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-AACL-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 2[1] LLM. Subsequently, we apply fine-tuning techniques to the Llama 2[2] and Dolly[3] 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.

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 72–78

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 limited 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 article[4] 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 service[5].

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 explorre 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/databrickslabs/dolly
[4] See 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
[9] https://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 Parabole.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, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

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)*.

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 Anthropic[1] 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:

- 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 an sample in the template's 'example' variable

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

[1] https://docs.anthropic.com/claude/docs/constructing-a-prompt

B Prompt Template using LangChain

The below template was used in experiments for semantic similarity approacha 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.