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 Library [1]. 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 FinBERT [2] 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

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 62–65

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-3.5[3] 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 Malakasiotis, 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.