BigBird for ESG Reports - A Sparse Attention Model For Differences Between Languages In Textual Analysis

Hunter Ng
Baruch College, City University of New York
hunterboonhian.ng@baruch.cuny.edu

Shuai Xu
Shanghai University of Finance and Economics
Singapore Management University
shuai.xu.2023@phdacc.smu.edu.sg

September 15th, 2024

Abstract

This paper introduces BigBird-ESG, a domain-specific transformer architecture pre-trained on manually classified paragraphs from Chinese ESG reports between 2016 to 2020. Our results show that BigBird-ESG, with its sparse attention mechanisms, more efficiently processes Chinese ESG reports due to innate qualities of the Chinese language and its difference from the English language. We show that BigBird-ESG outperforms BERT and FinBERT under specific conditions in both sentiment and category classification tasks. The findings suggest that language specificity affects the accuracy of LLM models in parsing textual data. Our findings affect the future advancement of multi-modal LLMs and transfer learning, which should consider language-specific qualities when interpreting sources of financial information. We also use state-of-the-art OCR to coherently extract paragraphs from the ESG reports, which preserves the meaning of the textual data and we use this new methodology to show that tone in Chinese ESG reports is correlated with ESG ratings.

Keywords: BigBird, Large Language Model (LLM), Natural Language Processing (NLP), Environmental, Social, and Governance (ESG), Chinese ESG Reports, Paragraph Parsing, Sentiment Classification, Domain-Specific Model, Transformer Models, Financial Text Analysis

JEL Classifications: C45, G32, M14, M41, Q56

1. Introduction

Finance and accounting research has long been at the forefront of exploring natural language processing (NLP) algorithms to analyze vast amounts of textual data (Li 2010; Loughran and McDonald 2016). Recently, advancements in large language models (LLMs) have introduced new methods that fully leverage cutting-edge NLP techniques to capture contextual relationships within text, establishing them as the gold standard in the field (Huang et al., 2022).

In this study, we ask the question of whether processing non-English financial or non-financial information requires understanding of the other language's features. In particular, we focus on ESG reports, which have been the subject of intense academic interest (Christensen et al., 2021). We further focus on Chinese ESG reports as China-related research has mirrored the country's increased importance on the global stage. (Lennox and Wu, 2022). ESG reports provide an information-rich source of accounting information and understanding its effects on financial indicators has been a cornerstone of ESG research (Christensen et al., 2021).

Hence, we introduce BigBird-ESG, a sparse attention transformer architecture specifically pre-trained on ESG-related paragraphs extracted from Chinese corporate ESG reports between 2016 to 2020. BigBird is designed for processing long texts and is particularly suitable for analyzing ESG reports, which often feature dense, lengthy disclosures (Zaheer et al. 2020). Previous research has demonstrated that standard transformer models, such as BERT, struggle with long documents due to their quadratic complexity versus input length ratio (Vaswani et al. 2017). BigBird mitigates this issue through its sparse attention mechanisms, allowing it to efficiently handle longer inputs while maintaining high performance.

Using a dataset of 25,740 manually labeled paragraphs from Chinese ESG reports, we evaluate BigBird-ESG's performance in classifying sentiment, as well as environmental, social, and governance-related content. We compare its accuracy and effectiveness against classic NLP methods such as Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM) neural network architecture and support vector machines (SVM). Our results demonstrate that BigBird-ESG outperforms these algorithms in identifying nuanced ESG topics in Chinese corporate disclosures. We perform several robustness test to demonstrate the incremental performance of BigBird-ESG and present linguistic evidence that parsing of Chinese language documents or other non-English documents may require different architectures which take into account the language's features. In particular, in this paper, our evidence is in line with the theory that the Chinese language's complexity and context-dependency make it a suitable candidate for different attention models such as BigBird.

To showcase BigBird-ESG's improved performance, we measure it across standard indicators of performance. In further robustness test, we translate the Chinese ESG reports to English using ChatGPT, a state-of-the-art artificial intelligence that has proven to be more superior than traditional translation techniques (Hendy et al., 2023). We then process the translated reports using FinBERT. We show that BigBird-ESG improves over FinBERT under specific conditions, which shows that some meaning may be lost in translation and that it may be better to work with the original language when processing accounting information such as ESG reports.

The contributions of this study are twofold. First, we manually label a dataset of 25,740 paragraphs from Chinese ESG reports and use this to train an implementation of BigBird, which is a modification of the original transformer model such as BERT or GPT. We document its superiority over existing NLP methods due to its suitability to the context of Chinese ESG reports and thus, highlight the importance of model adaptation to specific financial contexts, particularly

in emerging areas such as ESG reporting in countries where the main language is not English. In tandem with the advent of transfer learning and RAG architecture, the paper's results of the BigBird model is a step in a future world where multi-modal LLM may one day be the norm (Shen et al., 2024). In other words, as the future LLMs continue to develop, their multi-modal nature may encompass many smaller LLMs, of which this paper contributes in terms of a dataset and contextual methodology.

Secondly, our findings offer practical implications for policy-makers and regulators who often rely on company filings to make policy decisions. With regulators increasing use of machine learning and technology in assessing regulatory risks (Bauguess, 2017), there is the constant need to improve on existing models to efficiently process large amounts of data. Through understanding language differences and increasing the accuracy of textual analysis of reports, policy-makers can extract higher quality information from alternative sources of accounting information.

The rest of the paper is as follows. Section 2 introduces the linguistics literature and transformer architecture history, as well as our main hypotheses. Section 3 presents our methodology and empirical findings. Section 4 presents our conclusion and future work to be done.

2. Background

2.1 Chinese versus English Language ESG Reports - Complexity and Context-Dependency

Textual analysis has been an important tool in the finance and accounting literature to extract information from firm disclosures (Loughran and McDonald, 2014; Li, 2008; Brown and Tucker, 2011). It continues to improve with the proliferation of artificial intelligence (AI) and LLMs. However, many of these techniques are used for English textual data. Although they can be implemented for other languages, such as building a textual dictionary using the Loughran and McDonald approach, other LLM methods that are trained on corpuses on English data may not be easily transferred to other languages. Thus, in using the setting of Chinese ESG reports, we document several distinct features of the Chinese language which would warrant a different approach in textual parsing compared to English language reports.

Chinese, as a logographic language, encapsulates substantial meaning within individual characters, each representing complex words or ideas. Unlike English, where words are formed by combining letters that individually carry limited semantic weight, Chinese characters are inherently rich in meaning. For example, the character "心" embodies the concepts of "heart" or "mind" in a single symbol, conveying complex ideas with just one character. This quality allows Chinese text to convey extensive information and nuanced concepts in relatively few words, adding layers of complexity to its interpretation.

Moreover, Chinese grammar often relies heavily on context, omitting elements like articles, auxiliary verbs, and sometimes even pronouns or tense markers that are mandatory in English. For example, the English sentence "He is going to the zoo" is expressed in Chinese as "他去动物

元," which depends on contextual understanding to convey the full meaning. This high degree of contextual dependency and representation of ideas contribute to the linguistic complexity of Chinese, making it a challenging language for tasks such as natural language processing and machine translation, where capturing subtle meanings is crucial.

Due to these linguistic differences, sentence-level parsing in Chinese may be less effective than paragraph-level parsing for NLP tasks. The two properties of complexity and context-dependency above show that paragraph-level parsing in Chinese may be more appropriate for capturing the full breadth of meaning, reflecting the language's compact and context-driven nature.

2.2 Structural Integrity of Text from ESG Report

Thus, to develop our model, we need to preserve the structural integrity of textual information from the ESG reports. This is opposed to simple counts of positive versus negative words, which do not require structural integrity from the words. The extraction of textual data from ESG reports is a non-trivial task. A ESG report contains many tables, pictures, graphs, headers and other non-structured data. Unlike the more structured and standardized narrative sections of MD&A disclosures from 10-K filings, ESG reports often lack a uniform format, complicating the extraction of coherent, structured text for further analysis.

To address these challenges, we utilize Tesseract, an open-source OCR engine, which allows for the preservation of paragraph-level coherence during text extraction. Tesseract's advanced layout analysis helps retain the spatial structure of paragraphs, thus maintaining the semantic integrity crucial for NLP tasks. By ensuring that paragraphs remain intact, we can preserve the original narrative flow, which is essential for capturing the contextual meaning in subsequent textual analyses. This structural preservation is even more valuable when we use transformer architectures or LLMs such as BERT or AI, which can parse the context of the content. For information-rich ESG reports, the narrative coherence of sections may further provide critical insights into a company's ESG practices.

2.3 Difference between BERT, FinBERT and BigBird

In the context of Natural Language Processing (NLP) for financial reporting and ESG data, transformer-based models like BERT, FinBERT and BigBird offer distinct advantages. These models belong to the transformer architecture family, but they diverge in their design and underlying mechanisms, leading to varying capabilities in handling large-scale text data.

Firstly, BERT (Bidirectional Encoder Representations from Transformers) is a foundational transformer-based model developed by Google (Devlin et al., 2018). It employs a bidirectional training approach, allowing it to consider the context from both preceding and succeeding words in a sentence, which enhances its understanding of language nuances. When processing Chinese text, BERT base models employ a different tokenization strategy due to the Chinese's unique characteristics. Unlike English, which uses spaces to separate words, Chinese text is continuous, and each character represent a distinct meaning. BERT handles this by treating each Chinese

character as an individual token, effectively using character-level tokenization. This approach allows the model to capture the nuances of the language without relying on word boundaries. However, this can lead to longer token sequences for Chinese documents, especially in the context of financial reports and ESG disclosures that often contain complex and specialized terminology. Like the English version, BERT's maximum sequence length is limited to 512 tokens, which poses a challenge for processing longer Chinese texts.

Next, FinBERT, which is a deep learning model using the transformer architecture has been increasingly cited and used by researchers to parse financial reports (Huang et al., 2023; Yang et al., 2020). FinBERT uses a large, expert-labeled training set and it is also hosted on the python library for use as an API. Its credibility and accessibility make it an invaluable tool for researchers and policy-makers. FinBERT is trained on a large corpus of English financial textual information, such as analyst reports, earnings calls, and financial news. It offers sentiment analysis and ESG classification of sentences. Its key strength lies in its financial domain adaptation and has proven to be more effective than BERT when dealing with financial information (Huang et al., 2023). However, it does not currently have cross-language capabilities to parse textual information in other languages. Additionally, it is also limited to 512 tokens, making it less suitable for analyzing longer texts.

In contrast, BigBird (Big Transformer with Sparse Attention) is designed to overcome the problems of processing long sequences. BigBird introduces sparse attention mechanisms that reduce the complexity from quadratic to linear. This enables BigBird to efficiently handle documents with thousands of tokens, making it an ideal choice for processing large ESG reports that may contain multiple paragraphs of detailed disclosures. Moreover, BigBird's sparse attention mechanism allows it to capture long-range dependencies and contextual relationships within longer paragraphs in ESG reports, which can further preserve the coherence and structure of the text. It is especially relevant to the Chinese context because of the complexity and context-dependency, which can lead to long paragraphs that relate to one single topic.

2.4 Multi-modal LLMs

From the last section, we see that there are a variety of models that parse textual data. They do not conflict with each other because the future of large language models (LLMs) is trending towards multi-modal architectures that integrate various specialized models, each optimized for different types of inputs or specific tasks (Shen et al., 2024). Rather than relying on a single, monolithic model, future LLMs are likely to consist of several smaller, task-specific models that work in unison to provide contextually rich responses across multiple domains and languages. This modular approach allows for the combination of models that are fine-tuned to handle different challenges.

In this context of non-English accounting information, the BigBird methodology offers significant advantages over specialized models like FinBERT, especially when extended into languages such as Chinese. For instance, FinBERT has proven effective for financial analysis in English but is limited by its quadratic complexity and its inability to efficiently process long sequences

of text. BigBird, with its sparse attention mechanism, offers a more scalable solution for longer texts, which is useful for extracting information from longer paragraphs in languages such as Chinese, where the character-based structure introduces complexity.

By understanding the performance of different attention models in the Chinese setting, this paper contributes to the future of multi-modal LLMs. The ability to manage lengthy documents, such as ESG reports or financial disclosures, in multiple languages will is paramount in a rapidly evolving AI landscape. As multi-modal LLMs grow in popularity, models like BigBird that are optimized for long-document processing can be integrated as a specialized component within a larger architecture. Multi-modal systems can combine models like BigBird for non-English analysis or other purposes and domain-specific models like FinBERT for financial text understanding. This combinatory approach has been documented in the computer science and AI literature (Shen et al., 2024).

2.5 Sentiment Analysis and ESG Classification

For this paper, we propose to train and compare several models for sentiment analysis and classification of ESG-related text. The goal of this analysis is to assess the performance of these models in extracting sentiment from ESG reports and classifying content based on Environmental, Social, Governance (ESG) or None categories. Thus, we provide our hypotheses in the alternate forms below.

H1a: BigBird will outperform BERT and other transformer-based models in terms of accuracy for sentiment classification in Chinese ESG reports.

H1b: BigBird will outperform BERT and other transformer-based models in terms of accuracy for ESG classification in Chinese ESG reports.

The rationale behind this hypothesis stems from BigBird's sparse attention mechanism, which is better suited for processing long-form, unstructured text that is typical of ESG disclosures. Unlike FinBERT, which is constrained by its 512-token limit, BigBird's ability to handle thousands of tokens allows it to capture the contextual nuances and long-range dependencies within lengthy ESG reports. This structural advantage should lead to more accurate sentiment classification and superior performance on ESG-related tasks.

Next, we further develop our next hypothesis in the alternate form as follows.

H2: Chinese ESG report sentiment is positively correlated to ESG ratings.

$$\text{rating} = \alpha + \beta_1 \text{tone}_i + \text{controls} + \varepsilon$$

Rating refers to the esg ratings and tone is measured using the various NLP models. This hypothesis is based on the theory that tone and sentiment expressed in ESG reports can be predictive of a company's overall ESG ratings. Prior literature has indicated that ESG sentiment is linked to ESG performance and ratings in the Chinese setting (Sun et al., 2024) but they rely on classic methods of document parsing using textual dictionaries, which have been proven to be less effective compared to new BERT models (Huang et al., 2022). Thus, we test this hypothesis

using our methodology or extracting coherent paragraphs and analyze using different, newer NLP methods.

3. Data and Results

3.1 Chinese ESG Reports

For this study, we start with Chinese ESG reports from 2016 to 2020. 2016 is the first year that Chinese firms started issuing ESG reports[1]. We collect the ESG reports from Juchao Information Network, the ESG ratings from the iFinD finance database and the Wind Financial Terminal, and their corresponding financial numbers from China Stock Market Accounting Research (CSMAR).

After collecting the ESG reports, we use Tesseract 4, a state-of-the-art optical character recognition (OCR) software to extract text from the ESG reports. We first convert each page of the ESG report into a 300dpi tiff file, which is a lossless type of image format to preserve as much visual content as we can from the ESG reports. We then run Tesseract 4, which uses a new neural net (LSTM) based OCR engine and improves over Tesseract 3 (Weil et al., 2024). Tesseract also supports many languages and we use the simplified Chinese setting. The detailed specifications are shown in Appendix 1.

Next, we clean up the extracted paragraphs. As we are interested in the textual data, we set up filters such that (1) only rows with more than an arbitrary four Chinese characters and (2) the text chunk has at least two consecutive rows. This is how a typical paragraph is displayed in a Chinese ESG report.To further eliminate redundancies, if a ESG report has more than 10% of its pages with all blank or one-character rows, we eliminate the ESG report from our sample. We end up with 111 ESG reports with 25,740 paragraphs. After this cleaning, we still end up 109 paragraphs that were not extracted coherently and consists of random Chinese characters. We kept these paragraphs in our dataset to build robustness to the models that we train.

Next, we manually label the paragraphs in two parts. The first is to label whether a paragraph is positive(乐观), neutral(中立) or negative(悲观). The second is to label whether a paragraph relates to environmental(环境), social(社会), governance(管理) or none (都不是). After labelling, we proceed to train the models using varying levels of training and test ratios.

We report our results for the sentiment and ESG analysis in Tables 1 and 2. To compare the accuracies of the models, the NLP literature often uses four objective metrics. (1) Accuracy measures the correctness of the model through the ratio of correct predictions to total predictions. (2) Recall measures the sensitivity of the model. (3) Precision measures the relevance of the model. (4) F1 score measures the harmonic mean of precision and recall and balances the two.

[1]Prior to 2016, they issued CSR and sustainability reports, which are not the subject of this paper.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FN}$$

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP is True Positive, TN is True Negative, FP is False Positive and FN is False Negative. We also provide additional metrics explained in Appendix.

3.2 Comparisons between the NLP Models

In this study, we train four separate models, namely BigBird-ESG, BERT, LSTM and SVM. Past finance and accounting studies have documented that machine learning algorithms such as BERT, LSTM and SVM perform well on tasks involving financial text (Brown et al. 2024). Huang et al. (2022) further show that a LLM customized for financial texts leads to further improvement or benefits.

To train our models, we split the dataset into train and test sets of three different percentages of 80-20, 60-40 and 40-60. For the Big Bird model, we use standard parameters based on best practices in the literature (Devlin et al., 2019; Sun et al., 2019). We use a learning rate of $2e^{-5}$, batch size of 16, 3 epochs and 0.01 weight decay. More details can be found in the Appendix. For the Big Bird model, we use the Big Bird tokenizer *chinese-bigbird-wwm-base-4096* provided by Li (2024), which is Whole Word Masking (WWM). WWM offers improvements over standard masking in the Chinese language due to innate differences between Chinese and English (Cui et al., 2021; Dai et al., 2022). For the BERT and other models, we follow standard specifications. The use of the *bert-base-chinese* is also fair towards BERT and LSTM because both these tokenizers have similar specifications of about 105-110M parameters and other attributes. More details can be found in the Appendix.

Table 1 Sentiment classification performance of BigBird, BERT, LSTM and SVM. All models are based on the standard 80-20 train-test split. Parameters are detailed in the Appendix. Other variable definitions are detailed in Section 3.1.

Model	Accuracy	Precision	Recall	F1 score	Positive (Recall)	Neutral (Recall)	Negative (Recall)
BigBird	0.8803	0.8866	0.8803	0.8808	0.9170	0.8641	0.3214
BERT	0.8783	0.8840	0.8783	0.8785	0.9019	0.8698	0.3214
LSTM	0.8285	0.8223	0.8285	0.8252	0.7595	0.8777	0.0000
SVM	0.6933	0.6804	0.6933	0.6721	0.4107	0.8680	0.0000

Table 2 ESGN classification performance of BigBird, BERT, LSTM and SVM. All models are based on the standard 80-20 train-test split. Parameters are detailed in the Appendix. Other variable definitions are detailed in Section 3.1.

Model	Accuracy	Precision	Recall	F1 score	Environmental (Recall)	Social (Recall)	Governance (Recall)	None (Recall)
BigBird	0.8668	0.8689	0.8668	0.8665	0.9448	0.8881	0.8496	0.7985
BERT	0.8596	0.8598	0.8596	0.8593	0.9284	0.8783	0.8319	0.8098
LSTM	0.7363	0.7409	0.7363	0.7376	0.7607	0.7116	0.7089	0.7639
SVM	0.5274	0.5743	0.5274	0.5226	0.4264	0.4444	0.4199	0.7768

Table 3 ESGN classification performance of Accuracy measure for BigBird, BERT, LSTM and SVM across different train-test split ratios. The header percentages represent the training sample.. Parameters of the models are detailed in the Appendix.

Training Sample	80%	60%	Difference (80%-60%)
BigBird	0.8668	0.8576	0.0092
BERT	0.8596	0.8531	0.0065
LSTM	0.7392	0.7363	0.0029
SVM	0.5274	0.4862	0.0964

3.3 Does Content of Chinese ESG Reports determine ESG Ratings?

Next, we perform fixed effect OLS regression to determine if the textual analysis. We present these results in Table XX and see that from Column (1), when we use the original sentiment classification done manually, we find evidence that tone is correlated with ESG performance in the form of ratings.

Our results echo literature that find that positive ESG reporting tone has correlations with ESG performance and ratings, both for English ESG reports and Chinese ESG reports (Sun et al., 2024;). We differentiate from these findings as we use a context-coherent extraction method using advanced OCR and we further find the sentiment using BigBird-ESG, which has improved accuracy over traditional methods such as text dictionaries or count of positive versus negative words. The results also provide further credibility to our main empirical methods of extracting text in the form of context-coherent paragraphs as it is in line with the classic view that disclosure tone is correlated with ESG performance.

Table 4 OLS Regression of ESG Ratings on sentiment and other control variables.

	(1)
	Original Classification
Constant	8.29
	(0.339)
neutral	0.0258***

	(0.00929)
positive	−0.0365**
	(0.0263)
negative	−0.158
	(0.672)
size	1.71***
	(3.09E-05)
lev	−9.94***
	(0.00613)
ROA	−0.0217
	(0.954)
Accruals	0.0192
	(0.741)

3.4 Comparison with FinBERT

To further determine the efficacy of BigBird-ESG, we compare it to FinBERT. FinBERT is a specialized variant of the BERT model that is trained on a large corpus of financial textual data that is expertly labelled. FinBERT is trained on english texts and meant to process sentences. It is widely cited and used in sentiment analysis as well as ESG classification tasks (Huang et al., 2023).

We first translate our extracted paragraphs to English. We do so using start-of-the-art translation provided by ChatGPT. Using an API, we translate using the latest version *chatgpt-4o-mini* as of this writing. For texts that are untranslatable, we drop these rows. We then use *yiyanghkust/finbert-tone* and *yiyanghkust/finbert-esg*, while remapping their original labelling to our dataset's labelling. The comparison between BigBird-ESG and FinBERT is shown in Table XX.

From Table XX, we see that FinBERT is able to classify the sentiment. We caution that the task may be unfair to FinBERT because FinBERT imposes a maximum of 512 tokens to its input, thus, we have cut off rows that exceed this token limit. This may have resulted in a lot of lost meaning, which show in the ESG classification compared to the sentiment classification. FinBERT is also trained primarily on proper English sentences in financial data, while our paragraphs are extracted from Chinese ESG reports, where there may be slight for. The lower classification rates may also be due to translation inefficiencies. We caution against overly relying on this result to say that FinBERT is ineffective for translated data, but rather, models that are trained on the data in their original languages and take into account the language's properties can be more effective.

Table 5 Sentiment classification performance of FinBERT vs BigBird-ESG. The FINBERT scores are derived from AI translation from Chinese to English using *chatgpt-4o-mini* and untranslatable portions are not used. Other variable definitions are detailed in Section 3.1.

Model	Accuracy	Precision	Recall	F1 score	Positive (Recall)	Neutral (Recall)	Negative (Recall)

FinBERT	0.7483	0.7491	0.7483	0.7467	0.6105	0.8344	0.3214
BigBird	0.8803	0.8866	0.8803	0.8808	0.9170	0.8641	0.3214

4. Conclusion

Our study makes three key contributions. Firstly, we introduce BigBird-ESG in the context of Chinese ESG reports, representing an improvement over current BERT methods. We make BigBird-ESG available for researchers to classify ESG reports into sentiment categories and environmental, social, or governance (E, S, or G) categories.

Secondly, we empirically show that BigBird-ESG outperforms BERT and FinBERT under specific conditions in processing Chinese ESG reports. We attribute this to specific linguistic attributes of the Chinese language compared to English. In particular, in the Chinese setting, the complexity and context-dependence of the language allows for BigBird to derive incremental performance. We also show the use of advanced OCR to coherently extract paragraphs from the ESG reports.

Lastly, we show the advantage of using a language-specific transformer architecture over traditional models. In a future where multi-modal LLMs become the norm, incremental improvements in smaller LLM models can be added to overall bigger models. This has implications on policy-makers and stakeholders who look towards maximizing information extracted from different sources of financial and accounting data.

References

Bauguess, S. (2017). SEC.gov | The Role of Big Data, Machine Learning, and AI in Assessing Risks: A Regulatory Perspective. https://www.sec.gov/newsroom/speeches-statements/bauguess-big-data-ai

Brown, S. V., & Tucker, J. W. (2011). Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications. Journal of Accounting Research, 49(2), 309–346. https://doi.org/10.1111/j.1475-679X.2010.00396.x

Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. Review of Accounting Studies, 26(3), 1176–1248. https://doi.org/10.1007/s11142-021-09609-5

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 3504–3514.

Dai, Y., Li, L., Zhou, C., Feng, Z., Zhao, E., Qiu, X., Li, P., & Tang, D. (2022). "Is Whole Word Masking Always Better for Chinese BERT?": Probing on Chinese Grammatical Error Correction (arXiv:2203.00286). arXiv. http://arxiv.org/abs/2203.00286

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (arXiv:1810.04805). arXiv. http://arxiv.org/abs/1810.04805

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation (arXiv:2302.09210). arXiv. https://doi.org/10.48550/arXiv.2302.09210

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. Contemporary Accounting Research, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

Lennox, C., & Wu, J. S. (2022). A review of China-related accounting research in the past 25 years. Journal of Accounting and Economics, 74(2), 101539. https://doi.org/10.1016/j.jacceco.2022.101539

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics, 45(2–3), 221–247.

Li, L. (2024). LowinLi/chinese-bigbird [Python]. https://github.com/LowinLi/chinese-bigbird (Original work published 2021)

Loughran, T., & Mcdonald, B. (2014). Measuring Readability in Financial Disclosures. The Journal of Finance, 69(4), 1643–1671. https://doi.org/10.1111/jofi.12162

Shen, W., Li, C., Chen, H., Yan, M., Quan, X., Chen, H., Zhang, J., & Huang, F. (2024). Small LLMs Are Weak Tool Learners: A Multi-LLM Agent (arXiv:2401.07324). arXiv. https://doi.org/10.48550/arXiv.2401.07324

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu (Eds.), Chinese Computational Linguistics (Vol. 11856, pp. 194–206). Springer International Publishing. https://doi.org/10.1007/978-3-030-32381-3_16

Sun, Y., Zhao, D., & Cao, Y. (2024). The impact of ESG performance, reporting framework, and reporting assurance on the tone of ESG disclosures: Evidence from Chinese listed firms. Journal of Cleaner Production, 466, 142698. https://doi.org/10.1016/j.jclepro.2024.142698

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention Is All You Need. arXiv.Org. https://arxiv.org/abs/1706.03762v7

Yang, Y., UY, M. C. S., & Huang, A. (2020). FinBERT: A Pretrained Language Model for Financial Communications (arXiv:2006.08097). arXiv. http://arxiv.org/abs/2006.08097

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., & Yang, L. (2020). Big bird: Transformers for longer sequences. Advances in Neural Information Processing Systems, 33, 17283–17297.

Appendix 1A

Other Definitions

True Positive (TP)	Metric is correctly classified as the specific class (e.g., correctly classified as positive OR E correctly classified as E).
True Negative (TN)	Metric is correctly classified as not belonging to the class (e.g., correctly classified as negative or neutral OR something not E is classified as not E)
False Positive (FP)	Metric is incorrectly classified as the class (e.g., incorrectly classified as positive when it should be negative or neutral OR classified as E but actually S, G, or N)
False Negative (FN)	Metric is incorrectly classified as not being the class. (e.g., actually positive but classified as negative, OR actually E but classified as S, G, or N))

Appendix 1b

Explainer on BigBird architecture

BigBird is an innovative extension of the transformer architecture that addresses the limitations of processing long sequences in models such as BERT (Zaheer et al., 2020). In BERT, the self-attention mechanism requires every token in a sequence to attend to every other token. This self-and-full attention approach captures extensive contextual relationships but the complexity scales quadratically with the sequence length. In other words, the computational power requirement is extremely intensive when text chunks are long. As a result, processing longer documents becomes more difficult.

BigBird solves this via a sparse attention mechanism that reduces the complexity from quadratic to linear, with respect to the sequence length. Instead of having every token attend to all others, BigBird employs a combination of global, random, and local (window) attention patterns.

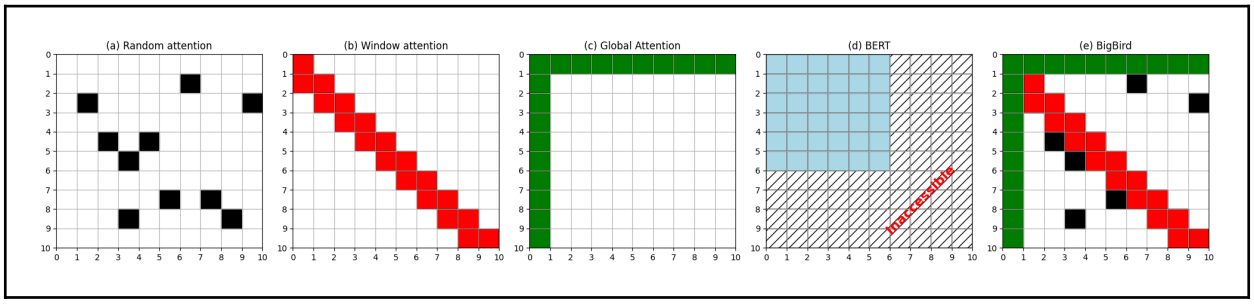

Figure A1. A comparison of the five types of attention models using a visualization. (a) shows a random attention model, where grids are randomly selected to pay attention to. (b) shows a window attention model, where grids are chosen along a diagonal line. (c) shows a global attention model, where grids are chosen along the borders. (d) shows the BERT full attention, which uses all available grids but can only cover a smaller maximum area due to its computational cost requirements. (e) shows the Big Bird model, where the first 3 models are combined to form a sparse attention model. The diagrams are adapted from Zaheer et al., 2020.

The three types of separate attention models are explained here. Firstly, random attention allows each token to attend to a fixed number of randomly selected tokens throughout the sequence. This introduces long-range connections between tokens that might be distant from each other, enabling the model to capture dependencies without the need for full attention across all token pairs. The randomness allows for effective learning of complex patterns.

Secondly, local or window attention means that each token attends to its immediate neighbors within a predefined window size. This maintains the local context and sequential information essential for understanding language.

Lastly, global attention uses a select set of tokens, designated as global tokens, that attend to all other tokens in the sequence and are also attended to by them. These tokens often include special classification tokens that are crucial for capturing overall context. This allows important information to flow across the entire sequence.

By combining these three types of attention, BigBird creates a sparse attention pattern that reduces computational demands while preserving the model's ability to learn from long sequences. The sparse attention mechanism allows BigBird to process sequences that are longer than what BERT can manage, without substantial loss in performance.

Quadratic Complexity of Full Self-Attention

In BERT, for a sequence of length n, each of the n tokens computes attention scores with all n tokens, resulting in $n * n = n^2$computations. The model needs to store the attention weights for all token pairs, which also scales with n^2. This is in contrast to n linear computations for Big Bird.

Maximum Sequence Length

BERT models are typically configured with a maximum sequence length (e.g., 512 tokens). Texts longer than this limit must be cut off, which can lead to loss of contextual information. This is in contrast to Big Bird which can exceed this token limit.

In summary, BigBird represents a significant advancement in transformer models by addressing the scalability issues in BERT. By effectively capturing both local and global dependencies without incurring large computational costs, BigBird is more likely to outperform BERT in scenarios where the main idea is distributed across longer texts. In this paper, BigBird is shown empirically to outperform BERT for Chinese ESG reports.

Appendix 2

Examples of paragraphs extracted from Chinese ESG Reports

012 CIMC

可持續發展管理

與利益相關方溝通

本集團重視與利益相關方關係。我們的主要利益相關方主要包括投資者、監管機構、員工、客戶、供應商和社區等。通過公司公告、年報、社會責任報告、意見調查、會議等渠道，我們與利益相關方保持持續溝通。下表簡要列出我們與各利益相關方的溝通方法及他們所關注的議題。

利益相關方	溝通方法	議題
投資者	股東會、年報、公告、券商投資年會	公司運營、企業管治
客戶	滿意度調查、投訴受理機構	產品質量、產品創新、客戶服務
供應商	定期會議、供應商評審	誠信合作、公平競爭、安全生產
員工	員工文化活動、員工培訓、員工懇談會、內部通訊	福利待遇、安全健康、培訓發展
政府、監管機構	定期會議、政策宣導、合作培訓	合規運營

未來，我們將繼續關注各利益相關方的訴求，逐步加強溝通，以及回應他們關注的議題，藉此審視我們可持續發展的管理方法，開展針對性工作。

Figure A2. A screenshot of page 12 of CIMC's ESG report in 2016. This page describes the company's communication with its various stakeholders and the feedback from them.

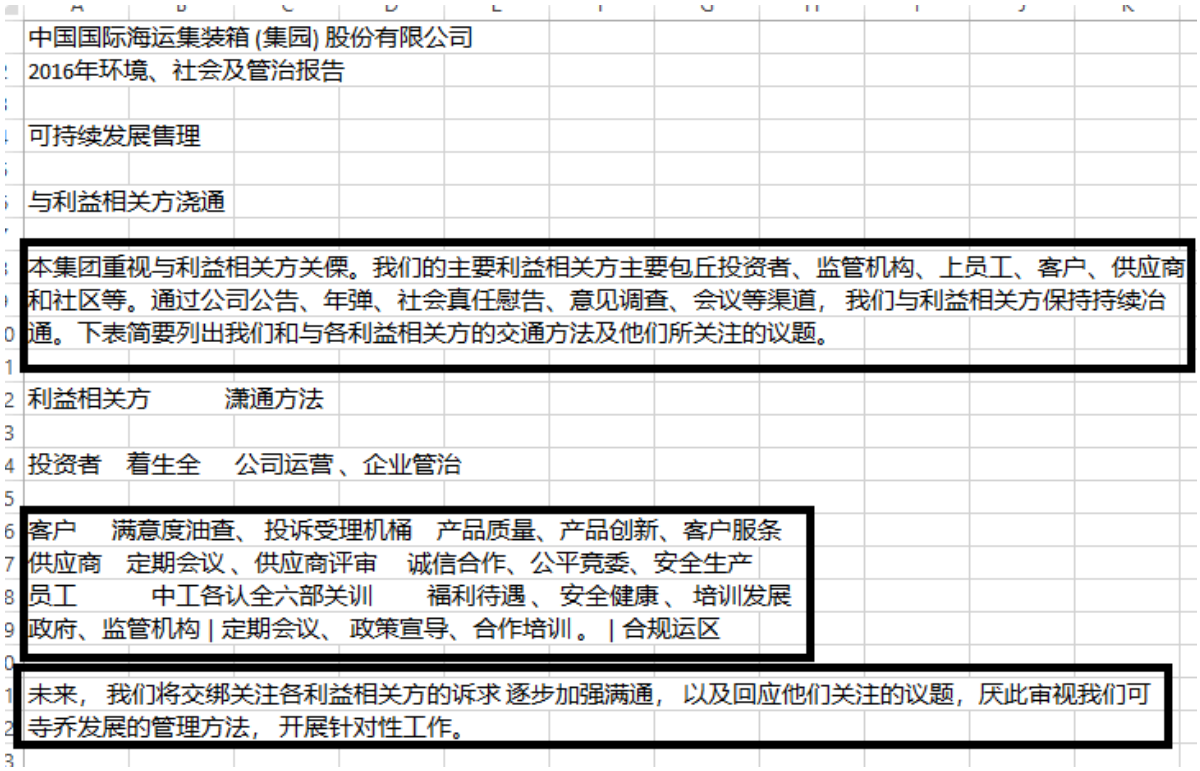

Figure A3. The previous screenshot is converted using Tesseract to sentences which preserve the coherence of the information in the ESG report page. While it is not 100% accurate, three main text chunks have been correctly identified, which are the main contents of this page. The other content areas are discarded as they do not have more than 1 consecutive row, and the title, which repeat on the other pages, is also eliminated.

Figure A4. Another screenshot of page 3 of another company, GCL-ET's ESG report in 2020. This page describes an introduction to the company's business and ESG strategy, and several key metrics of the company.

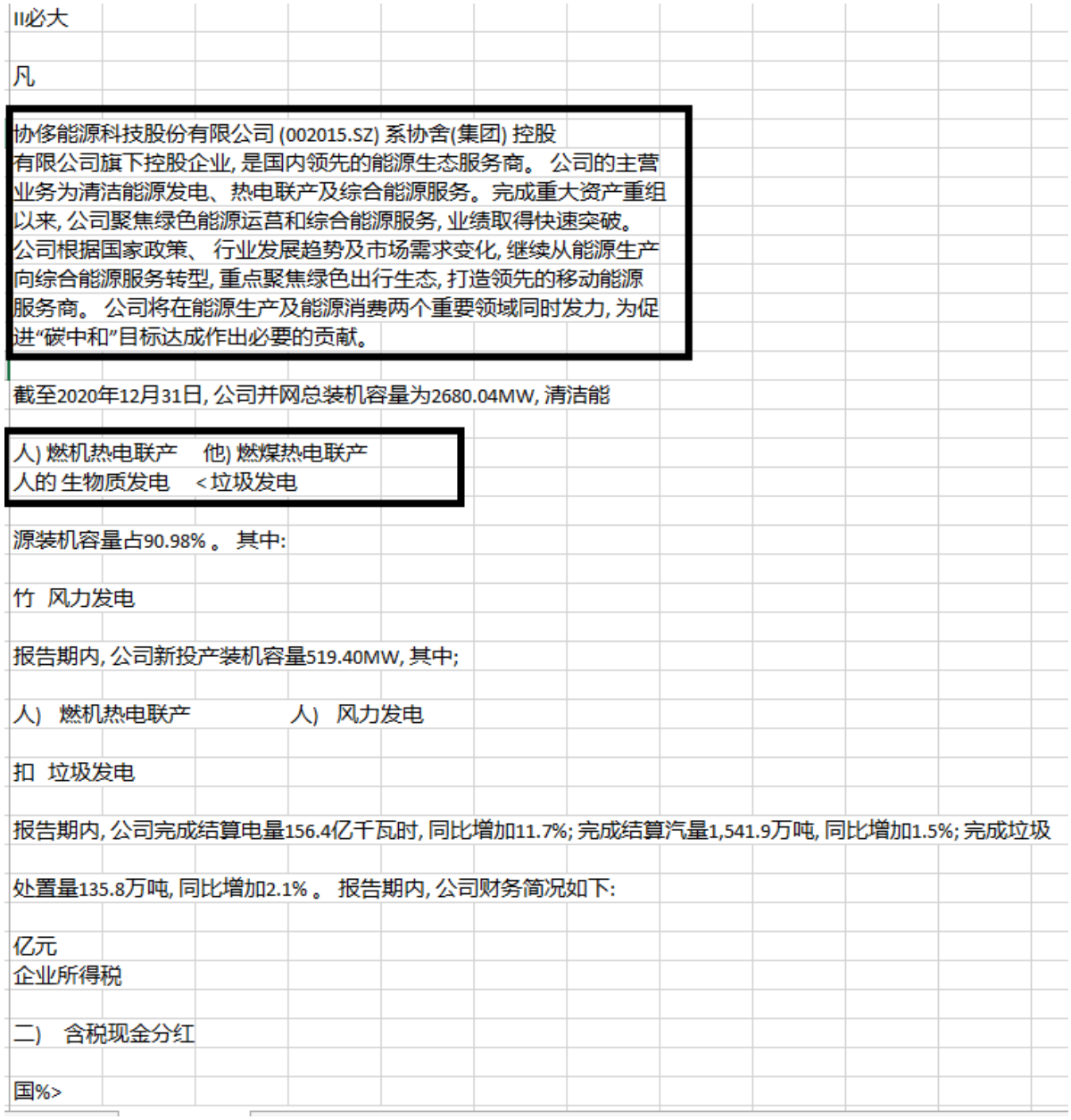

Figure A5. The previous screenshot is converted using Tesseract to sentences which preserve the coherence of the information in the ESG report page. While it is not 100% accurate, two main text chunks have been identified, which is the company's introduction, and a small portion which is erroneously extracted but does not confound the results because the data metrics provided is relevant to the ESG analysis. The other content areas are discarded.

Appendix 3

Model Specifications

Table A1.Specifications for BigBird-ESG	
tokenizer	Lowin/chinese-bigbird-wwm-base-4096
learning_rate	2e-5
train_batch_size	16
eval_batch_size	16
num_train_epochs	3
weight_decay	0.01
fp16	True
Specifications for BERT	
tokenizer	bert-base-chinese
learning_rate	2e-5
train_batch_size	16
eval_batch_size	16
num_train_epochs	3
weight_decay	0.01
fp16	True
Specifications for LSTM	
tokenizer	bert-base-chinese
hidden_dim	256
output_dim	len(df[‘label2’].unique())
num_layers	2
dropout	0.3
Specifications for SVM	
max features	5000
kernel	linear