a resilient and malleable approach. While numerous rating systems exist, our focus leans towards the MSCI ESG rating guidelines as the foundation for the Multi-Lingual ESG (ML-ESG) shared task series.

In this backdrop, the ML-ESG shared task series was initiated to cultivate a deeper understanding of news-driven ESG ratings. In a world growing more intertwined, the imperative of incorporating multiple languages in these assessments is paramount. In the inaugural ML-ESG, our focus was on pinpointing ESG issues (Chen et al., 2023). Progressing to the second iteration, ML-ESG-2 hones in on the classification of impact types, featuring datasets in Chinese, English, French, and Japanese.¹ The objective of ML-ESG-2 is discerning whether news pieces indicate risks or opportunities for company operations through an ESG lens.

This paper aims to shed light on the methodologies, datasets, and discoveries of ML-ESG-2. Through the concerted effort of diverse teams, we aspire to offer a comprehensive perspective on the latest advancements in multi-lingual ESG impact type identification. Our endeavor is to equip both scholars and industry professionals with insights into this emerging domain.

2 Dataset and Task Setting

Table 1 presents the dataset statistics for ML-ESG-2. The Chinese (Tseng et al., 2023) and Japanese (Kannan and Seki, 2023) datasets are previously published, while the English and French datasets make their debut in ML-ESG-2.

2.1 Chinese, English, and French

The Chinese dataset is derived from ESG-centric news articles found on ESG-BusinessToday (in Chinese)². Meanwhile, the English and French datasets amalgamate articles from sources including ESGToday (in English)³, RSEDATANEWS (in

¹Japanese dataset used different guidelines and data sources from other datasets. Please refer to Section 2.2 for details.

²<https://esg.businessstoday.com.tw/>

³<https://www.esgtoday.com/category/esg-news/companies/>

		English	French	Chinese	Japanese
Train	Opportunity (Positive)	694	458	536	460
	Risk (Negative)	114	360	58	49
	Other	-	-	666	387
Development	Opportunity (Positive)	-	-	60	-
	Risk (Negative)	-	-	6	-
	Other	-	-	74	-
Test	Opportunity (Positive)	191	89	67	115
	Risk (Negative)	27	111	7	13
	Other	-	-	82	97
Total		1,026	1,018	1,556	1,121

Table 1: Data statistics.

Team	Best Performing Model
LIPI	FinBERT, Translate to English, T5-Based Data Augmentation
231	ChatGPT Summarization, RoBERTa-Chinese, Convert to Simplified Chinese
FinNLU	FlauBERT, mBERT, ALBERT, TF-IDF features, and LSA features

Table 2: Best performing model proposed by each team for Chinese dataset.

French)⁴, and Novethic (in French)⁵.

The annotation schemes for the Chinese, English, and French datasets categorize news articles as either opportunities or risks. Within the Chinese dataset, additional delineations are made for articles that do not fit the aforementioned categories: they are classified as “Cannot Distinguish (related to company)”, “Related to ESG, but not related to company”, and “Not related to ESG topic”. In total, there are 1,556 instances for the Chinese dataset, 1,026 for the English, and 1,018 for the French.

2.2 Japanese

The Japanese dataset is sourced from EDINET⁶, which were published from Financial Services Agency of Japan (JFSA). Unlike the other datasets, the Japanese collection emphasizes company annual reports (called annual securities reports, and known as *Yuuka Shoken Houkokusho* in Japanese).

During the annotation phase, annotators allocated sentiment labels at the sentence level. These labels encompassed categories such as “Positive,” “Negative,” “Neutral,” and “None” (indicating the sentence’s irrelevance to ESG topics).

In delving deeper into the nuances of these annotations, we observed that the majority of sentences labeled as “positive” or “negative” resonate with the “opportunity” or “risk” ESG impact types, respectively. Interestingly, some of sentences deemed “neutral” also aligned with the “opportunity” impact

type. Our examination of the inter-annotator agreement, gauged via Cohen’s κ coefficient between the type of ESG impact and sentiment annotations, yielded a result of 0.980, signifying an almost perfect agreement.

For the Japanese analysis, a total of 1,121 instances are available. For an exhaustive description and analysis, we direct the reader to [Kannan and Seki \(2023\)](#).

3 Method

3.1 English and French

The participants bring forth a variety of methodologies, including pre-trained transformers, fine-tuning, prompt engineering, and ensemble learning, demonstrating a diverse set of approaches. Most participant systems underscore the significance of mitigating class imbalances through various data augmentation techniques, notably through translation. The translation of data from French, Japanese, or Chinese into English serves to enhance model performance by augmenting the sample size. Furthermore, we observe that applying the same architectural pipeline across all four languages yields successful results for some languages but not for others. This implies the need to consider more language-specific resources or models to improve overall performance. [Qiu et al. \(2023\)](#) and [Vardhan et al. \(2023\)](#) use pre-trained transformer models, incorporating various data and feature augmentation techniques to enhance performance, including translation, summarization, and data paraphrasing. [Polyanskaya and Brillet \(2023\)](#) demonstrates the superiority of GPT-3.5 Turbo over BERT for En-

⁴<https://www.rsedatanews.net/>

⁵<https://www.novethic.fr/actualite/environnement.html>

⁶<https://disclosure2.edinet-fsa.go.jp/WEEEK0010.aspx>

glish dataset which was enhanced by the translated French dataset. (Winatmoko and Septiandri, 2023) explores ST5 and SBERT for generating embeddings and in Mishra (2023), fine-tuning Llama2 on the English dataset with prompts detailing the classification criteria gives the best result, on both the English and French datasets. The Veeramani et al. (2023) introduces an ensemble learning method with mBERT, FlauBERT, ALBERT, and MLP models, incorporating feature representations (LSA and TF-IDF), demonstrating superior performance with early fusion ensemble across all four languages. Billert and Conrad (2023) describes an adapter-based framework designed to enhance the capture of ESG-aspect-specific knowledge and language-specific knowledge present in the training data.

3.2 Chinese and Japanese

In Table 2, the approaches proposed by different teams for the Chinese dataset are outlined. Team 231 (Qiu et al., 2023) suggests converting Traditional Chinese to Simplified Chinese and employing the summarization of ChatGPT as input, rather than the entire news article. On the other hand, Team LIPI (Vardhan et al., 2023), translates the entire dataset into English and uses a T5-based model⁷ for paraphrasing the data. Meanwhile, Team FinNLU (Veeramani et al., 2023) integrates four embeddings from various language models, combined with TF-IDF and LSA features, for their predictions.

LIPI and FinNLU also joined the Japanese subtask with the proposed methods, and SPEvFT (Mishra, 2023) applies prompt engineering to Japanese articles.

4 Experimental Results

Tables 3, 4, 5, and 6 report the performances of ML-ESG-2 participants on English, French, Chinese, and Japanese datasets, respectively.

Drawing from the outcomes of both English and French datasets, it becomes evident that fine-tuning RoBERTa using these datasets produces the most optimal results, as documented in AnakItik (Winatmoko and Septiandri, 2023). When juxtaposed with the outcomes of fine-tuning Llama-2 (SPEvFT) (Winatmoko and Septiandri, 2023), there is a pronounced disparity in performance for English, though this difference is less marked for

Submission	Micro-F1	Macro-F1	Weighted-F1
AnakItik_English_2	0.9817	0.9548	0.9810
BrothFink_English_3	0.9771	0.9445	0.9765
NeverCareU_English_2	0.9633	0.9227	0.9648
FinNLU_English_1	0.9633	0.9180	0.9639
231_English_3	0.9633	0.9127	0.9627
SPEvFT_English_3 (Late)	0.9587	0.9118	0.9602
231_English_1	0.9633	0.9096	0.9620
231_English_2	0.9633	0.9096	0.9620
BrothFink_English_2	0.9541	0.8870	0.9525
BrothFink_English_1	0.9450	0.8645	0.9430
AnakItik_English_1	0.9220	0.8537	0.9289
LIPI_English_2	0.9312	0.8335	0.9294
NeverCareU_English_1	0.9312	0.8211	0.9267
LIPI_English_3	0.9266	0.8127	0.9226
HHU_English_1	0.9174	0.8098	0.9174
HHU_English_3	0.9174	0.8098	0.9174
LIPI_English_1	0.9083	0.7741	0.9051
AnakItik_English_3	0.9083	0.6246	0.9495
SPEvFT_English_1	0.9174	0.5574	0.9256
SPEvFT_English_2	0.8716	0.4657	0.8160
HHU_English_2	0.4908	0.4225	0.5719

Table 3: Results in English dataset.

Submission	Micro-F1	Macro-F1	Weighted-F1
SPEvFT_French_3 (Late)	0.8700	0.8661	0.8686
AnakItik_French_2	0.8550	0.8547	0.8554
AnakItik_French_1	0.8400	0.8368	0.8393
HHU_French_1	0.7550	0.7548	0.7555
LIPI_French_2	0.7550	0.7547	0.7556
HHU_French_3	0.7500	0.7457	0.7493
LIPI_French_3	0.7200	0.7182	0.7157
LIPI_French_1	0.7100	0.7090	0.7109
HHU_French_2	0.6250	0.6169	0.6231
AnakItik_French_3	0.7500	0.5545	0.8310
FinNLU_French_1	0.5500	0.5292	0.5184
SPEvFT_French_1	0.7100	0.4918	0.7367
SPEvFT_French_2	0.4450	0.3080	0.2741

Table 4: Results in French dataset.

French.

In the evaluation of the Chinese dataset, Team LIPI (Vardhan et al., 2023) achieved the highest performance during the formal evaluation. Their strategy employed all datasets available in ML-ESG-2 to facilitate data augmentation. The predictions from their model are generated using FinBERT with English inputs. A noteworthy observation from their study is the superior performance achieved using translated inputs compared to the original data. Subsequent to the formal evaluation, Team 231 (Qiu et al., 2023) embarked on further exploration. Their findings indicate that a simple conversion from Traditional Chinese to Simplified Chinese can enhance performance. Furthermore, they observed that utilizing a summary from ChatGPT yields better results than employing the complete news content. However, there were contrasting findings among the participants: while Team 231’s experiments suggest that TF-IDF fea-

⁷https://huggingface.co/humarin/chatgpt-paraphraser_on_T5_base

Submission	Micro-F1	Macro-F1	Weighted-F1
LIPI_Chinese_3	0.6859	0.5279	0.6773
LIPI_Chinese_2	0.7564	0.4585	0.7321
LIPI_Chinese_1	0.6731	0.2897	0.6508
231_Chinese_1	0.3718	0.1853	0.3725
231_Chinese_3	0.3654	0.1833	0.3593
231_Chinese_2	0.3590	0.1792	0.3593
FinNLU_Chinese_1	0.4103	0.1728	0.3881

Table 5: Results in Chinese dataset.

Submission	Micro-F1	Macro-F1	Weighted-F1
LIPI_Japanese_2	0.6889	0.6340	0.6786
LIPI_Japanese_1	0.6400	0.5436	0.6242
LIPI_Japanese_3	0.6222	0.5366	0.6033
SPEvFT_Japanese_1	0.4800	0.3792	0.4776
FinNLU_Japanese_1	0.5378	0.3043	0.4943

Table 6: Results in Japanese dataset.

tures offer limited utility, Team FinNLU’s results underscore their significance.

In the evaluation of the Japanese dataset, Team LIPI’s performance emerged superior compared to other participating teams. This notable advantage is attributed to their T5-based paraphrasing module, which demonstrably bolstered the system’s efficacy. Notably, despite the evident disparity in document genres — transitioning from news to annual securities reports — their methodology retained its effectiveness. This suggests that certain topics, especially those pertinent to “opportunity” (positive) and “risk” (negative) annotations, exhibit shared characteristics across different document genres and across languages to some degree. Team FinNLU adopted an early ensemble fusion approach, which proved to be equally effective for the Japanese context. The inclusion of LSA features also enhanced their performance metrics, mirroring the positive outcomes observed in other languages. On the other hand, Team SPEvFT explored a prompt engineering approach tailored to the Japanese language. They employed a sentence similarity method with training examples, revealing its efficacy, albeit to a limited extent.

5 Conclusion

In this study, we provided an overview of the ML-ESG-2 shared task featured at FinNLP@IJCAI-2023. Our findings indicate that for the English and French datasets with paragraph-based annotations, the top-performing models were fine-tuned RoBERTa or Llama-2. Conversely, for the Chinese and Japanese datasets that used article-based and sentence-based structures, translating the content

into English yielded commendable outcomes. Such findings bring forth pivotal questions. Our investigation into baseline models suggests that their adeptness, even in their simplicity, might surpass certain specialized methods. Additionally, it hints at the possibility that massive datasets might not be essential for fine-tuning (large) language models in ESG impact type identification, particularly for languages like English and French.

Having delved into ESG issue identification and ESG impact type identification tasks, our next venture will be into ESG impact duration inference under ML-ESG-3. Through the series of ML-ESG shared tasks, we pave the way for a holistic approach to dynamic ESG ratings based on news articles.

Acknowledgments

This research is supported by National Science and Technology Council, Taiwan, under grants 110-2221-E-002-128-MY3, 110-2634-F-002-050-, and 111-2634-F-002-023-. 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).

References

- Fabian Billert and Stefan Conrad. 2023. Exploring knowledge composition for esg impact type determination. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.
- 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)*.
- 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 (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*.
- Soumya Mishra. 2023. Predicting esg impact types of multi-lingual news articles: Leveraging strategic

prompt engineering and llm fine-tuning. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Anna Polyanskaya and Lucas Fernández Brillet. 2023. Gpt-based solution for esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Le Qiu, Bo Peng, Jinghang Gu, Yu-Yin Hsu, and Emanuele Chersoni. 2023. Identifying esg impact with key information. In *Proceedings of the Sixth Workshop on Financial Technology 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)*.

Harsha Vardhan, Sohom Ghosh, Ponnurangam Kumaraguru, and Sudip Naskar. 2023. A low resource framework for multi-lingual esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023. Enhancing esg impact type identification through early fusion and multilingual models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Yosef Ardhito Winatmoko and Ali Septiandri. 2023. The risk and opportunity of data augmentation and translation for esg news impact identification with language models. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing*.

Identifying ESG Impact with Key Information

QIU Le, PENG Bo, GU Jinghang, HSU Yu-Yin and Emmanuele CHERSONI

The Hong Kong Polytechnic University

11 Yuk Choi Rd, Hung Hom, Hong Kong SAR

lani.qiu@connect.polyu.hk

{peng-bo.peng, jinghang.gu, yu-yin.hsu, emmanuele.chersoni}@polyu.edu.hk

Abstract

This paper presents a concise summary of our work for the ML-ESG-2 shared task, exclusively on the Chinese and English datasets. ML-ESG-2 aims to ascertain the influence of news articles on corporations, specifically from an ESG perspective. To this end, we generally explored the capability of key information for impact identification and experimented with various techniques at different levels. For instance, we attempted to incorporate important information at the word level with TF-IDF, at the sentence level with TextRank, and at the document level with summarization. The final results reveal that the one using summarization yields the best predictions.

1 Introduction

Environmental, Social, and Governance (ESG) factors have been deemed essential for a company's prosperity in the long run and emerged as a crucial consideration for investment and corporate operations (Tseng et al., 2023; Kannan and Seki, 2023). Spontaneously, ESG has garnered increased attention among the FinNLP community. In 2023 FinNLP, in conjunction with IJCAI, has proposed a shared task of Multi-Lingual ESG Impact Type Identification (ML-ESG-2), releasing a multi-lingual dataset that consists of news articles in four languages — (traditional) Chinese, English, French, and Japanese (Tseng et al., 2023). The objective is to determine if the given news is an opportunity or a risk for the company from the ESG aspect.

ML-ESG-2 presents itself as a text classification problem, which involves extracting features from raw textual data and predicting categories based on such features. Research around this topic in recent years centers on the attention mechanism, among others (see Li et al., 2022). In particular, Transformer models such as BERT (Devlin et al., 2018) are widely exploited, and further encourage the

trend of using more data and large language models for text classification tasks (Minaee et al., 2021). In the case of long document classification where regular Transformers fail, more effective methods have been proposed, mostly involving pre-training another language model for long sequences or extracting key information to feed into the model. For example, Beltagy et al. (2020) revised the attention mechanism in BERT and developed a Longformer that increases the input capacity up to 4, 096 tokens, and Ding et al. (2020a) proposed CogLTX that jointly trains two BERT or RoBERTa (Liu et al., 2019) models - one for key sentence extraction and the other for the final task. However, a survey by Park et al. (2022) suggests that complicated approaches don't necessarily bring better outcomes, meanwhile demanding more investment (e.g. Longformer requires more GPU memories, and CogLTX costs much more runtime). Inspired by such findings, we also used pre-trained language models (PLMs) for the ESG task. Specifically, we also exploited the ChatGPT series as a translation engine for data augmentation and to discern the important information for long document classification.

2 Related Work

In the last decade, text classification tasks have gradually embraced the deep learning approach, as it relieves the burden of feature designing. Multi-layer perceptions (Khalil Alsmadi et al., 2009) already outperform traditional models such as Naive Bayes, SVM, etc., CNN (convolutional neural network) and RNN (recurrent neural network) further advance the performance in this area (Li et al., 2022). The GNN (graph neural network) also takes a place but focuses on modeling the structural information within the text (Li et al., 2022; Liu et al., 2022). The introduction of BERT (Devlin et al., 2018) has especially promoted the fashion of ap-

plying PLMs in text classification tasks. Compared with previous methods such as TF-IDF (Rajaraman and Ullman, 2011) and Word2Vec (Mikolov et al., 2013), PLMs capture more effective representations and boost performance text classification tasks.

However, BERT and its variants such as RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019), etc., are intrinsically incapable of processing long sequences, and a brutal truncation does not necessarily provide benefits. The predicament sees the appearance of more PLMs tailored for long sequences. Attention-based models like Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2020) employ sparse self-attention instead of full attention as in the BERT series and expand the input capacity up to 4, 096 tokens. Hierarchical Transformers such as ToBERT (Pappagari et al., 2019) produce chunk-level representations and thus can take input of any length. ERNIE-DOC (Ding et al., 2020b) enhances the recurrence mechanism as employed in Transformer-XL (Dai et al., 2019) and XLNet (Yang et al., 2019), etc. and introduces a retrospective feed mechanism to directly model the text at the document level. Another type of approach aims at selecting important sentences from the document for classification, e.g. CogLTX (Ding et al., 2020a) and more traditional approaches such as TextRank Mihalcea and Tarau (2004). Techniques utilizing summarization for classification also fall within this category (e.g. Basha et al., 2019).

It should be noted that sophisticated models such as those described above do not guarantee better performances, as a regular Transformer model may surpass them with simple augmentation (see Li et al., 2022). Sparse attention cannot fully exploit the global information for each segment when modeling long documents; the recurrence mechanism introduces latency (Mamakas et al., 2022), and hierarchical Transformers have the problem of *context fragmentation* (Ding et al., 2020b).

3 Methods

As evident in Table 1, the Chinese track is a multi-class long document classification task, while the English track is a binary classification task. Also, a severe imbalance can be observed in the Chinese dataset, with the "Opportunity" and "ESG but not company related" samples occupying about 90% of the entire set.

Train set	Class distribution (0: 1: 2: 3: 4)	Text length (avg.) ¹
Chinese	536: 58: 23: 593: 50	1349.88
English	694: 114: 0: 0: 0	412.48

Table 1: Data statistics of the training sets. For reference: 0 = "Opportunity", 1 = "Risk", 2 = "Cannot Distinguish (company related)", 3 = "ESG but not company related", 4 = "Non-ESG".

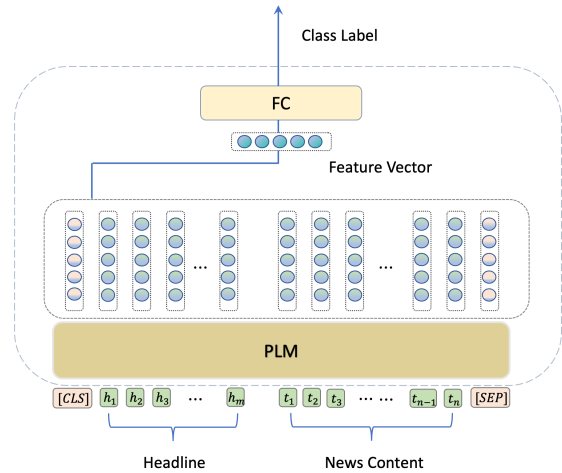


Figure 1: Model architecture

Headline	Content
Fed Launches Climate Risk Exercise for Big Banks	The U.S. Federal Reserve Board released details and instructions for its inaugural climate scenario analysis exercise for the six largest U.S. banks, designed to assess the banks' climate-related risk management practices and their resilience to a range of climate outcomes. Results from the exercise are to be submitted by the banks by the end of July.

Figure 2: An example of the English data

For this task, we adopted a vanilla architecture as shown in Figure 1. The underlying idea of our method is to solely utilize text representations as features for classification. As shown in Figure 2, each sample provides a headline alongside the content. Seeing that headlines are exploited for news classification (see Rana et al., 2014), we also include them in our method. In the Chinese task, particularly, we replaced the original news content with a summarized text.

For the English task, we managed to expand the dataset by including more samples translated from the French data considering the original one is rather small and only contains a training set initially. We chose the French data instead of others

¹The value may vary by a small margin due to the pre-processing methods.



Figure 3: An example of the GPT-4 summary

because the French task is also a binary classification one, and the fact that the two languages share a majority of vocabulary could ease the translation process. For the Chinese task, we converted the text from traditional Chinese to simplified Chinese. In this case, we utilized GPT-3.5 (i.e., ChatGPT) for translation and GPT-4 (OpenAI, 2023) for summarization (refer to Figure 3 for an example). In terms of PLM, we used *roberta-large* (Liu et al., 2019) for English, and both *bert-base-chinese* (Devlin et al., 2018) and *chinese-roberta-wwm-ext* (Cui et al., 2019) for Chinese. The input sequence length is set to 512 tokens. We applied over-sampling and under-sampling strategies during training to alleviate the data imbalanced problem. For better outcomes, we adopted an ensemble learning strategy in the final submission. Specifically, we aggregated the results of several models (three for each submission and six in total, to be precise) based on hard voting.

The method was justified with ablation experiments that will be presented in the following section. We explored the contributions of different components or pre-processing techniques, especially on the Chinese task. To be specific, we started with a regular truncation, which is inevitable considering that the Chinese data consists of long sequences, then an irregular truncation that involves assembling sentences roughly extracted from the beginning, middle, and end of the text (referred to as a *sandwich* text hereinafter), and a key sentence selection that is implemented with the

TextRank algorithm (Mihalcea and Tarau, 2004). Besides the headlines mentioned earlier, we tried to concatenate important words extracted with the TF-IDF metric in the input (Rajaraman and Ullman, 2011), in an attempt to incorporate more information at a lexical level. Additionally, to further attend to the imbalance issue, we also involved the focal loss function (Lin et al., 2017) in our experiments. In short, the focal loss works by decreasing the loss contribution of easy cases and forcing the model to focus on the hard cases. That gives it the potential to address the imbalance issue. Previous studies (e.g. Liu et al., 2021; Nan et al., 2021) also confirmed its positive influence on NLP tasks.

4 Results and Discussion

Task	Model	Micro F1	Macro F1	Weighted F1
En.	<i>AnakItik's</i> ²	0.9817	0.9548	0.9810
	<i>BrothFink's</i>	0.9771	0.9445	0.9765
	<i>NeverCareU's</i>	0.9633	0.9227	0.9648
	Ours	0.9633	0.9127	0.9627
	Ours	0.9633	0.9096	0.9620
Ch.	<i>LIPi's</i>	0.6859	0.5279	0.6773
	<i>LIPi's</i>	0.7564	0.4585	0.7321
	<i>LIPi's</i>	0.6731	0.2897	0.6508
	Ours	0.8654	0.7325	0.8686
	Ours	0.8846	0.7245	0.8856
		0.8782	0.6770	0.8745

Table 2: Evaluation scores of submitted results on both tracks.

Method	Micro F1	Macro F1	Weighted F1
bbc + CE	0.8333	0.7237	0.8379
wwm + CE	0.8718	0.7027	0.8617
wwm + CE	0.8526	0.6949	0.8522
bbc + CE	0.8141	0.6780	0.8185
wwm + CE	0.9038	0.6618	0.8970
wwm + FL	0.8590	0.6142	0.8508

Table 3: Performance of our submitted results without ensemble learning on the Chinese track. For reference, bbc = bert-base-chinese, wwm = chinese-roberta-wwm-ext, CE = Cross-Entropy loss, FL = Focal loss.

Table 2 presents the F1 scores of the top models on the leaderboard and of our three outputs. Note that we aggregated the predictions of three models via hard voting for submission. Evidently,

²The name of the team, the same below.

Method	Micro F1	Macro F1	Weighted F1
bbc, CE	0.8787	0.7393	0.8759
bbc, FL	0.8929	0.7398	0.8919
wwm, CE	0.8786	0.7383	0.8792
wwm, FL	0.9071	0.7519	0.9041

Table 4: Performance of models with cross-entropy and focal loss. For comparison, we used the same setup for both experiments. Best F1 scores are reported within 5 epochs.

Features	Micro F1	Macro F1	Weighted F1
content, tra	0.8214	0.6223	0.8329
headline + content, tra	0.8429	0.6385	0.8501
headline + content, sim	0.8643	0.7129	0.8633
headline + <i>sandwiched</i> content, sim	0.8571	0.6115	0.8496
headline + key content, sim	0.9000	0.7485	0.8970
headline + summary, sim (the proposed method)	0.9143	0.7624	0.9093
headline + summary + tf-idf words, sim	0.9071	0.7591	0.9024

Table 5: Performance of models with different features on the Chinese dev set. For reference, tra = traditional Chinese, sim = simplified Chinese. For demonstration, we used the same PLM — chinese-roberta-wwm-ext and reported the best F1 score within 5 epochs.

the scores on the English set including ours have achieved a high level in general, which can be expected considering that the English task is relatively simple. The tally on the Chinese task, on the other hand, shows that our models outperform the others by a notable margin. The models without ensemble learning (see Table 3 for their F1 scores) also appear to be competitive. Although the model with a focal loss, which is expected to yield improvement, ends up with the lowest scores in submission, the contribution of the function has been confirmed with experiments as shown in Table 4.

Regarding the Chinese task, we also investigated other possibilities and reported their evaluation results on the dev set in Table 5, which justified our method. The experiments with the original headline and the news content in traditional Chinese set the baselines. Additionally, we managed to incorporate other information in an attempt to further advance the performance. The results reveal that the sentences extracted via TextRank (Mihalcea and Tarau, 2004) and words extracted via TF-IDF

(Rajaraman and Ullman, 2011) have positive influences. The method with the summarized content genuinely boosts the performance. Nevertheless, the takeaway from these experiments could be that the key information, in this case including the headlines (which in some sense foretell or summarize the article), the keywords, or the summary (which explains all the information in an effective and precise way) plays a crucial role in the ESG impact identification task.

5 Conclusion

To recap, we employed a simple architecture for the ML-ESG-2 shared task on ESG impact type identification and ended up with a fair result. Particularly, we employed a summarising technique to address the document classification problems as in the Chinese track with the widely popular AI bot — GPT-4 (OpenAI, 2023) as a text summarizer. Note that the summarization-based approach is a consequence of multiple experiments. Before settling down on summarization, we investigated the influences of other components including news headlines, key sentences, and words. The results reveal that the key formation as such is useful for text classification. Our method turns out to be effective in that GPT-4 captures the essential meaning of the texts.

However, we failed to compare the summarization performance of GPT-4 and other possible methods, nor did we examine other approaches to keywords or key sentence extraction besides TextRank (Mihalcea and Tarau, 2004) and TF-IDF (Rajaraman and Ullman, 2011). The evaluation results show that there is still room for improvement in the Chinese task. A further and deeper investigation could produce some more sparkles and lead to more interesting findings.

References

- S Rahamat Basha, J Keziya Rani, and JJCP Yadav. 2019. A novel summarization-based approach for feature reduction enhancing text classification accuracy. *Engineering, Technology & Applied Science Research*, 9(6):5001–5005.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

- Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.
- 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*.
- Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020a. CogLtx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.
- Siyu Ding, Junyuan Shang, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020b. Ernie-doc: A retrospective long-document modeling transformer. *arXiv preprint arXiv:2012.15688*.
- Naoki Kannan and Yohei Seki. 2023. Textual evidence extraction for esg scores. *Proceedings of the Joint Workshop of the 5th Financial Technology and Natural Language Processing (FinNLP)*.
- Mutasem Khalil Alsmadi, Khairuddin Bin Omar, Shahrul Azman Noah, and Ibrahim Almarashdah. 2009. Performance comparison of multi-layer perceptron (back propagation, delta rule and perceptron) algorithms in neural networks. In *2009 IEEE International Advance Computing Conference*, pages 296–299. IEEE.
- 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*.
- 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 Transactions on Intelligent Systems and Technology (TIST)*, 13(2):1–41.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Jianyi Liu, Xi Duan, Ru Zhang, Youqiang Sun, Lei Guan, and Bingjie Lin. 2021. Relation classification via bert with piecewise convolution and focal loss. *Plos one*, 16(9):e0257092.
- Tengfei Liu, Yongli Hu, Boyue Wang, Yanfeng Sun, Junbin Gao, and Baocai Yin. 2022. Hierarchical graph convolutional networks for structured long document classification. *IEEE Transactions on Neural Networks and Learning Systems*.
- 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*.
- 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*.
- Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning-based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.
- Fulai Nan, Jin Wang, and Xuejie Zhang. 2021. Mirror distillation model with focal loss for chinese machine reading comprehension. In *2021 International Conference on Asian Language Processing (IALP)*, pages 7–12. IEEE.
- OpenAI. 2023. [Gpt-4 technical report](#). *ArXiv*, abs/2303.08774.
- Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 838–844. IEEE.
- Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *arXiv preprint arXiv:2203.11258*.
- Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.
- Mazhar Iqbal Rana, Shehzad Khalid, and Muhammad Usman Akbar. 2014. News classification based on their headlines: A review. In *17th IEEE International Multi Topic Conference 2014*, pages 211–216. IEEE.
- Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Dynamicsesg: A dataset for dynamically unearthing esg ratings from news articles. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*.
- 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.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

A low resource framework for Multi-lingual ESG Impact Type Identification

N Harsha Vardhan*, Sohom Ghosh†, Ponnurangam Kumaraguru*, Sudip Kumar Naskar†

*International Institute of Information Technology, Hyderabad, India

†Jadavpur University, Kolkata, India

nemani.v@research.iiit.ac.in, sohom1ghosh@gmail.com

pk.guru@iiit.ac.in, sudip.naskar@gmail.com

Abstract

With the growing interest in Green Investing, Environmental, Social, and Governance (ESG) factors related to Institutions and financial entities has become extremely important for investors. While the classification of potential ESG factors is an important issue, identifying whether the factors positively or negatively impact the Institution is also a key aspect to consider while making evaluations for ESG scores. This paper presents our solution to identify ESG impact types in four languages (English, Chinese, Japanese, French) released as shared tasks during the FinNLP workshop at the IJCNLP-AAACL-2023 conference. We use a combination of translation, masked language modeling, paraphrasing, and classification to solve this problem and use a generalized pipeline that performs well across all four languages. Our team ranked 1st in the Chinese and Japanese sub-tasks.

1 Introduction

In recent times, the focus on Institutions' Environmental, Social, and Governance factors (ESG) has garnered increased interest from the global investment and corporate governance communities. People have also grown to be socially responsible and environmentally conscious while investing. ESG serves as a third dimension beyond risk and return. Research also indicates that Institutions with better ESG performance directly correlate to better stock performance and risk management (Whelan and Atz, 2021). Keeping this in mind, many rating agencies quantify the nature and impact of ESG aspects of an institution and publish ratings (Serafeim and Yoon, 2022). Apart from ESG investing, Impact investing (Berk and van Binsbergen, 2021) has also gained traction where investors, instead of investing solely based on ESG benefits, would look for a combination of better returns as well as a positive influence in society. Hence, impact identification is crucial to determine whether statements

are an opportunity or a risk for the Institution.

Most of the scoring processes involved in ESG and Impact assessments are extremely time-consuming and require expert involvement and manual annotations. To automate this, we propose a generalized pipeline capable of predicting the impact types of ESG-related news articles (as shown in Figure 1). This generalized pipeline can be scaled to other low-resource datasets as well.

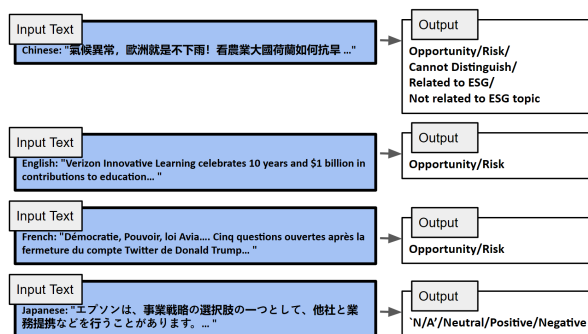


Figure 1: The Multilingual ESG Impact Assessment Task.

The labels primarily indicate if the given news is an opportunity or a risk from the ESG aspect. In this shared task, we participated in all four languages and were ranked 1st in Chinese and Japanese sub-tasks, 4th in French, and 7th in English.

2 Related Work

With the advent of green investing, many approaches and models have been developed to automate processes in Financial and ESG-based NLP research, including the development of models like FinBERT (Araci, 2019), ESGBERT (Mehra et al., 2022), etc. While there has been much work on ESG-type classifications, including on multilingual datasets, more work needs to be done on impact-type classifications. FinNLP 2023 (Chen et al., 2023) focuses on a similar task where participants

were required to classify multilingual data into the ESG issue type, where the best results were obtained by using language-specific BERT models along with data augmentations using Large Language Models. Furthermore, it’s important to note that extensive research has been conducted on sentiment analysis (Pasch and Ehnes, 2022; Aue et al., 2022), which can be considered a fundamental aspect of impact identification. Attempts have also been made for impact identification in Chinese (Tseng et al., 2023) and Japanese (Kannan and Seki, 2023).

3 Task Description

The task is primarily a classification task where given a text, classify whether the text poses a risk or an opportunity for the company. As shown in Figure 1, there are multiple languages with differences in classes.

4 Data

The dataset primarily contains news articles collected from four different languages, English (en), Chinese (zh), Japanese (ja), and French (fr), along with their impact types.

Language	Train	Test	C	W_c	W_h
English	808	218	2	412.48	76.83
Chinese	1400	156	5	-	33.68
Japanese	896	225	4	-	78.82
French	818	200	2	564.88	96.17

Table 1: Metrics across languages. C denotes the number of Classes, W_c denotes the average character length of content and W_h denotes the average character length of headline. Chinese and Japanese datasets do not have content columns.

Given that the dataset across languages is small, and the classwise distribution is highly skewed. To overcome these challenges, we use a combination of translation and data paraphrasing on minority classes.

5 Approaches

We primarily used encoder-based models for this classification task. Given the limited sample size of the dataset, variations in languages, and disparity with class distribution among different languages, We tried to make a pipeline that accounted for such differences and performed consistently well across all languages. We tried a variety of approaches like

Masked Language Modelling (MLM), Paraphrasing for augmenting the minority classes, Translation, and Multilingual Models and used a combination to finalize our pipeline based on empirical experiments.

All of the experiments have been run using a batch size of 32, a learning rate of $2e^{-5}$, weight decay of 0.01, and for ten epochs. The reported metrics are based on 80 : 20 train-test set splits with a constant random seed and not on the final validation sets used for the leaderboard. The code, data, and models used for inferences are available at the link.

5.1 Masked Language Modelling

We performed several experiments to decide the necessary models for classification. Also, we experimented with pre-training the models beforehand on the ESG corpus, which was the English dataset for the Multi-Lingual ESG Issue Identification (ML-ESG) (Chen et al., 2023) and then using the fine-tuned models for classifications. We noticed that across all languages, the models pre-trained on the ESG corpus and then fine-tuned for classification outperformed those fine-tuned for classification.

Approach	Title	Content
Classification	74.89%	92.48%
MLM + Classification	85.48%	93.16%

Table 2: Comparison across Classification and MLM + Classification approaches along with news headlines and content using bert-base-cased (Devlin et al., 2019) model. These reported numbers are the weighted $F1$ with the English dataset.

From Table 2, we also observe that using news content for training over title performs better. The French dataset exhibits similar trends, and hence, for all further analysis, we use the news content for English and French and the news title for Chinese and Japanese since they do not have news content available in the dataset.

5.2 Translation and Multilingual Models

We have also experimented with specific language models vs. translating and English-based models primarily due to a larger number of specialized models pre-trained on ESG data being available in English. We used Google Translate to translate data from French, Chinese, and Japanese and leveraged this data as additional data while training for

models. Also, by using English, we were able to use paraphrasing tools to augment and extend the minority classes of our dataset.

Approach	F1
Translated	68.92%
Chinese	68.45%

Table 3: Comparison of weighted F1 scores while using translated Chinese to train a bert-base-cased model vs. using Chinese data to train a bert-base-multilingual-cased model (Devlin et al., 2019).

While the disparity between the translated text and the original language may not seem substantial, there exists a possibility that employing more specialized language models tailored to the Chinese language could have potentially delivered better results. However, this approach would have restricted our ability to employ paraphrasing-based techniques, as such tools are not as readily available in non-English languages. Furthermore, it would have limited our access to English models predominantly trained on ESG data. Accordingly, our primary strategy revolved around using translated text for classification.

5.3 Paraphrasing for Data Augmentation

Given that the dataset across languages is small and the classwise distribution is highly skewed, one of the approaches we considered for improving the classification task is to augment the minority classes and extend the dataset. While rule-based paraphrasers are popular and widely used for such tasks, the variation within sentences is frequently minor and only offers a slight improvement during training. Hence, we considered a T5-based paraphraser (Vladimir Vorobev, 2023), primarily fine-tuned on ChatGPT paraphrases. It offers a better range of sentence variations than any other approaches tried. We first translated the dataset from the respective languages to English and then generated paraphrased data on minority class data (For each minority instance, approximately 3-4 paraphrases were created, depending on the specific count of instances for that particular label. For the same reason we did not paraphrase for french language since the label distribution was already uniform. The paraphrased data can be accessed [here](#).) and used this along with the original data for training the classification model.

We observe that across languages, paraphrased data improved the F1 metrics of models to a great

Approach	F1 (en)	F1 (zh)
Paraphrased Data	98.91%	84.98%
Original Data	93.16%	68.45%

Table 4: Comparison of weighted F1 while using paraphrased text vs. original dataset for MLM + Classification on the English dataset and The original dataset for Chinese and the translated + Paraphrased version of the Chinese dataset. bert-base-cased model was used for English and bert-base-multilingual-cased for Chinese.

extent. This effect was more prominent in Chinese and Japanese datasets, where the number of classes was more prominent, and there was a wider class disparity. This supports our choice of using translated text rather than the original despite lackluster results while just translating and using that data for classification.

6 Final System Description

For the final system that was used, based on the empirical studies performed above, We used a pipeline that initially translated all of the given text into English using Google Translate. Then we use the T5-based paraphraser (Vladimir Vorobev, 2023) to generate new minority class instances. We also use an ESG corpus to initially pre-train a model on this corpus and then fine-tune it for classification on the translated and augmented dataset. Figure 2 shows the exact process.

We also performed more experiments to decide which models best performed on the English dataset and chose bert-base-cased (Devlin et al., 2019), Finbert (Araci, 2019), and finbertone (Huang et al., 2023). We used the same models for the other languages as well. The model hyper-parameters are the same as mentioned in the methodology.

We observe that despite using a generalized pipeline and models for all the languages, the results are good. Table 5 shows the performance of models for all of the languages and models used.

Models	F1(en)	F1 (zh)	F1 (ja)	F1 (fr)
BBC	98.91%	84.98%	89.64%	78.25%
FB	97.82%	85.31%	91.13%	78.55%
FBT	98.91%	82.26%	89.54%	70.79%

Table 5: Final weighted F1 metrics for the models used for submission. BBC = bert-base-cased, FB = FinBERT, FBT = FinBERT Tone

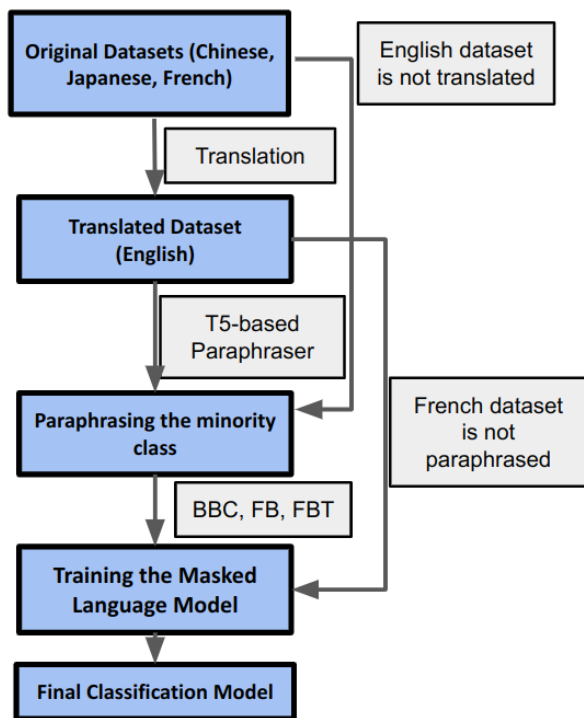


Figure 2: The final system pipeline. BBC = bert-base-cased, FB = FinBERT, FBT = FinBERT Tone

7 Conclusion

Comparing the performance of our models with that of other participants, we conclude that our models performed consistently well. We outperformed all other teams in the Chinese and Japanese sub-tasks. One unique feature is despite four different languages, we were able to use the same pipeline and same set of models and achieve consistently good results across languages, which leads us to believe that the pipeline is performant for low resource settings. All of the data generated and code used can be accessed [here](#).

Limitations

The primary challenge highlighted in the paper’s approaches is the translation process. While it expands the possibilities, it also comes with a drawback - the loss of language-specific nuances and information. Integrating language-specific paraphrasing tools and access to Environmental, Social, and Governance (ESG) datasets tailored to those languages could enable us to adapt the existing pipeline. This adaptation would involve incorporating regional language models instead of relying solely on English models, potentially resulting in improved performance.

We also did not evaluate larger models due to time and feasibility constraints, but larger models would have provided better results. Also, since the number of classes differed across languages, training a singular multilingual model or similar approaches resulted in poor metrics for some languages. Hence we did not pursue this direction.

One of the initial choices for selecting news content as the primary choice for the classification approach could also have been flawed. Since headlines are generally more read and captivating, it might have provided a polarized view of the instance and might have been easier to categorize as an Opportunity or Risk. Ideally, some form of ensemble modeling between headlines and content might improve the performance of the present approach.

Ethics Statement

In conducting this research, we have not encountered any significant ethical concerns or considerations that would require special attention in this paper. Our study focuses on impact type classification of ESG-related publically available news instances, and the data and methods employed adhere to established ethical guidelines and standards within the field of computational linguistics.

References

- Dogu Araci. 2019. [Finbert: Financial sentiment analysis with pre-trained language models](#).
- Tanja Aue, Adam Jatowt, and Michael Färber. 2022. [Predicting companies’ esg ratings from news articles using multivariate timeseries analysis](#).
- Jonathan B. Berk and Jules H. van Binsbergen. 2021. [The Impact of Impact Investing](#). Research Papers 3981, Stanford University, Graduate School of Business.
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. Overview of the FinNLP-2023 ML-ESG task: Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- 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.

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 (FinNLP)*.

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*. Academy and Industry Research Collaboration Center (AIRCC).

Stefan Pasch and Daniel Ehnes. 2022. **Nlp for responsible finance: Fine-tuning transformer-based models for esg**. pages 3532–3536.

George Serafeim and Aaron Yoon. 2022. **Stock price reactions to esg news: the role of esg ratings and disagreement**. *Review of Accounting Studies*, 28.

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. **Dynamicsesg: A dataset for dynamically unearthing esg ratings from news articles**. In *Proceedings of the 32nd ACM International Conference on Information & Knowledge Management, CIKM ’23*, New York, NY, USA. Association for Computing Machinery.

Maxim Kuznetsov Vladimir Vorobev. 2023. A paraphrasing model based on chatgpt paraphrases.

Tensie Whelan and Ulrich Atz. 2021. **Esg and financial performance : Uncovering the relationship by aggregating evidence from 1 , 000 plus studies published between 2015 – 2020**.

A Appendix

In this section, we present the classwise F1 metric for models based on the language. The models in consideration are:

- bert-base-cased (BBC)
- Finbert (FB)
- Finbert-Tone (FBT)

• English

Models	Label 0	Label 1
BBC (LIP1)	99%	99%
FB (LIPI2)	98%	98%
FBT (LIPI3)	99%	99%
Support	146	130

Table 6: English Language Model Metrics

Note: For English, Label 0 denotes Opportunity and Label 1 signifies Risk.

• French

Note: Label 0 stands for Opportunity and Label 1 represents Risk. The Support is 20% of the training set as the French dataset has an almost equal class distribution.

Models	Label 0	Label 1
BBC (LIP1)	82%	74%
FB (LIPI2)	81%	76%
FBT (LIPI3)	76%	65%
Support	88	76

Table 7: French Language Model Metrics

• Chinese

Models	0	1	2	3	4
BBC (LIP1)	85%	95%	75%	84%	83%
FB (LIPI2)	84%	93%	86%	83%	86%
FBT (LIPI3)	80%	89%	88%	80%	83%
Support	132	58	29	122	55

Table 8: Chinese Language Model Metrics

Note: In Chinese, Label 0 is Opportunity, 1 is Risk, 2 is Cannot Distinguish, 3 is Related to ESG but unrelated to the company, and 4 is Not Related.

• Japanese

Models	Label 0	Label 1	Label 2	Label 3
BBC (LIP1)	88%	86%	92%	97%
FB (LIPI2)	91%	88%	91%	96%
FBT (LIPI3)	90%	85%	88%	97%
Support	86	69	57	49

Table 9: Japanese Language Model Metrics

Note: For Japanese, Label 0 is Positive, 1 is "Not Available", 2 is Neutral, and 3 is Negative.

All metrics are derived from the paraphrased testset, which forms part of the publicly accessible training set. For further details on the training data, refer to [this repository](#).

GPT-based Solution for ESG Impact Type Identification

Anna Polyanskaya
StockFink, UPV-EHU
annap@stockfink.com

Lucas Fernández Brillet
StockFink
lucasfb@stockfink.com

Abstract

In this paper, we present our solutions to the ML-ESG-2 shared task which is co-located with the FinNLP workshop at IJCNLP-AACL-2023. The task proposes an objective of binary classification of ESG-related news based on what type of impact they can have on a company - Risk or Opportunity. We report the results of three systems, which ranked 2nd, 9th, and 10th in the final leaderboard for the English language, with the best solution achieving over 0.97 in F1 score.

1 Introduction

In an era characterized by increasing environmental, social, and governance (ESG) awareness, investors are becoming increasingly conscious of such issues, and companies' ESG performance affects their financial performance (Naeem et al., 2022). It has been shown that ESG-related news have become a significant driver of market volatility, as both good and bad news can have a considerable impact (Sabbaghi, 2022; Wong and Zhang, 2022).

2 Related work

Following the trend for ESG awareness, natural language processing (NLP) is commonly used to analyze texts, usually reports and news, related to these types of issues. New tools and data sets are being developed, going further than just being specific to the financial domain, e.g. ESG-BERT (Mukherjee, 2020) and DynamicESG (Tseng et al., 2023). Last year's Shared Task focused on ESG Taxonomy Enrichment and Sustainable Sentence Prediction (Kang and El Maarouf, 2022).

Li et al. (2023) shows that GPT-based models, while producing impressive results, still fall behind domain-specific large language models (LLMs), such as FinBERT (Araci, 2019). In this work, we wanted to test this hypothesis, as the proposed task differs from sentiment analysis in the complexity of the relation between a given text and its label. We believe that GPT models can reason better in this

task as they seem to be able to utilize the context (or factual knowledge) they already have about the world, thus being able to grasp complicated causation even in zero- and few-shot settings (Radford et al., 2019; Brown et al., 2020; Liu et al., 2023).

3 Data

The data was provided by the organizers, see (Chen et al., 2023). The English training data set contained 808 entries, each entry consisting of *URL*, *News Title*, *News Content*, and *Impact Type* (class).

The classes were quite imbalanced, so for our first system's training we also used an additional 360 entries of the Risk class from the French data set, automatically translated to English using DeepL Python Library¹. The dataset statistics are presented in Table 1.

Set	Risk	Opportunity	Total
Train	91	555	646
Train + Fr	451	555	1006
Dev	23	139	162
Test	27	191	218

Table 1: Number of entries in each data subset.

4 Methodology

4.1 System I: FinBERT

For our first submission, we used a pre-trained FinBERT² model and fine-tuned it for the binary classification with the Train and Train+Fr sets. We trained for 5 epochs with F1 being the metric for choosing and loading the best model. The scores on the development set are presented in Table 2.

As we can see, the addition of the translated French data helped improve the precision for the Risk class and the overall results. This model, trained on the combined set was used to produce the final submission #1.

¹<https://pypi.org/project/deepl/>

²<https://huggingface.co/ProsusAI/finbert>

System:
 You are an expert in the financial market, helping a client understand the impact of ESG-related news on the market. Given a section of a recent ESG news article, which will be provided to you by the user, decide whether it presents Opportunity or Risk, described respectively as follows:
 Opportunity: An event, whether good or bad, that could yield positive returns for ESG-related issues.
 Risk: An event or statement, whether good or bad, that could yield negative returns or threaten positive returns for ESG-related issues.
 Reply with only one word (Opportunity or Risk). Don't explain your answers.

User:
 <NEWS CONTENT>

Figure 1: The message structure sent via API.

Train set	Class	Precision	Recall	F1
Train	Opportunity	.98	.91	.94
	Risk	.62	.87	.73
	weighted avg.	.93	.91	.91
Train+Fr	Opportunity	.96	.98	.97
	Risk	.86	.78	.82
	weighted avg.	.95	.95	.95

Table 2: Dev scores of the fine-tuned FinBERT model.

4.2 System II: Zero-Shot GPT

For our second submission, we explored the zero-shot capabilities of GPT-3.5³ for this type of task. We used the *gpt-3.5-turbo* model with a temperature of 0.1 via the OpenAI's API. The final prompt design is shown in Figure 1.

For consistency purposes, we scored this approach on the same development subset. We also evaluated this approach using the Train subset (English only) to get a better picture of the model's zero-shot capabilities. The results are shown in Table 3. This approach was used to produce the final submission #2.

Set	Class	Precision	Recall	F1
Dev	Opportunity	.96	.99	.98
	Risk	.90	.78	.84
	weighted avg.	.96	.96	.96
Train	Opportunity	.95	.99	.97
	Risk	.89	.68	.77
	weighted avg.	.94	.94	.94

Table 3: Dev and Train scores of the *gpt-3.5-turbo* model (zero-shot classification).

³<https://platform.openai.com/docs/models/gpt-3-5>

4.3 System III: Few-Shot GPT

Seeing that GPT-based models are capable of producing high-quality results in a zero-shot setting, we wanted to explore if they can be further improved by using a few-shot approach. We used the same prompt and parameters, but before asking the model to produce the result for a given text, we added 12 random news (6 for each class) as examples. Thus, the message sequence was as shown in Figure 2.

System: <PROMPT>, see Figure 1
 User: <NEWS CONTENT>
 Assistant: Risk
 User: <NEWS CONTENT>
 Assistant: Opportunity

} × 6

Figure 2: The structure of the messages.

The Dev scores for this setup are shown in Table 4.

Set	Class	Precision	Recall	F1
Dev	Opportunity	.98	.98	.98
	Risk	.87	.87	.87
	weighted avg.	.96	.96	.96
Train	Opportunity	.97	.98	.98
	Risk	.87	.80	.83
	weighted avg.	.96	.96	.96

Table 4: Dev and Train (except entries used as examples) scores of the *gpt-3.5-turbo* model (few-shot classification).

System	Micro-F1	Macro-F1	Weighted-F1	Rank
fine-tuned FinBERT	.9450	.8645	.9430	10
zero-shot GPT	.9541	.8870	.9525	9
few-shot GPT	.9771	.9445	.9765	2

Table 5: Test scores of our systems, provided by the organizers, with ranks among 21 other submissions.

System	Class	Precision	Recall	F1
fine-tuned FinBERT	Opportunity	.9589	.9790	.9689
	Risk	.8260	.7037	.7600
	weighted avg.	.9425	.9450	.9430
zero-shot GPT	Opportunity	.9641	.9842	.9740
	Risk	.8696	.7408	.8000
	weighted avg.	.9523	.9541	.9525
few-shot GPT	Opportunity	.9793	.9948	.9870
	Risk	.9583	.8519	.9020
	weighted avg.	.9768	.9770	.9765

Table 6: Extended test scores of our systems.

During our experiments, we saw that increasing the number of examples provided better results. We limited it to 6 in our approach for speed reasons: API has a token-per-minute limit, so to use more examples we would need to slow down the requests by increasing the interval between them, which led to a significant increase in time costs even on such a small Dev and Test subsets. We also tried several random sets of examples, and all of them led to almost the same results with minor differences in scores. However, recent findings show that few-shot results can be improved by using representative samples selected by a human expert (Loukas et al., 2023). This is an interesting research direction for the future work.

The addition of the examples helped increase the recall for the Risk class, thus producing a more balanced result, compared to the zero-shot version. This approach was used to produce the final submission #3.

5 Results

The organizers provided the micro-, macro-, and weighted averaged F1 scores (see Table 5), and also the Test data set with labels. In Table 6 we report the full scores for our three submissions.

As we can see, the few-shot approach outperforms the other two and reaches over 0.97 F1 score, ranking second among 21 total submissions for the English language.

6 Conclusions and Further work

We conducted several experiments, showing that even with limited data pre-trained LLMs are capable of achieving high scores (> 0.94 weighted F1) in Risk vs. Opportunity classification. We show that GPT outperforms FinBERT in both zero- in few-shot settings.

For further work, we consider fine-tuning GPT and ESG-BERT models, while also exploring GPT’s capabilities to reasonably explain its classification decisions in such a task, especially GPT-4. We also consider applying and evaluating the same approaches with the data in other languages, namely French, as even the top scores for it are at least 0.1 lower than for English. Another exciting direction would be exploring alternative translation models such as GPT-3.5, GPT-4, and Flan-T5 (Chung et al., 2022).

References

- Dogu Araci. 2019. [Finbert: Financial sentiment analysis with pre-trained language models](#).
- 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](#).
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).
- 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.
- Xianzhi Li, Xiaodan Zhu, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. [Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? an examination on several typical tasks](#).
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. [Gpt understands, too](#). *AI Open*.
- Lefteris Loukas, Ilias Stogiannidis, Prodromos Malakiotis, and Stavros Vassos. 2023. [Breaking the bank with chatgpt: Few-shot text classification for finance](#).
- Mukut Mukherjee. 2020. [Esg-bert: Nlp meets sustainable investing](#).
- Nasruzzaman Naeem, Serkan Cankaya, and Recep Bildik. 2022. [Does esg performance affect the financial performance of environmentally sensitive industries? a comparison between emerging and developed markets](#). *Borsa Istanbul Review*, 22:S128–S140. Environmental, Social and Governance (ESG) and Sustainable Finance.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Omid Sabbaghi. 2022. [The impact of news on the volatility of esg firms](#). *Global Finance Journal*, 51:100570.
- 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)*, Birmingham, United Kingdom. ACM, New York, NY, USA.
- Jin Boon Wong and Qin Zhang. 2022. [Stock market reactions to adverse esg disclosure via media channels](#). *The British Accounting Review*, 54(1):101045.

The Risk and Opportunity of Data Augmentation and Translation for ESG News Impact Identification with Language Models

Yosef Ardhito Winatmoko

Nestlé
Jakarta, Indonesia
yosef.ardhito@id.nestle.com

Ali Akbar Septiandri

Nokia Bell Labs
Cambridge, UK
ali.septiandri@nokia-bell-labs.com

Abstract

This paper presents our findings in the ML-ESG-2 task, which focused on classifying a news snippet of various languages as “Risk” or “Opportunity” in the ESG (Environmental, Social, and Governance) context. We experimented with data augmentation and translation facilitated by Large Language Models (LLM). We found that augmenting the English dataset did not help to improve the performance. By fine-tuning RoBERTa models with the original data, we achieved the top position for the English and second place for the French task. In contrast, we could achieve comparable results on the French dataset by solely using the English translation, securing the third position for the French task with only marginal F1 differences to the second-place model.

1 Introduction

ESG factors have gained increasing prominence in recent years, not only among stakeholders but also in the decision-making processes of investors and financial institutions. As the awareness of ESG risks grows, so does the need for precise and real-time classification of these risks. Traditional ESG risk analysis has largely relied on structured data, such as company disclosures, financial reports, and pre-defined ESG metrics. However, these sources often provide an incomplete and lagging view of a company’s ESG footprint. Moreover, they are subject to reporting biases and may lack granularity in capturing the diverse dimensions of ESG risks.

The proliferation of online news media offers a fertile ground for harvesting a more comprehensive set of data on ESG issues. News articles, in particular, often capture real-time events, public sentiment, and expert opinions, providing a more immediate and multifaceted perspective on ESG risks than can be obtained from traditional structured data. Yet, leveraging this unstructured textual data to accurately classify ESG risks presents computational challenges. These challenges include

but are not limited to, natural language understanding, sentiment analysis, and the development of a robust taxonomy for ESG risk classification.

The FinNLP-2022 workshop introduced a FinSim4-ESG¹ shared task that centres on ESG issues. To deepen the understanding of these areas, FinNLP@IJCAI-2023 released a new dataset for the FinNLP community. This dataset is designed to explore the task of identifying key ESG issues in multiple languages, guided by the MSCI ESG rating framework, which includes 35 key issues for categorisation.

Expanding on this discourse, a new task dubbed Multi-Lingual ESG Impact Type Identification (ML-ESG-2) was introduced. The primary objective of this task is to identify the type of ESG impact a given piece of news may have. Specifically, models are tasked with determining whether the news presents an ESG-related opportunity or risk. This aspect of impact identification is structured as a single-choice question.

In this study, we present our methodology for tackling the ML-ESG-2 shared task, using datasets in both English and French. We conducted our experiments using two primary methods: fine-tuning language models on the dataset and training a logistic regression model using Sentence-BERT (SBERT) embeddings (Reimers and Gurevych, 2019). Our findings suggest that the optimal approach can be achieved solely by relying on the given datasets, without the need for any data augmentation. Furthermore, we found that comparable results could be obtained on the French dataset by first translating the text into English and then employing a model pre-trained specifically on English text. Using this strategy, our models secured first place in the English language dataset and second and third places in the French language dataset.

¹<https://sites.google.com/nlg.csie.ntu.edu.tw/finnlp-2022/shared-task-finsim4-esg>

2 Datasets

In ML-ESG-2, a new task called “ESG Impact Type Identification” was introduced to advance discussions on ESG ratings from the previous shared task (Chen et al., 2023). This task requires models to discern whether a given news article signifies an ESG “opportunity” or “risk”. Each data point consists of the URL, title, content, and the assigned impact type for a given article. An overview of the English and French datasets can be seen in Table 1.

	English	French
Training-Opportunity	694 (85.9%)	458 (56.0%)
Training-Risk	114 (14.1%)	360 (44.0%)
Test-Opportunity	191 (87.6%)	111 (55.5%)
Test-Risk	27 (12.4%)	89 (44.5%)

Table 1: Summary statistics of the datasets

Data Augmentation While the French dataset exhibits a relatively balanced distribution across its classes, the English dataset has significantly more instances labelled as “Opportunity” compared to “Risk”. Drawing inspiration from the successful approach employed by the previous ML-ESG task winner (Lee et al., 2023), we experimented with data augmentation using GPT3Mix (Yoo et al., 2021) to augment the “Risk” training data during the fine-tuning of the English models. For more details on this process, please refer to Appendix A.1. Our approach involved leveraging the text-davinci-003 model from OpenAI to generate the additional training data.

Data Translation As a pivotal component of our experimental approach, we employed large language models (LLMs) to facilitate the translation of training data from French to English (see Appendix A.2). This translation step was essential to ensure that our models, primarily designed for English text processing, could effectively comprehend and learn from the French-language content within the dataset.

3 Methods

We conducted experiments using two primary methods: fine-tuning language models on the dataset and training a logistic regression model using SBERT embeddings (Reimers and Gurevych, 2019). Given the absence of development sets and the presence of imbalanced training data, we em-

ployed 5-fold cross-validation to assess the effectiveness of our approaches. Additionally, we observed that the news titles are not unique and two news contents with the same title can be both “Opportunity” and “Risk”. Thus, we decided to disregard the titles altogether and used only the content as the input.

Baseline. To demonstrate the performance improvements achievable through the two methods mentioned earlier, we initially established a simple baseline. This benchmark was created by employing TF-IDF and logistic regression using scikit-learn (Pedregosa et al., 2011). We retained the top 1000 features identified by TF-IDF and conducted hyperparameter optimisation (see Appendix A.3).

Fine-tuning language models. We experimented with three well-known pre-trained encoder models: DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa (He et al., 2021). Specifically, the model names in Hugging Face (Wolf et al., 2020) were `distilbert-base-uncased`, `roberta-large`, `microsoft/deberta-v3-large`. We also experimented with the XLM-RoBERTa (`xlm-roberta-large`) model (Conneau et al., 2020) for the French dataset. We fine-tuned the pre-trained models on the ML-ESG-2 dataset and used Optuna (Akiba et al., 2019) to find the optimal hyperparameters of each model. The final list of the hyperparameters is shown in Table 2.

Model Name	Batch Size	Learning Rate	Epoch
DistilBERT	32	2.5e-5	4
RoBERTa	16	1.3e-5	2
XLM-RoBERTa	4	6.8e-6	4
DeBERTa	4	2.3e-5	4

Table 2: Hyperparameters used in the model fine-tuning.

Sentence-BERT. Unlike traditional word embeddings that represent individual words, SBERT (Reimers and Gurevych, 2019) is designed to generate embeddings for entire sentences or paragraphs. It leverages pre-trained transformer-based models, such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and fine-tunes them specifically for sentence-level tasks. This fine-tuning process enables SBERT to capture contextual information, semantic meaning, and the relationships between sentences. In our experiment, we used Sentence-T5 (ST5) (Ni et al., 2021), a variant based on T5

(Raffel et al., 2020), to generate the embeddings as input for logistic regression. We used ST5 because it outperformed SBERT in a range of natural language inference (NLI) and question-and-answer (Q&A) tasks (Ni et al., 2021). Additionally, we conducted experiments with different sizes of the ST5 model in this study.

4 Results

Our experiments revealed several key findings in the context of model performance. Firstly, RoBERTa demonstrated outstanding performance on the English dataset, achieving an impressive F1 score of 85.19% (Table 3)². In comparison, ST5-XXL also performed well with an F1 score of 81.29%. Remarkably, both models achieved the best results without the need for data augmentation and translation.

When subjected to the test set, RoBERTa continued to shine, outperforming ST5-XXL by a significant margin. RoBERTa achieved an F1 score of 92.00%, while ST5-XXL scored 75.36% (Table 5). This discrepancy underscores RoBERTa’s robustness and suitability for this task.

Shifting our focus to the French dataset, our cross-validation results favoured ST5-XXL (FR→EN) over XLM-RoBERTa (FR), with an F1 score of 76.52% versus 72.17% (Table 4). Nevertheless, when assessing the model’s performance on the test set, XLM-RoBERTa emerged as the winner, achieving an F1 score of 86.12% (Table 5). These findings highlight the importance of considering both cross-validation and test set results when evaluating model performance.

Furthermore, the test results between XLM-RoBERTa and ST5-XXL differ only slightly (<1%). Note that XLM-RoBERTa was trained on the original French dataset, but ST5-XXL used the translated dataset. The results for ST5-XXL raise the potential of translating other languages to English and used English-based models for text classification.

5 Related Work

In recent studies (Lehman et al., 2023; Xu et al., 2023), it was discovered that relatively compact specialised clinical models exhibit significantly

superior performance compared to all in-context learning approaches when applied to LLMs. This superior performance holds true even when these clinical models are fine-tuned on a limited amount of annotated data. Additionally, their research revealed that pretraining on clinical tokens enables the development of smaller, more parameter-efficient models that can either match or surpass the performance of much larger language models trained on general text.

On a similar note, Septiandri et al. (2020) found that classical NLP techniques, i.e. bag-of-words and TF-IDF, could produce comparable results to the more advanced word2vec (Bojanowski et al., 2017) and BiLSTM with a fraction of the training time. The study focused on a binary classification task, similar to ML-ESG-2. They suggested that even though the tiny improvement from complex models is crucial in a competition, one should consider allocating more resources to improve the quality of the dataset in practical settings.

6 Conclusion

In summary, within the English sub-task of ML-ESG-2, the RoBERTa model fine-tuned only with the original data secured the top position, even without any particular strategy to address the class imbalance. Our investigation revealed that data augmentation failed to improve F1 scores on the training set. Similarly, adding translated French news for English models did not contribute to improved performance. From these findings, we deduce that increasing data quantity through augmentation and translation may not consistently benefit model performance.

For the French sub-task, we observed that translating to English provided better results for the training set. However, the multi-lingual RoBERTa model (XLM-RoBERTa) fine-tuned on the original French dataset achieved higher F1 for the test set, albeit slightly. These results indicate an opportunity for future works: the potential of translating text to English as a preprocessing strategy in other languages.

7 Availability

The code is available at <https://github.com/aliakbars/esg-finnlp>.

²All F1 scores shown are calculated based on the F1-score of the “Risk” class. We also include the weighted-F1 scores for the test set as presented in the final leaderboard of ML-ESG-2 task for reference.

Model	Dataset		
	EN	EN+AUG	EN+FR
Baseline	48.45% \pm 9.30%	-	-
ST5-Base	78.63% \pm 5.66%	78.32% \pm 7.56%	72.82% \pm 11.40%
ST5-XXL	81.29% \pm 2.72%	80.01% \pm 2.98%	79.08% \pm 5.45%
DistilBERT	79.84% \pm 9.28%	76.01% \pm 14.8%	71.26% \pm 11.72%
RoBERTa	85.19% \pm 8.97%	83.68% \pm 8.30%	82.48% \pm 9.65%
DeBERTa	82.23% \pm 10.83%	76.92% \pm 14.47%	72.19% \pm 19.99%

Table 3: **F1 scores on the English training dataset.** Apart from using only the English dataset (EN), we also experimented with augmenting the dataset using large language models (EN+AUG), and the English translation of the French dataset (EN+FR). We were only augmenting the training set using the Risk-labelled data points.

Model	Dataset	
	FR	FR \rightarrow EN
Baseline	65.61% \pm 4.66%	63.95% \pm 3.38%
ST5-Base	-	71.13% \pm 4.89%
ST5-XXL	-	76.52% \pm 4.26%
RoBERTa	-	67.33% \pm 12.73%
XLM-RoBERTa	72.17% \pm 5.43%	71.05% \pm 6.09%

Table 4: **F1 scores on the French training dataset.** FR \rightarrow EN indicates the use of the English translation of the French dataset during training.

Model	F1		Weighted F1	
	EN	FR	EN	FR
ST5-XXL	75.36%	85.96%	92.89%	83.94%
RoBERTa	92.00%	-	98.10%	-
XLM-RoBERTa	-	86.12%	-	85.54%

Table 5: **F1 scores on the test sets using the best models.** All models were trained on the vanilla training set, except ST5-XXL (FR) which used the translated version of the French (FR) dataset.

References

- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- 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, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual ESG impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- 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.
- 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.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. [DeBERTa: Decoding-enhanced BERT with disentangled attention](#). In *International Conference on Learning Representations*.
- Hanwool Lee, Jonghyun Choi, Sohyeon Kwon, and Sungbum Jung. 2023. [EaSyGuide: ESG issue identification framework leveraging abilities of generative large language models](#).
- Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. 2023. [Do we still need clinical language models?](#) In *Proceedings of the Conference on Health, Inference, and Learning*, volume 209 of *Proceedings of Machine Learning Research*, pages 578–597. PMLR.
- 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*.
- Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.
- 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. volume 35, pages 27730–27744.
- 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. *Journal of Machine Learning Research*, 12:2825–2830.
- 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(1).
- 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.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. [DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter](#). *arXiv preprint arXiv:1910.01108*.
- Ali Akbar Septiandri, Yosef Ardhito Winatmoko, and Ilham Firdausi Putra. 2020. [Knowing right from wrong: Should we use more complex models for automatic short-answer scoring in Bahasa Indonesia?](#) In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 1–7, 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.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Hong Yu, James Hendler, Anind K Dey, and Dakuo Wang. 2023. Leveraging large language models for mental health prediction via online text data. *arXiv preprint arXiv:2307.14385*.

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.

A Additional Experimental Details

A.1 Data Augmentation Prompt

We ran the augmentation 100 times, each time using three distinct samples from the actual dataset and taking five generated texts. In total, we generated 500 additional “Risk” contents for the English dataset using the following prompt:

```
Each line in the following list contains a
    snippet taken from a news article and the
    respective ESG impact identification.
ESG impact is one of 'Risk' and 'Opportunity'.
Opportunity: (a random 'Opportunity' content)
Risk: (another random 'Risk' content)
Opportunity: (another random 'Opportunity'
    content)
Risk:
```

A.2 Translating French to English

We experimented with three models to translate the text from French to English: Flan-T5 (Chung et al., 2022), DeepL³, and GPT-3.5 Turbo (Ouyang et al., 2022)). We found that the English translation using GPT-3.5 Turbo would result in a better performance. Thus, the results reported in this paper were based on the GPT-3.5 Turbo translation only.

A.3 Hyperparameter Tuning

For the logistic regression model trained on the SBERT embeddings and the baseline approach, we tuned the hyperparameters using the values provided in Table 6.

Hyperparameter	Values tested
C	{0.1, 1, 10, 100}
class_weight	{1, 2, 5, 10, 20}

Table 6: Hyperparameters tested for logistic regression

³<https://www.deepl.com/translator>

ESG Impact Type Classification: Leveraging Strategic Prompt Engineering and LLM Fine-Tuning

Soumya Smruti Mishra

Amazon Web Services

soumyasmruti@gmail.com

Abstract

In this paper, we describe our approach to the ML-ESG-2 shared task, co-located with the FinNLP workshop at IJCNLP-AAACL-2023. The task aims at classifying news articles into categories reflecting either “Opportunity” or “Risk” from an ESG standpoint for companies. Our innovative methodology leverages two distinct systems for optimal text classification. In the initial phase, we engage in prompt engineering, working in conjunction with semantic similarity and using the Claude 2¹ LLM. Subsequently, we apply fine-tuning techniques to the Llama 2² and Dolly³ LLMs to enhance their performance. We report the results of five different approaches in this paper, with our top models ranking first in the French category and sixth in the English category.

1 Introduction

Natural Language Processing (NLP) is pivotal in the finance sector, extracting valuable semantic information from vast unstructured data sources like reports, news, and social media. This extraction is crucial for identifying scenarios and analyzing risks, especially in the rising field of Environmental, Social, and Governance (ESG) considerations, as the global economy shifts towards sustainability. Financial markets and investors play a central role in supporting companies with strong ESG principles amidst growing interest in corporate sustainability performance. Using NLP, investors can swiftly and efficiently analyze sustainability reports and news, simplifying the traditionally complex manual process. With automatic text classification and sentiment analysis, NLP models identify crucial ESG topics and sentiments, saving time and resources while enabling informed and timely sustainable investment decisions.

2 Multi-Lingual ESG Impact Type Identification shared task

The "Multi-Lingual ESG Impact Type Identification (ML-ESG-2)" shared task, introduced by [Chen et al. \(2023a\)](#), aims to discern the type of ESG impact news articles exert on companies, focusing on

articles written in Chinese ([Tseng et al. \(2023\)](#)), English, French, and Japanese ([Kannan and Seki \(2023\)](#)). Each article is systematically categorized as either "Opportunity", "Risk", "Cannot Distinguish", or in the case of Japanese texts, "Positive", "Negative", or "N/A", serving as indicative labels of potential impacts. This task is a sequel to the initial ML-ESG shared task ([Chen et al., 2023b](#)) and is meticulously designed under the rigorous ESG rating guidelines provided by MSCI, seeking the development and evaluation of systems adept at accurately classifying articles into specific impact types. Within resource and time constraints, our team focused on classifying ESG impacts in English and French articles with comprehensive strategies, while solely applying Prompt Engineering to Japanese articles, demonstrating a tailored approach to multilingual classification.

2.1 Dataset Details

The ESG dataset, comprising English and French languages, encompasses columns such as `news_title`, `news_content`, `impact_type`, and the URL of the respective news articles. For the methodologies employed in this study, only the `news_title`, `news_content`, and `impact_type` columns were utilized as primary text columns. Conversely, the Japanese dataset contained columns labeled `sentence`, `URL`, and `sentiment`. Due to constraints in resources and time, the URL column from all datasets was omitted, precluding the scraping of additional news articles. Comprehensive statistics pertaining to these datasets are delineated in Table 1.

As observed in Table 1, the dataset is notably biased towards Opportunity or Positive sentiment. Training models on this dataset without addressing the imbalance would inevitably yield biased outcomes. Therefore, in our prompt engineering approaches, we attempted to sample equal amounts of training data from both Opportunity and Risk impact type groups wherever feasible. Additionally, during the fine-tuning process, we utilized a weighted cross-entropy loss to further mitigate the effects of imbalance.

Label	Language		
	English	French	Japanese
Opportunity / Positive	694	458	460
Risk / Negative	114	360	49
Cannot Determine	0	0	387

Table 1: Shows statistics related to training datasets

3 Related Work

3.1 Prompt Engineering and ESG

Prompt engineering is a pivotal technique in leveraging the capacities of large language models (LLMs), focusing particularly on in-context learning (ICL) for tasks like few-shot classification and semantic similarity. [Brown et al. \(2020\)](#) pioneered this approach with GPT-3, introducing in-context few-shot learning that serves as a foundation for subsequent improvements in ICL efficacy.

Research indicates that selecting appropriate examples in few-shot scenarios can yield high performance and near state-of-the-art accuracy, as evidenced by [Gao et al. \(2021\)](#) and [Liu et al. \(2022\)](#). These studies use sentence embeddings to select examples that closely align with the input in the embedding space. [Tanwar et al. \(2023\)](#) underscored the significance of aligning both semantics and task-specific textual signals across source and target language inputs in prompts, showcasing enhanced performance in cross-lingual text classification tasks. Their findings accentuate the value of dynamic, similarity-based example selection in guiding LLMs to develop superior in-context predictors, applicable to various language pairs. Inspired by these approaches, our primary motivation is to select examples from training data that are semantically similar to the news article being predicted and incorporate them into the prompt template.

Despite the burgeoning advancements in prompt engineering and in-context learning, their application to ESG-related texts is conspicuously limited, remaining at a nascent stage of exploration and understanding.

3.2 Fine-Tuning and ESG

In recent years, the field of natural language processing has become adept at utilizing fine-tuning methods for pre-trained language models with lim-

ited task-specific data [Howard and Ruder \(2018\)](#). Such techniques are particularly valuable in specialized domains characterized by the use of unique terminology not commonly found in general texts, as is often the case with sentences pertaining to ESG.

[Pasch and Ehnes \(2022\)](#) presents a novel approach to fine-tuning transformer-based models for the ESG domain, demonstrating enhanced prediction of companies’ ESG behaviors through a unique sentiment model trained with ESG ratings and annual report texts, outperforming traditional classifiers by up to 11 percentage points.

[Mukut Mukherjee and Parabole.ai \(2020\)](#) developed ESG-BERT by pre-training Google’s BERT on extensive sustainability text, enhancing its understanding of the domain-specific vocabulary crucial for Sustainable Investing. Building on these foundational works and methodologies presented in the literature, our study explores the application and adaptation of fine-tuning techniques within the specific context of our competition. We aim to leverage and extend the promising results observed in the aforementioned studies for analyzing ESG-related texts.

4 Techniques Explored

This section elucidates the various techniques employed during the competition. Our discussion will not only delve into the specifics of each method utilized but also provide a comprehensive understanding of the applied techniques.

4.1 Prompt Engineering with Semantic Similarity

Constructing a prompt template is pivotal in prompt engineering as it lays the foundation for effectively steering the language model’s responses and interactions, providing a structured format and necessary context. Our design for the prompt template, inspired by Pinecone’s article⁴ and refined after multiple attempts, is detailed in Appendix A.

The template initiates with an Objective section, outlining the knowledge base essential for the classification task. This knowledge base encompasses the Classification Criteria, delineated into positive (Opportunity), negative (Risk), and neutral (Cannot Distinguish) sections, as suggested by [Kannan and Seki \(2023\)](#). Subsequently, the Training Data section is introduced, comprising news_title, news_content,

Techniques & Models	English (TSC = 218)			French (TSC = 200)		
	Micro	Macro	Weighted	Micro	Macro	Weighted
PE + Claude 2	0.8853	0.5195	0.8968	–	–	–
PE + SEM-SIM + Claude 2	0.9174	0.5574	0.9256	0.7100	0.4918	0.7367
FT + EN (only) + Dolly	0.8716	0.4657	0.816	0.4450	0.3080	0.2741
FT + EN (only) + Llama2	0.9587	0.9118	0.9602	0.8700	0.8661	0.8686
FT + EN + FR + Llama2	0.9174	0.5062	0.9062	0.7150	0.7118	0.7085

Table 2: ESG Impact Type Identification results based on F1 scores, TSC: the count of news articles in test dataset; PE: Prompt Engineering; SEM-SIM: Semantic Similarity; FT: Fine Tuning; EN: English; FR: French

and `impact_type` in JSON format. Test sets, consisting of `news_title` and `news_content`, were presented in small batches for impact type prediction, with experiments conducted using batch sizes of 1, 5, and 10. Ultimately, a batch size of 10 was selected for predicting the impact type of news articles to optimize processing time and cost-efficiency in making API calls to the LLM. The prompt concludes with a specification of the strict output format required by the LLM, with further details available in Appendix A. The decision to employ Claude 2 was primarily due to its expansive 100K context window and accessibility through the Amazon Bedrock service⁵.

Prompt Engineering with Balanced Few-Shot

Examples: In this approach, the template, encompassing training and test data, was directly submitted to Claude’s API. Experiments with imbalanced randomly selected few-shot training data revealed a bias towards Opportunity / Positive labels in the validation set. For the English language task, 70 news articles were provided, evenly split between Opportunity and Risk labels to maintain a balanced set of training examples, thereby preventing bias in the LLM. The French dataset, characterized by longer news articles, permitted only 20 training examples (10 for each label class).

Prompt Engineering with Semantically Similar

Few-Shot Examples: In this approach, we employ the technique of selecting few-shot examples from training dataset, which are semantically similar to the news article being predicted (test-set). This approach involves using a template, detailed in, Appendix B, for classifying English, French, and Japanese news articles into ESG Impact Type categories. Implemented using the Langchain framework, this method leverages classes such as,

`FewShotPromptTemplate`⁶, `PromptTemplate`⁷, `SemanticSimilarityExampleSelector`⁸, and `SentenceTransformerEmbeddings`⁹ (using `all-MiniLM-L6-v2`¹⁰). By combining the structured guidance offered by the prompt template with the power of semantic similarity, this approach facilitates more nuanced and accurate classification of news articles based on their ESG impact. We utilize the same number of few-shot examples as discussed in the previous paragraph. The results of both the approaches are provided in first two rows of the Table 2. For the Japanese dataset, we submitted output solely from this approach, securing a 4th place rank among all submissions with a weighted F1 score of 0.4776.

4.2 Fine-Tuning of Instruction Tuned LLMs

Fine-tuning Large Language Models (LLMs) is a vital step in our methodology for enhancing model performance. We choose to use Llama 2 (Touvron et al., 2023) and Dolly (Conover et al., 2023), because they’re designed to understand and respond to specific input instructions. Their extensive training on diverse prompts allows for effective handling of complex ESG-related language, improving prediction reliability and accuracy. The names of the models in Hugging Face (Wolf et al., 2020) were `meta-llama/Llama-2-13b-hf`¹¹ and `databricks/dolly-v2-12b`¹².

Our methodology utilizes the prompt template from the semantic similarity approach (Appendix B) to fine-tune LLMs. For this process, each training example was converted into text using the prompt template. Efficiency in tuning was achieved through Low-Rank Adaptation (LoRA) Hu et al. (2021), which significantly minimized trainable parameters and GPU memory requirements.

In our initial experiment, both LLMs were

trained exclusively with English news articles, and these fine-tuned models were used to predict English and French news articles in both test and validation sets. Subsequently, in a second experiment, we fine-tuned Llama only using both English and French news articles. The choice to exclusively fine-tune Llama was informed by the significant performance disparities observed between Llama and Dolly during the initial experiment on the validation set. It’s crucial to note that fine-tuning Llama demanded five times more GPU clock time than Dolly, a characteristic also observed during the inference process.

The structured input, provided by the prompt template, enhances the models’ ability to understand subtle text nuances, crucial for ESG classification accuracy. This combination of instruction-tuned LLMs and prompt templates supports efficient and effective ESG text classification in our competition application. The results of the all the approaches discussed in this section are provided in last 3 rows of the Table 2 and hyperparameters used in Table 3.

Hyperparameters	Values
Batch Size	64
Gradient Accumulation Steps	8
Learning Rate	1e-5
Epoch	10
LORA-R	512
LORA-ALPHA	1024
LORA-DROPOUT	0.05
Optimizer	Adam
Warmp Up	15%
Max Grad Norm	0.3

Table 3: Hyperparameters used in fine-tuning of LLM’s

5 Result

The performance of our models is demonstrated in Table 2, where the numbers provided represent the F1 scores on the test dataset. We submitted the top three out of our five models, which were the best performers on our validation set. Due to page limits, we do not provide the validation set F1 scores and results here. The organizers published the official results for each language, labeling our models as SPEvFT. Our models ranked 6th best in the English category and were the best in the French category. For the English and French datasets, we submitted three runs: 1) Prompt Engineering with

Semantic Similarity (PE + SEM-SIM); 2) Fine Tuning with the English dataset on Dolly (FT + EN(only) + Dolly); 3) Fine-tuning with the English dataset on Llama (FT + EN(only) + Llama2). Run 3 achieved superior performance compared to the other runs, as highlighted in bold in Table 2. These results suggest that fine-tuning with prompt templates yields the best outcomes compared to other approaches. The relatively weaker performance of Dolly might be attributable to the number of tokens used in the original pre-trained model.

6 Lessons Learned

Several key lessons emerged from our work. We omitted translating the French dataset, and training on non-translated data didn’t enhance classification scores for French samples. This raises questions about the potential impact of initially translating news articles and integrating them into templates, as well as the appropriateness of the French and Japanese translations in the prompt templates. Further experimentation and investigation are warranted to clarify these issues. Second, while our approach was potent, it was also resource-intensive. The demanding nature of LLMs prompts consideration of using more cost-effective alternatives. Additionally, considering the superior performance often exhibited by fine-tuned pre-trained models for discriminative tasks, we should explore the use of models like RoBERTa. Furthermore, future efforts should also involve experimenting with diverse and concise prompt templates.

7 Conclusion

This paper presents a nuanced approach developed for the FinNLP 2023 ML-ESG-2 shared task, with a focus on classifying news articles into “Opportunity” or “Risk” categories for companies from an ESG perspective. Our dual-strategy method seamlessly integrated Prompt Engineering with Semantic Similarity and subsequent fine-tuning of LLMs, notably enhancing their performance. With our models, we achieved a commendable 6th position in the English category and 1st in the French category. Particularly noteworthy was the superior performance demonstrated by the fine-tuned Llama 2 model, highlighting its promising application potential in ESG texts. Our findings offer valuable insights, contributing significantly to the field of ESG impact classification and suggesting promising directions for future research in this domain.

Notes

¹<https://www.anthropic.com/index/claude-2>

²<https://ai.meta.com/llama/>

³<https://github.com/databricks/dolly>

⁴See <https://www.pinecone.io/learn/series/langchain/langchain-prompt-templates/> for the inspiration behind our prompt template design.

⁵<https://aws.amazon.com/bedrock/>

⁶<https://js.langchain.com/docs/api/prompts/classes/FewShotPromptTemplate>

⁷<https://js.langchain.com/docs/api/prompts/classes/PromptTemplate>

⁸<https://js.langchain.com/docs/api/prompts/classes/SemanticSimilarityExampleSelector>

⁹https://python.langchain.com/docs/integrations/text_embedding/sentence_transformers

¹⁰<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

¹¹<https://huggingface.co/meta-llama/Llama-2-13b-hf>

¹²<https://huggingface.co/databricks/dolly-v2-12b>

References

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](#).

Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. Multi-lingual esg impact type identification. In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.

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 (FinNLP)*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. [Free dolly: Introducing the world's first truly open instruction-tuned llm](#).

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. [Making pre-trained language models better few-shot learners](#). 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 3816–3830, Online. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. [Universal language model fine-tuning for text classification](#).

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](#).

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)*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. [What makes good in-context examples for GPT-3? In Proceedings of Deep Learning Inside Out \(DeeLIO 2022\): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures](#), pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Charan Pothireddi Mukut Mukherjee and Parable.ai. 2020. [Esg-bert: Nlp meets sustainable investing](#).

Stefan Pasch and Daniel Ehnes. 2022. [Nlp for responsible finance: Fine-tuning transformer-based models for esg](#). In *2022 IEEE International Conference on Big Data (Big Data)*, pages 3532–3536.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. [Multilingual llms are better cross-lingual in-context learners with alignment](#).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrutu 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](#).

Yu-Min Tseng, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2023. [Dynamicsesg: 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)*.

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.

A Initial Prompt Template

This prompt template is formatted in compliance with the standards set by Anthropic¹ for Claude 2 LLM. The template utilizes Human: and Assistant: formatting to facilitate the conversational agent Claude in speaker identification, with Human: serving for prompts or instructions and Assistant: (Claude) for responses.

Template:

Human:

Objective: Based on the provided training data, classify news articles concerning their ESG implications for a company. Prioritize and give significant weight to the training examples when making your classification, in addition to your inherent knowledge base.

Classification Criteria:

Opportunity:

- An event that could potentially yield positive returns in areas related to environment, social impact, governance, etc.
- Contains phrases stating that the company is improving or is likely to improve in the future its long-term values.
- Contains other phrases that are considered positive from an ESG perspective.

Risk:

¹<https://docs.anthropic.com/claude/docs/constructing-a-prompt>

- An event that could potentially yield negative returns or threaten the positive returns concerning ESG issues.
- Contains phrases that could be viewed as potentially detrimental to the long-term value of the company.
- Contains other phrases that could be considered negative from an ESG perspective.

Cannot Distinguish:

- If the article does not distinctly indicate it as either an opportunity or risk.

Training Data (Highly Important):

train_set

This training data is critical. Ensure to use these examples as your primary reference for classification.

Articles for Classification:

test_set

Instruction: Carefully classify the unlabeled articles, giving high importance to the training examples provided.

Output Format:

For each article, produce a JSON object with the ID and the classification. The output should be a list of JSON objects like this example below.

example

Only submit the JSON output. Do not include any additional explanations or text.

Assistant:

Listing 1: Example of the json output expected from the Assistant this is also provided as a sample in the template's 'example' variable

```
[
  {
    "article_number": "1",
    "impact_type": "Opportunity"
  },
  {
    "article_number": "2",
    "impact_type": "Risk"
  }
]
```

B Prompt Template using LangChain

The below template was used in experiments for semantic similarity approach and also in the fine-tuning approaches.

Template:

Question:

Given below is a news article in English, you need to take up the role to classify the given news is an opportunity or risk from the ESG (environmental, social and governance) aspect.

Classification Criteria: Opportunity:

- An event that could potentially yield positive returns in areas related to environment, social impact, governance, etc.
- Contains phrases stating that the company is improving or is likely to improve in the future its long-term values.
- Contains other phrases that are considered positive from an ESG perspective.

Risk:

- An event that could potentially yield negative returns or threaten the positive returns concerning ESG issues.
- Contains phrases that could be viewed as potentially detrimental to the long-term value of the company.
- Contains other phrases that could be considered negative from an ESG perspective.

Cannot Distinguish:

- If the article does not distinctly indicate it as either an opportunity or risk.

Based on the classification criteria given above, How will you classify the news article provided below?

News title: news_title

News Content: news_content

Impact Type: impact_type

Output Format: The output should be a key value object like these examples below.

Impact Type: Opportunity

Impact Type: Risk

Impact Type: Cannot Determine

Only reply with the key value output. Do not include any additional explanations or text.

Exploring Knowledge Composition for ESG Impact Type Determination

Fabian Billert^{1,2} and Stefan Conrad¹

¹Heinrich-Heine University of Düsseldorf

²GET Capital AG

{fabian.billert, stefan.conrad}@hhu.de

Abstract

In this paper, we discuss our (Team HHU’s) submission to the Multi-Lingual ESG Impact Type Identification task (ML-ESG-2). The goal of this task is to determine if an ESG-related news article represents an opportunity or a risk. We use an adapter-based framework in order to train multiple adapter modules which capture different parts of the knowledge present in the training data. Experimenting with various Adapter Fusion setups, we focus both on combining the ESG-aspect-specific knowledge, and on combining the language-specific-knowledge. Our results show that in both cases, it is possible to effectively compose the knowledge in order to improve the impact type determination.

1 Introduction

The substantial rise in Environmental, Social, and Governance (ESG) research over the past few years underscores the growing significance of sustainability in the corporate world (Zumente and Bistрова, 2021). Both investors and companies have a growing interest in ESG issues, as it becomes clearer that they are vital for a company’s brand and, consequently, also value (Schramm-Klein et al., 2016), (Islam et al., 2021). Studies show that investor interest in a company’s stock depends significantly on whether an estimation of their ESG-practices is available or not (Zumente and Lāce, 2021).

Different agencies have developed a variety of ESG-scoring mechanisms in order to quantify the sustainability practices of companies which offers investors an easier way to determine the ESG-related risk a company represents in their portfolio. Popular examples of this are MSCI¹ and Sustainalytics². However, inconsistencies in

the scoring practices and discrepancies between different score-providers cast doubt on the available ESG-scores (Berg et al., 2019), (Clément et al., 2023). This doubt could be alleviated by providing a more detailed overview of the various aspects of sustainability a company is involved in. Researchers in the field of natural language processing (NLP) have recently made efforts to enhance conventional ESG scoring methodologies by incorporating alternative data sources, aimed at offering a clearer understanding of a company’s sustainability practices. (Nicolas et al., 2023) underscores the impact of ESG-risk events found in social media data, illustrating how they can result in adverse financial returns. Meanwhile, (Aue et al., 2022) focuses on analyzing news articles to construct a predictive model for determining ESG scores.

The ML-ESG series of shared tasks aims to advance research in this area of using natural language processing to provide more transparency in ESG topics. ML-ESG-1 focused on determining the ESG-aspects of news articles (Chen et al., 2023b). In the next step of the ML-ESG series, ML-ESG-2 aims to find the impact type of the information in news articles, meaning detecting if they represent opportunities or risks (Chen et al., 2023a).

In this paper, we present our submission for the French and English part of the ML-ESG-2 shared task. We train adapters on different parts of the dataset and experiment with several approaches in order to combine their knowledge using Adapter Fusion. We show that this knowledge composition yields a significant improvement of the performance compared to simply finetuning an adapter for the whole dataset. Our submitted approaches achieve fourth place for French and fifteenth place for English, however we investigated additional approaches after the task deadline which improve on

¹<https://www.msci.com/our-solutions/esg-investing/esg-ratings>

²<https://www.sustainalytics.com/esg-data>

those results.

2 Task Description and Dataset

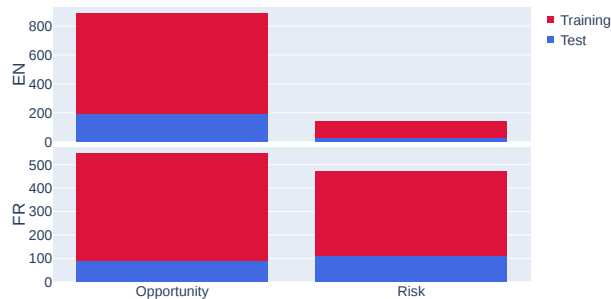


Figure 1: Occurrences of the different labels in the training- and test-data of the task. The top row represents the English data, the bottom row the French data. Training data is shown in red, test data in blue. **Note:** The test data was released after the task deadline and was not used during training.

ML-ESG-2 presents data in four different languages: English, French, Japanese and Chinese. Because the Japanese and Chinese labels are slightly different than the English and French ones, we chose to only participate in the latter languages. We display information about the training dataset in Figure 1 in red. For English, the training dataset contains 808 news articles while there are 818 news articles for French. Each news article has a news title, a news content and an associated impact type label. Different samples can have the same news title but a different news content and a different impact type, which happens a total of 100 times in the two datasets. For this reason, we didn't use the news title in our approaches.

The impact type is labelled as either "Opportunity" or "Risk". While the data is quite imbalanced for English, with the "Opportunity" label being the majority class, it is approximately even for French. The average length of the news content is 412 characters for English and 565 characters for French, which should fit well within the maximum length of 512 tokens of BERT-based models. Otherwise, the samples are truncated.

2.1 ESG-Aspect Dataset

The previous ML-ESG shared task focused on determining one out of 35 key issues defined by MSCI³, where each key issue belongs to one ESG-aspect (E/S or G) (Chen et al., 2023b). The dataset

³<https://www.msci.com/our-solutions/esg-investing/esg-industry-materiality-map>

shared there consists of 1200 news articles in English and French each, some of which overlap with the news articles used in the dataset we work on here. We present some approaches of using the knowledge available in the ML-ESG-1 dataset to improve the determination of "Opportunity" or "Risk" of the current task in subsection 3.3.

3 Experimental Approach

The basis of our approach consists of mBERT (Devlin et al., 2018), a multilingual BERT-based model. For fine-tuning, we use bottleneck-adapter-modules (Houlsby et al., 2019), which are a small set of weights inserted into the layers of the base-model. In order to account for the specific language we are training in, we implement an approach described in (Pfeiffer et al., 2020b), where a pre-trained language adapter is inserted before the adapter that is being fine-tuned (known as the task-adapter). We use the language adapters from the adapterhub (Pfeiffer et al., 2020a), which are available for all task-languages for the mBERT architecture. Adapter-modules allow us to be more flexible in training our model, as we will describe in subsection 3.2.

To train the adapters, we split the dataset and use 10% as the evaluation set. We use a learning rate of $5e - 5$. Finally, to account for the label imbalance in the training data we can observe in Figure 1, we use a weighted loss function.

3.1 Data Augmentation

To augment the data available in the two different languages, we translate both of them to the other language respectively. Since translating the whole text at once sometimes produced artifacts, we split the news content into sentences using nltk's sentence tokenizer (Bird et al., 2009) and translate a single sentence at a time. In order to translate, we use the OPUS-MT models (Tiedemann and Thottingal, 2020) from the huggingface-hub⁴.

3.2 Adapter Fusion

In section 3, we explain our adapter-based approach, where we use a base model, a pre-trained language adapter and finally a task-adapter for fine-tuning. In this configuration, the different components are aligned in series. Adapters can also be used in parallel, which allows for the combination of the task-specific knowledge of each adapter.

⁴<https://huggingface.co/Helsinki-NLP>

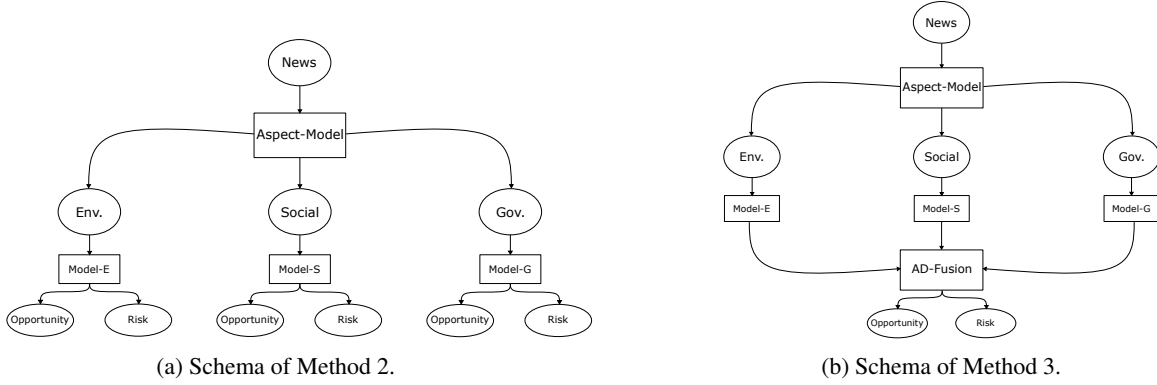


Figure 2: Schemata for Methods 2 and 3. The Aspect Model is obtained using data from ML-ESG-1 (Chen et al., 2023b). The E/S/G-Models in Method 3 are the same ones we train in Method 2.

This is known as Adapter-Fusion. Adapter-Fusion demonstrates particular advantages in scenarios involving limited datasets, as it can place greater emphasis on the knowledge obtained from the task-adapters trained on larger datasets (Pfeiffer et al., 2021). In addition, Adapter-Fusion has shown to lead to performance-improvements in multilingual-scenarios as it can use knowledge from training in other languages (Billert and Conrad, 2023). Since the concept of "Opportunity" and "Risk" can mean different things depending on the context, we aim to use the learnings from ML-ESG-1 (Chen et al., 2023b) in order to train aspect-specific adapters. These adapters will learn more specifically to determine "Opportunity" and "Risk" in the environmental, social and governance context. In practice this means that we will first train a model to determine the rough ESG-aspect of a news-article (environmental, social, governance), before training three different aspect-specific adapters to determine the ESG impact type for articles of each aspect. We achieve an F_1 -Macro score of around 0.86 for English and 0.79 for French for the adapters classifying the ESG-aspect, meaning that there is still the possibility of misclassifying the aspect in the first place, which might be detrimental for the downstream adapters predicting the impact type.

After training the three aspect-specific adapters, it is then possible to train an Adapter Fusion layer to make a final prediction for the ESG impact type.

3.3 Configurations

In order to test the premise that the ESG impact type depends on the aspect (as we describe in the previous subsection), we designed three different configurations.

- Method 1: Train an adapter on the two impact type labels directly.
- Method 2: Determine the ESG-aspect first, then train an adapter for the news articles belonging to each aspect separately in order to determine the impact type.
- Method 3: Use the same approach as in Method 2, but use Adapter-Fusion in order to combine the knowledge of the three separate adapters.

Method 2 and method 3 are depicted in Figure 2. Note that for the governance aspect-adapters, we augment the evaluation data with the governance-evaluation data from the other language respectively, because the evaluation sets have a very small amount of samples.

4 Results

The test data is shown in Figure 1 in blue. For English, the label imbalance in the training set is also present in the test set. For French, while there is still no clear imbalance, the "Risk" label now occurs more often as opposed to the "Opportunity" label in the training set.

In Table 1, we show the F_1 -Macro scores of the various approaches. The three methods we submitted can be seen on the left side of the table. For English, method 1 and method 3 have the same score, with method 2 being much worse than both of them. For French, method 1 shows the best performance, with method 3 being slightly worse. Again, method 2 trails far behind.

Interestingly, method 2 shows a big discrepancy to method 3, even though the same adapters are used, the only difference being the Adapter-Fusion layer.

Language	Pre-Deadline Results			AD-Fusion Experiments	
	Method 1	Method 2	Method 3	w/o gov	Lang-Fusion
EN	<u>0.8098</u>	0.4225	<u>0.8098</u>	0.8771	0.8557
FR	<u>0.7548</u>	0.6169	0.7457	0.7726	0.8084

Table 1: F₁-Macro scores of the trained adapters for the test-set. On the left side, the submitted results. On the right side, there are additional experiments that were done after the deadline. Bold values represent the best achieved score, while underlined values represent the best submitted score for each language.

4.1 Aspect-Fusion

Since there was not a lot of training data for the governance adapter in methods 2 and 3, we suspect that it is detrimental to the adapter fusion performance. We do several experiments without the governance adapter, using only the environmental and social adapters for the Adapter Fusion. The results are depicted on the right side of Table 1. We can see that the performance surpasses both the previous score of method 3 as well as the score of method 1, which confirms our previous suspicions about the governance adapter.

For English, we note an improvement of almost 7 points while for French, we see an increase slightly lower than 2 points. In both cases, the performance is significantly higher than the previous highest result, demonstrating that the concept of ESG impact type depends on the ESG aspect and that the trained Adapter Fusion layer successfully composes the knowledge learned by our aspect-specific adapters.

4.2 Language-Fusion

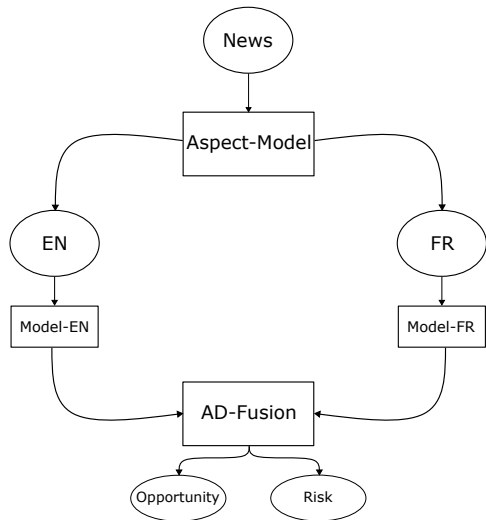


Figure 3: Adapter Fusion on the different language adapters from Method 1.

As mentioned in subsection 3.2, it is also possible to use Adapter Fusion to combine informa-

tion learned from tasks in different languages. To test if this also holds true in our case, we train an Adapter Fusion layer consisting of both language-specific adapters of method 1. The setup is displayed schematically in Figure 3. This is a similar approach as was used in (Billert and Conrad, 2023), although we only work with two languages in this case.

In the rightmost column of Table 1, we can see that this configuration achieves an outperformance of about 5 points over just using the language-specific adapters of method 1. Compared to using the aspect-fusion approach described in subsection 4.1, the results are conflicting. For English, the aspect-fusion shows better results, but for French, using the language-fusion seems superior.

5 Conclusion

In this work, we present multiple approaches using adapters to determine the ESG impact type of news articles. We show that it is possible to gain significant performance increases by using Adapter Fusion, which allows us to combine the knowledge present in the separate adapter modules, by experimenting with implementing Adapter Fusion in two different ways: Firstly, using the prior knowledge about ESG aspects, we train an adapter for each aspect and fuse them to determine the ESG impact type. Secondly, we fuse the adapters trained on the different languages in order to combine the knowledge learned from the different language datasets. Further studies could focus on how to combine the two Adapter Fusions, which we did not attempt in this work due to time constraints. In addition, it might be interesting to use the Japanese and Chinese datasets to try to improve the language-fusion approach even further. Finally, putting more focus on augmenting the training data, e.g. by using paraphrases, could bridge the gap between this study and the top-placed systems.

References

- Tanja Aue, Adam Jatowt, and Michael Färber. 2022. [Predicting Companies' ESG Ratings from News Articles Using Multivariate Timeseries Analysis](#). *arXiv*.
- Florian Berg, Julian F Kölbl, and Roberto Rigobon. 2019. [Aggregate Confusion: The Divergence of ESG Ratings](#). *SSRN Electronic Journal*.
- Fabian Billert and Stefan Conrad. 2023. [HHU at SemEval-2023 Task 3: An Adapter-based Approach for News Genre Classification](#). *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1166–1171.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023a. [Multi-Lingual ESG Impact Type Identification](#). In *Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing (FinNLP)*.
- 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 (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin)*, pages 111–115, Macao.
- Alexandre Clément, Élisabeth Robinot, and Léo Trepspeuch. 2023. [The use of ESG scores in academic literature: a systematic literature review](#). *Journal of Enterprising Communities: People and Places in the Global Economy*, ahead-of-print(ahead-of-print).
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). *arXiv*.
- 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](#). *arXiv*.
- Tahir Islam, Rauf Islam, Abdul Hameed Pitafi, Liang Xiaobei, Mahmood Rehmani, Muhammad Irfan, and Muhammad Shujaat Mubarak. 2021. [The impact of corporate social responsibility on customer loyalty: The mediating role of corporate reputation, customer satisfaction, and trust](#). *Sustainable Production and Consumption*, 25:123–135.
- Maxime L D Nicolas, Adrien Desroziers, Fabio Caccioli, and Tomaso Aste. 2023. [ESG Reputation Risk Matters: An Event Study Based on Social Media Data](#). *arXiv*.
- Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. [AdapterFusion: Non-Destructive Task Composition for Transfer Learning](#). *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.
- Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020a. [AdapterHub: A Framework for Adapting Transformers](#). *arXiv*.
- Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020b. [MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer](#). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673.
- 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.
- Jörg Tiedemann and Santhosh Thottingal. 2020. [OPUS-MT – Building open translation services for the World](#). *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*.
- 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.
- Ilze Zumente and Natalja Lāce. 2021. [ESG Rating—Necessity for the Investor or the Company?](#) *Sustainability*, 13(16):8940.

Enhancing ESG Impact Type Identification through Early Fusion and Multilingual Models

Hariram Veeramani
Department of Electrical
and Computer Engineering,
UCLA, USA
hariram@ucla.edu

Surendrabikram Thapa
Department of Computer
Science, Virginia Tech,
Blacksburg, USA
sbt@vt.edu

Usman Naseem
College of Science and
Engineering, James Cook
University, Australia
usman.naseem@jcu.edu.au

Abstract

In the evolving landscape of Environmental, Social, and Corporate Governance (ESG) impact assessment, the ML-ESG-2 shared task proposes identifying ESG impact types. To address this challenge, we present a comprehensive system leveraging ensemble learning techniques, capitalizing on early and late fusion approaches. Our approach employs four distinct models: mBERT, FlauBERT-base, ALBERT-base-v2, and a Multi-Layer Perceptron (MLP) incorporating Latent Semantic Analysis (LSA) and Term Frequency-Inverse Document Frequency (TF-IDF) features. Through extensive experimentation, we find that our early fusion ensemble approach, featuring the integration of LSA, TF-IDF, mBERT, FlauBERT-base, and ALBERT-base-v2, delivers the best performance. Our system offers a comprehensive ESG impact type identification solution, contributing to the responsible and sustainable decision-making processes vital in today's financial and corporate governance landscape.

1 Introduction

In the rapidly evolving landscape of finance and corporate governance, Environmental, Social, and Corporate Governance (ESG) considerations have gained unprecedented prominence (Linhares Pontes et al., 2022). Investors, stakeholders, and businesses increasingly recognize the importance of ESG factors in decision-making processes, portfolio management, and corporate strategy (Li et al., 2021; Gillan et al., 2021). The “E” in ESG shines a spotlight on environmental sustainability, compelling companies to address pressing issues such as climate change and resource conservation. Recognizing that eco-friendly practices are ethical and financially savvy, businesses are embracing environmental responsibility as a key driver of long-term success (Alsayegh et al., 2020; Newell and Marzuki, 2022). Simultaneously, the “S” in ESG emphasizes social responsibility, pushing corpora-

tions to foster diversity, inclusivity, and community engagement (Buallay, 2019). Companies that prioritize these social factors not only enhance their reputation but also attract top talent and build resilience in a socially conscious world. Finally, the “G” underscores the importance of sound corporate governance, providing the foundation for trust and stability in financial markets. Investors now realize that transparent decision-making, ethical conduct, and effective risk management are essential elements for ensuring a company's long-term viability (Tian et al., 2022).

In this context, the ML-ESG-2 shared task is pivotal in facilitating a deeper understanding of how ESG information can be extracted and utilized from the vast sea of textual data, contributing to the broader endeavor of responsible and sustainable finance. This specific task of ESG impact type identification, where machine learning models classify news articles or sentences as “Opportunity”, “Risk”, or “Cannot Distinguish”, carries profound significance. For investors, this knowledge translates into well-informed decisions with the potential to impact portfolios significantly (Amel-Zadeh and Serafeim, 2018). Positive ESG signals can guide investments toward companies poised for sustainable growth while identifying ESG risks enables proactive risk mitigation. In the corporate sphere, grasping the ESG implications of textual data like news articles and annual reports fosters strategic decision-making and risk management. Companies can leverage positive ESG news to enhance their sustainability efforts and attract responsible investors while swiftly addressing negative news to minimize reputational damage and regulatory challenges (Goel et al., 2022). Furthermore, automated text analysis tools expedite these processes, enabling scalable and efficient ESG information extraction.

As global issues like climate change, social inequality, and ethical governance become increas-

ingly prominent, harnessing ESG information from textual sources is a means to hold corporations accountable. By scrutinizing relevant text like news articles or annual securities reports for their ESG impact, stakeholders can encourage responsible behavior, promote sustainable practices, and ensure that businesses prioritize societal and environmental concerns (Kannan and Seki, 2023). This task not only empowers decision-makers with the ability to navigate the complexities of ESG information but also contributes to a broader cultural shift towards responsible investment and corporate citizenship. In essence, ESG information extraction from text represents a vital step towards a more sustainable and equitable future, where financial decisions align with environmental and social responsibility goals.

In Natural Language Processing (NLP), text classification is a critical enabler for tasks like ESG impact type identification, bridging the gap between the vast expanse of textual data and actionable insights. NLP techniques have been widely used in the financial domain (Jaggi et al., 2021; Adhikari et al., 2023). Leveraging NLP techniques, machine learning models can autonomously categorize news articles or sentences into predefined ESG impact types, such as “Opportunity”, “Risk”, or “Cannot Distinguish”. These models harness the power of feature extraction, supervised learning, feature engineering, and even advanced transfer learning from pre-trained language models to accurately assess the implications of textual content on a company’s ESG performance. The synergy between NLP and ESG, impact type identification, is instrumental in empowering stakeholders, including investors and corporations, to efficiently navigate the complexities of ESG information, make informed decisions, and contribute to a more sustainable and responsible financial and corporate governance landscape (Kang and El Maarouf, 2022).

This paper presents the system description of our submission to ML-ESG-2, a shared task hosted by the FinNLP workshop in conjunction with IJCNLP-AACL 2023. In this task, we delve into the critical intersection of Natural Language Processing (NLP) and ESG impact type identification.

2 Task Description

ML-ESG-2 introduces a new task called ESG impact type identification.

2.1 Objective

Given textual data, the objective is to classify it into one of three categories from the ESG aspect: “Opportunity”, “Risk”, or “Cannot Distinguish” (or “Positive”, “Negative”, or “N/A” in the Japanese dataset).

2.2 Dataset

The ML-ESG-2 task utilizes a carefully curated dataset by Kannan and Seki (2023) to support the ESG impact type identification objective. This dataset was curated using Japanese annual securities reports as a primary source of information to extract insights into a company’s ESG (Environmental, Social, and Corporate Governance) initiatives. In the context of ESG impact type identification, each news article or sentence is categorized into one of three distinct labels. The label “Opportunity” is assigned to articles or sentences that suggest a positive impact or opportunity for the company from an ESG perspective. Conversely, the label “Risk” is assigned to articles or sentences that convey a negative impact or potential risk to the company concerning ESG factors. However, in cases where the available content does not provide sufficient clarity to definitively classify the article as either an opportunity or a risk, the label “Cannot Distinguish” is used. It is worth noting that in the Japanese dataset, these labels are represented as “Positive”, “Negative”, and “N/A” to align with the language and terminology commonly used in Japanese ESG contexts.

3 System Description

In this section, we provide a detailed overview of our system architecture and methodology for the ML-ESG-2 shared task on ESG impact type identification. Our approach is founded on the principles of ensemble learning, utilizing both early fusion and late fusion techniques to harness the strengths of multiple language models and feature representations. In order to approach this task, we use four different models.

3.1 Models Used

In our system, we leverage the predictive capabilities of four different models, each contributing unique strengths to the ESG impact type identification process:

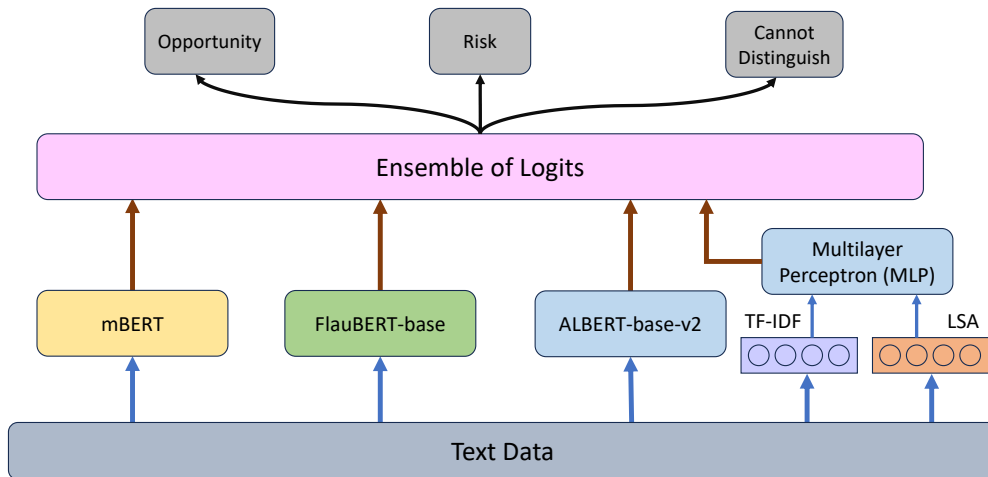


Figure 1: Late fusion technique that uses logits from all models to make the final decision.

3.1.1 mBERT (Multilingual BERT)

To ensure robust multilingual text analysis, we fine-tune mBERT (Devlin et al., 2019; Wu and Dredze, 2020) on the task-specific dataset. This model adapts to contextual information and linguistic nuances in the news articles, making it a valuable component for handling multilingual content (Veeramani et al., 2023c,b,d) effectively.

3.1.2 FlauBERT-base (French Language Model)

Given the presence of French textual content in the task, we incorporate FlauBERT-base (Le et al., 2020), a specialized French language model. This addition ensures accurate analysis and classification of French content, enhancing the overall system’s robustness.

3.1.3 ALBERT-base-v2

We integrate ALBERT-base-v2 (Lan et al., 2019), known for its efficient parameterization and performance, to diversify our model components. This model further enhances our system’s capability to capture nuanced ESG-related nuances and impact types.

3.1.4 MLP (Multi-Layer Perceptron)

Complementing the language models, we employ a Multi-Layer Perceptron (MLP) (Murtagh, 1991) that incorporates feature representations derived from Latent Semantic Analysis (LSA) (Dumais, 2004) and Term Frequency-Inverse Document Frequency (TF-IDF) (Bafna et al., 2016; Adhikari et al., 2021). These features capture semantic infor-

mation and term importance within the documents. The MLP facilitates the modeling of intricate feature interactions (Veeramani et al., 2023a,e), enhancing the depth of our system’s analysis.

3.2 Ensemble Techniques

Our ensemble strategy combines both early fusion and late fusion techniques, capitalizing on the collective intelligence of individual models and feature representations.

3.2.1 Late Fusion Ensemble

In the late fusion approach, we aggregate predictions from individual models and feature representations at the logits level (Kanagasabai et al., 2023). This technique effectively combines the diverse outputs generated by mBERT, FlauBERT-base, ALBERT-base-v2, and the MLP with LSA and TF-IDF features. Late fusion ensures comprehensive integration of the predictive capabilities of these components, resulting in a more robust and accurate prediction for ESG impact type identification. Figure 1 shows the schematic overview of the late fusion ensembling technique we employ.

3.2.2 Early Fusion Ensemble

In the early fusion approach, we integrate representations from mBERT, FlauBERT-base, ALBERT-base-v2, LSA, and TF-IDF features at an earlier stage, prior to inputting them into the MLP for prediction. Early fusion allows for seamless information integration from multiple sources, empowering the MLP to capture intricate relationships among these elements. This approach ensures the

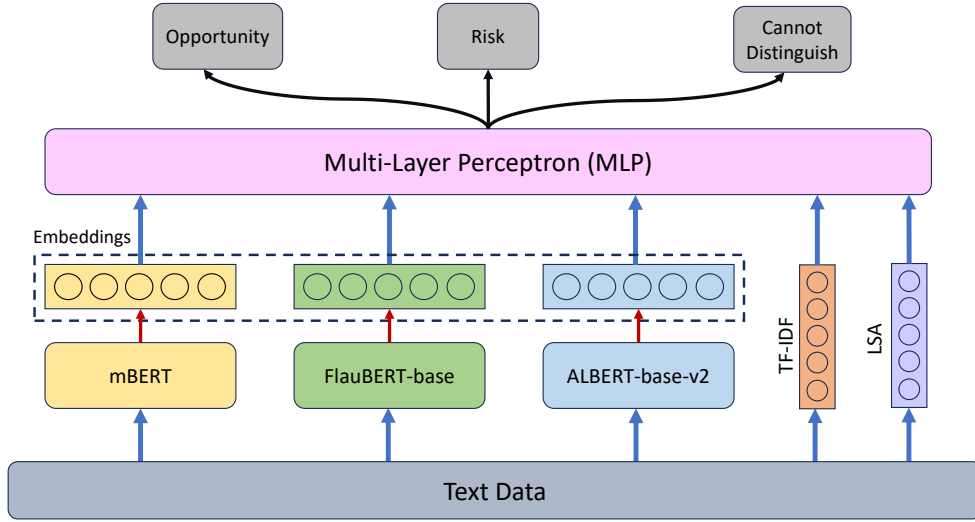


Figure 2: Early fusion ensemble takes the different representations and uses MLP for the final classification.

Embeddings	Language	Micro-F1	Macro-F1	Weighted-F1
FlauBERT + mBERT + ALBERT + TF-IDF	English	0.9587	0.9128	0.9587
	French	0.545	0.520	0.5111
	Japanese	0.5323	0.2933	0.4905
	Chinese	0.403	0.1667	0.3782
FlauBERT + mBERT + ALBERT + TF-IDF + LSA	English	0.9633	0.918	0.9639
	French	0.5501	0.5292	0.5155
	Japanese	0.5378	0.3043	0.4943
	Chinese	0.4102	0.1728	0.3881

Table 1: Ablation study for different combinations of embeddings with MLP layers as the classification layer (early fusion).

extraction of richer semantic and contextual features, enhancing the system’s ability to make nuanced predictions for identifying ESG impact types in news articles. Figure 2 shows the schematic overview of the early fusion ensembling technique we employ.

By adopting both early and late fusion ensemble techniques, our system maximizes accuracy and robustness, offering a powerful solution for the task of ESG impact type identification in textual data.

4 Results

In the ML-ESG-2 shared task for ESG impact type identification, we evaluated the performance of our system across multiple languages using different ensemble configurations and fusion techniques. The micro F1, macro F1, and weighted F1 scores offer a comprehensive assessment of our system’s

effectiveness in capturing nuances within different linguistic contexts.

Our system secured notable rankings in the shared task, claiming the fourth position for English, the third position for both Japanese and Chinese, and the fifth position for the French language. We employed various ensemble configurations, including late fusion models combining FLAUBERT, mBERT, and ALBERT, incorporating TF-IDF and Latent Semantic Analysis (LSA) features. The results demonstrate significant improvements when transitioning to early fusion techniques, particularly when combining TF-IDF, LSA, and language model representations, as shown in Table 1. The late fusion techniques, adding more information with multiple models, seemed to help, as shown in Table 2.

In the English language, our best early fusion ensemble achieved a micro F1 score of 0.9633, a

Models	Language	Micro-F1	Macro-F1	Weighted-F1
FlauBERT + mBERT + ALBERT	English	0.9403	0.8899	0.9357
	French	0.525	0.501	0.4933
	Japanese	0.5155	0.2755	0.465
	Chinese	0.378	0.1346	0.3525
FlauBERT + mBERT + ALBERT + MLP (TF-IDF)	English	0.945	0.8944	0.9403
	French	0.530	0.505	0.5022
	Japanese	0.520	0.280	0.475
	Chinese	0.384	0.141	0.3589
FlauBERT + mBERT + ALBERT + MLP (TF-IDF + LSA)	English	0.9495	0.8991	0.9449
	French	0.535	0.512	0.5066
	Japanese	0.5244	0.2899	0.480
	Chinese	0.391	0.1474	0.3717

Table 2: Ablation study for different combinations of models (late fusion).

Team	Micro-F1	Macro-F1	Weighted-F1
AnakItik	0.9817	0.9548	0.9810
BrothFink	0.9771	0.9445	0.9765
NeverCareU	0.9633	0.9227	0.9648
FinNLU	<u>0.9633</u>	<u>0.9180</u>	<u>0.9639</u>
231	0.9633	0.9127	0.9627
SPEvFT	0.9587	0.9118	0.9602
LIPI	0.9312	0.8335	0.9294
HHU	0.9174	0.8098	0.9174

Table 3: Leaderboard for ESG identification in English language

macro F1 score of 0.918, and a weighted F1 score of 0.9639, showcasing its proficiency in ESG impact type identification. Similarly, for other languages, the early fusion ensemble, which integrates TF-IDF and LSA features with language models, achieved the highest performance among all early fusion and late fusion techniques.

While our ensemble system demonstrates commendable performance in English (as shown in Table 3), it is evident that its effectiveness diminishes when applied to other languages such as French, Japanese, and Chinese. This discrepancy underscores the importance of adapting our approach to address language-specific nuances and complexities. To remedy this, future research efforts may focus on language-specific model fine-tuning, dataset augmentation with more diverse linguistic content, and the development of specialized language models tailored to the unique challenges posed by each language. Additionally, incorporating more extensive language-specific features and linguistic resources into our ensemble configurations could

further enhance the system’s adaptability and robustness across diverse linguistic contexts. These measures hold the potential to improve our system’s performance and make it a valuable tool for multilingual ESG impact type identification.

5 Conclusion

In the rapidly evolving finance and corporate governance landscape, the ML-ESG-2 shared task has presented a unique opportunity to explore ESG impact type identification. Our comprehensive system leverages the collective intelligence of multiple language models and feature representations through advanced ensemble learning techniques, resulting in high performance. While our system excels in English, it faces hurdles in accurately identifying ESG impact types in Japanese, Chinese, and French languages. These challenges underscore the importance of further research and fine-tuning to adapt our approach to diverse linguistic contexts.

Limitations

While our ensemble learning approach has demonstrated strong performance, we acknowledge certain limitations in our study. The system’s performance variation across languages, with lower accuracy for Japanese, Chinese, and French texts, highlights the need for further language-specific model development and fine-tuning. Additionally, the reliance on single-source textual data alone may not capture all relevant context and nuances, which may affect the system’s ability to identify ESG impact types accurately in certain cases. Furthermore,

the generalization of our approach to other ESG-related tasks and domains may require additional adaptation and validation. We also recognize the evolving nature of language models and the need for continuous updates and refinements to maintain their effectiveness.

Ethics Statement

Throughout our research, we recognize that ethical considerations, particularly those related to bias mitigation and fairness in model predictions, represent ongoing and critical areas for improvement in the development of responsible AI systems. It is important to note that, in this study, we have not undertaken specific measures to actively mitigate bias within our models or predictions.

References

- Surabhi Adhikari, Surendrabikram Thapa, Usman Naseem, Hai Ya Lu, Gnana Bharathy, and Mukesh Prasad. 2023. Explainable hybrid word representations for sentiment analysis of financial news. *Neural Networks*, 164:115–123.
- Surabhi Adhikari, Surendrabikram Thapa, Priyanka Singh, Huan Huo, Gnana Bharathy, and Mukesh Prasad. 2021. A comparative study of machine learning and nlp techniques for uses of stop words by patients in diagnosis of alzheimer’s disease. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Maha Faisal Alsayegh, Rashidah Abdul Rahman, and Saeid Homayoun. 2020. Corporate economic, environmental, and social sustainability performance transformation through esg disclosure. *Sustainability*, 12(9):3910.
- Amir Amel-Zadeh and George Serafeim. 2018. Why and how investors use esg information: Evidence from a global survey. *Financial analysts journal*, 74(3):87–103.
- Prafulla Bafna, Dhanya Pramod, and Anagha Vaidya. 2016. Document clustering: Tf-idf approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE.
- Amina Buallay. 2019. Is sustainability reporting (esg) associated with performance? evidence from the european banking sector. *Management of Environmental Quality: An International Journal*, 30(1):98–115.
- 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.
- Susan T Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, 38:189–230.
- Stuart L Gillan, Andrew Koch, and Laura T Starks. 2021. Firms and social responsibility: A review of esg and csr research in corporate finance. *Journal of Corporate Finance*, 66:101889.
- Tushar Goel, Vipul Chauhan, Suyash Sangwan, Ishan Verma, Tirthankar Dasgupta, and Lipika Dey. 2022. **TCS WITM 2022@FinSim4-ESG: Augmenting BERT with linguistic and semantic features for ESG data classification**. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 235–242, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Mukul Jaggi, Priyanka Mandal, Shreya Narang, Usman Naseem, and Matloob Khushi. 2021. Text mining of stocktwits data for predicting stock prices. *Applied System Innovation*, 4(1):13.
- Rajaraman Kanagasabai, Saravanan Rajamanickam, Hariram Veeramani, Adam Westerski, and Kim Jung Jae. 2023. Semantists at ImageArg-2023: Exploring Cross-modal Contrastive and Ensemble Models for Multimodal Stance and Persuasiveness Classification. In *Proceedings of the 10th Workshop on Argument Mining*, Online and in Singapore. Association for Computational Linguistics.
- Jujeon 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.
- 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. -.
- 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. In *International Conference on Learning Representations*.
- Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2479–2490.

- Ting-Ting Li, Kai Wang, Toshiyuki Sueyoshi, and Derek D Wang. 2021. Esg: Research progress and future prospects. *Sustainability*, 13(21):11663.
- Elvys Linhares Pontes, Mohamed Ben Jannet, Jose G. Moreno, and Antoine Doucet. 2022. [Using contextual sentence analysis models to recognize ESG concepts](#). In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 218–223, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Fionn Murtagh. 1991. Multilayer perceptrons for classification and regression. *Neurocomputing*, 2(5-6):183–197.
- Graeme Newell and Muhammad Jufri Marzuki. 2022. The increasing importance of environmental sustainability in global real estate investment markets. *Journal of Property Investment & Finance*, 40(4):411–429.
- Ke Tian, Zepeng Zhang, and Hua Chen. 2022. [Automatic Term and Sentence Classification Via Augmented Term and Pre-trained language model in ESG Taxonomy texts](#). In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 224–227, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023a. Automated Citation Function Classification and Context Extraction in Astrophysics: Leveraging Paraphrasing and Question Answering. In *Proceedings of the second Workshop on Information Extraction from Scientific Publications*, Online. Association for Computational Linguistics.
- Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023b. DialectNLU at NADI 2023 Shared Task: Transformer Based MultiTask Approach Jointly Integrating Dialect and Machine Translation Tasks. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.
- Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023c. KnowTellConvince at ArAIEval 2023: Disinformation and Persuasion Detection using Similar and Contrastive Representation Alignment. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.
- Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023d. LowResContextQA at Qur’an QA 2023 Shared Task: Temporal and Sequential Representation Augmented Question Answering Span Detection. In *Proceedings of The First Arabic Natural Language Processing Conference (ArabicNLP 2023)*, Singapore. Association for Computational Linguistics.
- Hariram Veeramani, Surendrabikram Thapa, and Usman Naseem. 2023e. Temporally Dynamic Session-Keyword Aware Sequential Recommendation system. In *2023 International Conference on Data Mining Workshops (ICDMW)*.
- Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Author Index

- Apte, Manoj, 31
Arno, Henri, 11
- Baeck, Joke, 11
Billert, Fabian, 79
Brillet, Lucas Fernández, 62
- Caillaut, Gaetan, 1
Chen, Chung-Chi, 46
Chen, Hsin-Hsi, 46
CHERSONI, Emmanuele, 51
Conrad, Stefan, 79
- Day, Min-Yuh, 46
Demeester, Thomas, 11
- Ghosh, Sohom, 57
GU, Jinghang, 51
- HSU, Yu-Yin, 51
- Kang, Juyeon, 46
Kumaraguru, Ponnurangam, 57
- Lhuissier, Anaïs, 46
Liu, Jingshu, 1
- Mishra, Soumya, 72
Mulier, Klaas, 11
- Nakhlé, Mariam, 1
Naseem, Usman, 84
Naskar, Sudip, 57
Nishida, Shunsuke, 22
- Palshikar, Girish, 31
Parikh, Ankur, 42
Pawar, Sachin, 31
Pawde, Aditi, 31
PENG, Bo, 51
Polyanskaya, Anna, 62
- Qader, Raheel, 1
QIU, Le, 51
- Rajpoot, Pawan, 42
Ramrakhiyani, Nitin, 31
- Rodriguez Inserte, Pau, 1
- Seki, Yohei, 46
Septiandri, Ali, 66
- Tamura, Takuya, 22
Thapa, Surendrabikram, 84
Tseng, Yu-Min, 46
Tu, Teng-Tsai, 46
- Utsuro, Takehito, 22
- Vaishampayan, Sushodhan, 31
Vardhan, Harsha, 57
Veeramani, Hariram, 84
- Wang, Xiaotian, 22
Winatmoko, Yosef Ardhito, 66
- Zenimoto, Yuki, 22