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-

Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing, pages 51–56
November 1, 2023. ©2023 Association for Computational Linguistics

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".

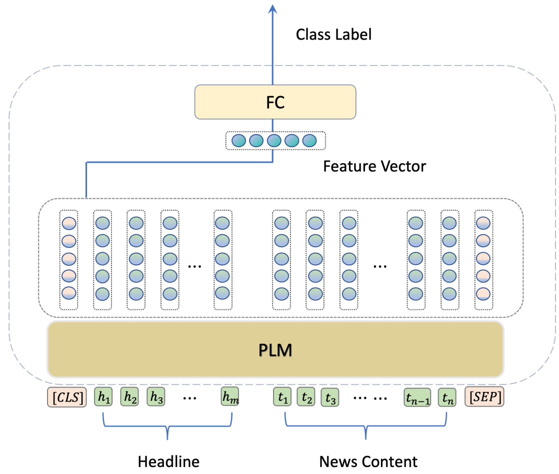
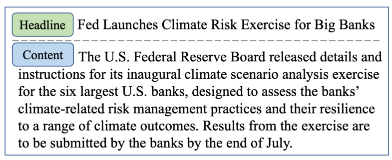


Figure 1: Model architecture



Headline Fed Launches Climate Risk Exercise for Big Banks

Content 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

[1]The value may vary by a small margin due to the preprocessing methods.

Content

各國代表也終於將2015年《巴黎協定》的規則手冊定案，包括關於全球碳市場機制的「第六條」（Article 6），將有助強化國際之間的減碳合作。 這份「格拉斯哥氣候協定」（Glasgow Climate Pact）在談判最後關頭遭到削減力度。針對「未使用碳捕捉技術的燃煤發電」，在中國與印度積極要求之下，將原本草案中的「分階段淘汰」（phase out）改為力度較低的「分階段削減」（phase down），引發歐洲與氣候脆弱國家的不滿。然而，這仍是首份提及化石燃料的聯合國氣候協議。 13日下午，大會主席沙瑪召開全員出席大會，盤點最新版本的草案內容，但會中各國談判代表仍充滿歧見。印度代表、環境部長亞達夫（Bhupender Yadav）率先開炮，認為針對協議所要達成的共識，「仍然難以找到」。 印度代表團要求，將關於「淘汰化石燃料與補貼」的字眼，從協議文本中刪除。亞達夫表示，這場氣候危機肇因於不永續的生活方式和資源浪費，暗批西方國家才是罪魁禍首。他強調，拿特定產業開刀是不必要的：「世界必須認清現實，化石燃料與其使用，讓世界一些地區，取得了高度的財富與幸福。」 中東產油國伊朗也支持印度對於化石燃料的立場，對於協議內容感到不滿。伊朗被視為發展中國家，國內生產毛額（GDP）約6820億美元（19兆新台幣），其中大部分來自於出口石油與天然氣。 稍早，瑞典氣候少女童貝里在推特抨擊氣候大會，表示「現在隨著COP26即將落幕，我們必須留意漂綠（greenwash）海嘯來襲，以及媒體編織謊言，將（談判）結果塑造為『良好』、『進步』、『帶有希望』或『往對的方向前進一步』」。 延伸閱讀：氣候峰會敲定碳市場規則 有助數兆美元釋出抗暖化 COP26達氣候協定 環保署重申：將調整2030減碳目標 ※本文授權自聯合新聞網，原文見此。

Summary

全球各國代表在《巴黎協定》下達成「格拉斯哥氣候協定」，強化國際間的碳減排合作。這包括有關全球碳市場的「第六條」，在最後關頭被削減力度。原草案中的「分階段淘汰」燃煤發電，應中印兩國的要求，被改為「分階段削減」，引發不滿。雖然有歧見，這是首份提及化石燃料的聯合國氣候協定。印度代表亞達夫對協定內容提出異議，要求刪除「淘汰化石燃料與補貼」的字眼。他指責西方國家的生活方式和資源浪費為氣候危機的主因，認為特定行業無須進行調整。他認為化石燃料帶來財富和幸福，伊朗等國亦贊成此觀點。瑞典氣候活動家童貝里則批評協定，警告COP26將帶來「漂綠」風潮，及媒體將談判結果美化的做法。文章相關訊息來源自聯合新聞網。

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.	*AnakItik*'s [2]	0.9817	0.9548	0.9810
	BrothFink's	0.9771	0.9445	0.9765
	NeverCareU's	0.9633	0.9227	0.9648
	Ours	0.9633	0.9127	0.9627
		0.9633	0.9096	0.9620
		0.9633	0.9096	0.9620
Ch.	*LIPI*'s	0.6859	0.5279	0.6773
	LIPI's	0.7564	0.4585	0.7321
	LIPI's	0.6731	0.2897	0.6508
	Ours	0.8654	0.7325	0.8686
		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,

[2] The 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 + *sandwiched* 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. Dynamicesg: 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.