DynamicESG: A Dataset for Dynamically Unearthing ESG Ratings from News Articles

Yu-Min Tseng
Data Science Degree Program, National Taiwan University
Taiwan
ymtseng@nlg.csie.ntu.edu.tw

Chung-Chi Chen
Artificial Intelligence Research Center, AIST
Japan
c.c.chen@acm.org

Hen-Hsen Huang
Institute of Information Science, Academia Sinica
Taiwan
hhhuang@iis.sinica.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University
Taiwan
hhchen@ntu.edu.tw

ABSTRACT

This paper introduces the DynamicESG dataset, a unique resource for dynamically extracting ESG ratings from news articles. The ESG rating, a novel metric employed annually to gauge a company's sustainability, relies heavily on corporate disclosure and other external information, especially news narratives. Our dataset, comprising a wide spectrum of news over a twelve-year span, annotates articles in accordance with MSCI ESG ratings methodology and SASB standards, with relevance to ESG issues. DynamicESG provides a comprehensive means of investigating the relationship between public discourse, ESG-related events, and subsequent ESG rating adjustments. We detail our data collection, curation, annotation procedure, and inter-rater agreement, ensuring high data quality and usability. Importantly, our dataset includes a temporal dimension, enabling the analysis of longitudinal trends in ESG ratings and their correlation with news coverage. Moreover, the dataset incorporates an opportunity/risk tendency, thus permitting analysis from diverse perspectives to discern if the news is beneficial or detrimental to the company. We believe this dataset will serve as a valuable resource for researchers in fields such as corporate social responsibility, sustainable investing, machine learning, and natural language processing. Initial analysis using the dataset underscores its potential to facilitate new insights into the dynamics of ESG ratings and the influence of news media on these ratings.

CCS CONCEPTS

• **Applied computing**;

KEYWORDS

ESG, ESG Rating, Social Good

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615118

ACM Reference Format:
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), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615118

1 INTRODUCTION

Corporate Social Responsibility (CSR) has increasingly become a crucial component of company operations. Nowadays, the impacts towards Environmental, Social, and Governance (ESG) serve as a third dimension, beyond return and risk, considered by investors when making corporate and personal investment decisions. Several frameworks and policies have been proposed to assess companies' ESG-related activities, with endeavors to quantify these considerations into scores. Yet, the scoring process necessitates substantial expert involvement and numerous manual annotation procedures for relevant events. To streamline this process and enhance experts' efficiency, this paper employs the guidelines utilized by experts in Morgan Stanley Capital International (MSCI), a globally recognized authority in formulating financial indexes, to annotate news articles and share these annotations with the research community. We anticipate that this dataset will stimulate the proposal of more sophisticated methods.

A shortcoming of the current ESG rating approach is the annual update frequency of ESG ratings. The financial market is dynamic and fast-paced; an annual update cycle is too delayed for decision-making based on the latest information. To address this, we select news articles as the resource for capturing the most recent events and deducing the impact of these events on a company's ESG rating. Our annotations span three aspects: impact type (risk/opportunity), impact duration, and ESG key issues. The impact type helps infer if a given news item will augment the ESG rating. Impact duration aids in understanding the duration of an event's influence on the ESG rating. Differentiating key issues across industries is crucial as the MSCI guidelines suggest varying weights for each issue. It is crucial to ascertain the issue addressed by the news; for example, the carbon emissions issue has a 15.6% weighting in the oil and gas drilling industry, contrasting with a mere 4.7% weighting in the specialized finance industry. Based on this annotation scheme,

we can capture the influence of sentiment, temporal, and topic dimensions for dynamically understanding the possible change of a company's ESG rating.

Another challenge with ESG rating is the recent fragmentation and lack of standardization across different frameworks, complicating the effective navigation of the ESG landscape. In addition to adhering to MSCI ESG guidelines, we incorporate another standard guideline, The Sustainability Accounting Standards Board (SASB) Standards,[1] to augment our proposed label scheme. This expanded label set enhances the applicability of our proposed dataset to most ESG rating guidelines and allows extension to other ESG-related analyses. We believe that incorporating various perspectives and best practices into a single scheme will yield a more robust and encompassing ESG-rating process. To sum up, this paper presents the DynamicESG dataset, a unique resource designed to dynamically glean ESG ratings from news articles. Our work underscores the importance of timely, multi-faceted analysis in understanding the implications of news narratives on ESG ratings and provides a foundation for future explorations in this domain.

2 RELATED WORK

The implications of ESG factors on corporate performance and investment decisions have been the focus of an expanding body of research. Foundational research conducted by MSCI [8] identified governance (G) as the predominant pillar for short-term analysis, representing event risks, while environmental (E) and social (S) indicators became increasingly crucial in the long term, reflecting cumulative risks to performance, such as carbon emissions. Mehra et al. [15] contributed to the field by introducing ESGBERT, a tool developed by fine-tuning the BERT model for sequence classification and performing a Masked Language Model (MLM) task on an ESG corpus. The experimental results demonstrated ESGBERT's success in learning the ESG context, highlighting its value for various ESG-specific text classification tasks. Further, Raman et al. [16] automatically generates an ESG relevance score for any corporate discourse, providing a mechanism to gauge and assess the emphasis placed on sustainable business practices. The other interesting study by Lee et al. [12] pointed out the challenges associated with calculating ESG scores for small businesses, which often lack extensive data records compared to larger corporations. Distinct from the aforementioned studies, our DynamicESG dataset extends the ESG discourse by focusing on daily news articles to dynamically assess changes in a company's ESG rating. This approach addresses the inherent limitations of the traditional annual update cycle, providing a more nuanced understanding of ESG impacts influenced by real-time events and developments.

3 DATASET

We constructed this dataset with a meticulous five-step pipeline to guarantee its utility and value. Initially, we gathered news data from the Business Today website, a well-regarded Taiwanese magazine known for its content on finance, business, and investment. Subsequently, we aligned the SASB Standard's ESG issues guideline with that of MSCI, culminating in forty-four ESG categories forming our definitive guideline. Annotators, guided by these categories, were tasked to ascertain three main aspects of the news articles: the impact type, impact duration, and ESG category. We ensured annotation consistency through biweekly discussion sessions, during which disagreements were resolved and answer sets adjusted as required. Lastly, we calculated the agreement and distribution of annotations across the entire dataset.

Table 1: Labels in ESG category identification task.

Topic	Theme	# of Key Issues
Environmental (E)	Climate Change	4
	Natural Capital	3
	Pollution & Waste	3
	Environmental Opportunity	3
Social (S)	Human Capital	5
	Product Liability	9
	Stakeholder Opposition	2
	Social Opportunity	4
Governance (G)	Corporate Governance	8
	Corporate Behavior	3

3.1 Data Collection

After surveying numerous ESG news websites, we selected the ESG page of Business Today[2] as our data source, owing to its reputation for comprehensive, diverse content spanning finance, business, and investment topics. We gathered a total of 2,472 Chinese news articles covering a broad spectrum of ESG topics from this source, ranging from January 1, 2011, to December 31, 2022.

3.2 Task Design

Various methodologies and guidelines for ESG ratings have been introduced globally to assess the implementation outcome of ESG practices by companies and organizations. These frameworks provide a standardized approach for evaluating and comparing the ESG performance of different entities. We selected the MSCI and SASB Standards as our primary guidelines among all the existing ESG frameworks.

Regarding ESG ratings, MSCI has outlined 35 key issues across three topics: Environmental, Social, and Governance. Additionally, MSCI's ESG rating methodology categorizes the expected time frame for risk or opportunity to materialize into two distinct periods: short-term ("less than 2 years") and long-term ("more than 5 years"). The methodology also differentiates between the importance of ESG key issues across industries, assigning the highest weight to short-term, high-importance factors and the lowest weight to long-term, low-importance factors. Based on this, MSCI assesses the ESG ratings of companies and organizations annually, using data from corporate or social disclosures, open government information, and other sources. SASB aims to improve the reporting of ESG issues. SASB has identified 26 significant issues spanning five aspects: Environment, Social Capital, Human Capital, Business Model & Innovation, Leadership & Governance. Furthermore, SASB has released the Materiality Map that details the relative importance of each issue across 77 industries.

[1] https://sasb.org/standards/materiality-map/

[2] https://esg.businesstoday.com.tw/catalog/180686/

Table 2: Agreement among annotators. Fleiss Kappa value exceeding 0.20 indicates fair agreement among annotators [11].

Task		Cohen	Fleiss	Krippendorff
Impact Type		0.49	0.49	0.54
Impact Duration		0.21	0.21	0.29
ESG Category	Topic (3)	0.52	0.49	0.55
	Theme (10)	0.35	0.34	0.37
	Key Issues (44)	0.28	0.28	0.28

Though SASB does not assess ESG ratings, Taiwan's Financial Supervisory Commission requires listed companies to report in alignment with SASB's ESG issues. Upon conducting a comparative analysis of two guidelines, we discover that MSCI places greater emphasis on environmental aspects, whereas SASB encompasses elements related to business models and innovation that MSCI overlooks. Therefore, for the purpose of generalization of ESG issues, we merged SASB's 26 important issues with MSCI's 35 key issues to formulate our final set of 44 ESG category guidelines for the ESG category task. Like MSCI, our guidelines are structured into three topics and ten themes but encompass 44 key ESG issues instead of 35. The statistics of our guidance are provided in Table 1.[3].

Guided by this methodology, we crafted three tasks to help classify news articles:

Impact Type Identification: This single-choice question aims to ascertain the type of impact a news article might have on the company. The possible labels are "Opportunity", "Risk", and "Cannot Distinguish".

Impact Duration Inference: This single-choice question seeks to determine the duration of the impact a news article might have on the company. Based on the distinction between short-term and long-term defined above, we present three labels: "Less than 2 years", "2 to 5 years", and "More than 5 years".

ESG Category Idenfication: This multiple-choice question is designed to identify the ESG categories related to a news article. The labels include the 44 ESG key issues, and an additional "None of the Above" choice.

3.3 Agreement and Statistics

We employ three measures, namely Cohen's Kappa [3], Fleiss Kappa [7], and Krippendorff's α [10], to evaluate the agreement among annotators. Table 2 presents the agreement for all tasks. As per Landis and Koch [11], a Fleiss Kappa value exceeding 0.20 indicates fair agreement among annotators. During our biweekly meetings, we made discussions on the instances that got different labels from annotators to clarify the rationale. After excluding articles that did not align with our objective, we obtained 2,220 instances from the 2,472 collected news articles. Table 3 shows the distribution of different labels in the proposed dataset.

In the released DynamicESG, we furnish the news headlines accompanied by a URL, and the annotations corresponding to the proposed three tasks. Due to copyright concerns, users are required to gather the news content themselves. To facilitate this, we provide the web crawler code that enables the reconstruction of the full dataset along with the annotations.[4] The annotations are released under the CC BY-SA 4.0 license.

4 EXPERIMENT

4.1 Longformer

Given that some news articles surpass the input length limitation (512 tokens) of conventional language models such as BERT [6] and RoBERTa [14], we employed Longformer [1] to overcome this issue. The Longformer was fine-tuned with the learning rate and weight decay set to 1e-5 and 0.03, respectively.

We present the experimental results in terms of precision, recall, and weighted F1-Score (Weighted F1) in Table 4. In the task of impact type identification, the models trained with the proposed DynamicESG dataset attain near-perfect performance in discerning opportunities and risks from an ESG perspective. Even though the proposed dataset is not extensive, the experimental results suggest its effective use in industrial applications, particularly for the impact type identification task, without significant concerns about the model's performance.

In the task of impact duration inference, the Longformer achieves a Weighted F1-score of 0.728. This underscores the importance of our biweekly meetings with annotators to resolve instances of disagreement. After addressing this issue, the model exhibits good performance when trained with the proposed DynamicESG dataset. These experimental results mitigate concerns about agreement during the annotation process, which merely meets the criteria for fair agreement. In this context, we conduct experiments under three-topic settings for the ESG issue identification task. Given that a news article may include discussions on two or even all topics, we experiment within a multi-label task setting. However, the experimental results indicate that the model still finds it challenging to identify ESG issues, even under the simplified three-topic setting.

4.2 Open Challenge for ESG Issue Identification

As discussed in Section 4.1, ESG issue identification is the most challenging among the proposed tasks. To foster advancement in this area, we organized a shared task as part of the 5th Workshop on Financial Technology and Natural Language Processing (FinNLP-2023) [2]. The shared task attracted 26 research teams, out of which three teams presented their solutions for the 44-issue ESG identification task [9, 13, 18]. This section outlines their methods and reports the state-of-the-art performance.

Wang et al. [18] employs MacBERT [5], a contrastive learning framework, and uses unlabeled and pseudo-labeled data to enhance the performance in this task. Linhares Pontes et al. [13] investigates the performance of SVM [4] with SentenceBERT's embeddings [17] (SBERT). Glenn et al. [9] illustrates how to leverage synthetic data from a large language model, ChatGPT, to enhance the performance of multilingual BERT (mBERT).

Table 5 displays the performance of the ESG category identification task under the 44-class task setting. Experimental results indicate there is ample room for further improvement in the

[3]Please refer to the guideline in our dataset for the details of all ESG issues.

[4]DynamicESG: https://github.com/ymntseng/DynamicESG

Table 3: Dataset Statistics.

	Impact Type			Impact Duration			ESG Category		
	Opportunity	Risk	Cannot Distinguish	2 years <	2 to 5 years	> 5 years	E	S	G
Train	70.25%	7.60%	3.01%	19.95%	14.22%	46.56%	25.92%	33.45%	27.12%
Dev	7.86%	0.79%	0.39%	2.29%	1.61%	5.28%	2.12%	2.36%	1.84%
Test	8.78%	0.92%	0.39%	2.52%	1.83%	5.73%	2.72%	2.24%	2.20%

Table 4: Longformer's performances.

	Impact Type	Impact Duration	ESG Category
Precision	0.877	0.722	0.671
Recall	0.872	0.772	0.450
Weighted F1	0.865	0.728	0.467

Table 5: Results of 44-class ESG issue identification task.

Method	Micro F1	Macro F1	Weighted F1
mBERT with LLM Synthetic Data [9]	0.121	0.038	0.091
SVM with SBERT Embeddings [13]	0.279	0.137	0.263
MacBERT with Data Augmentation [18]	0.391	0.180	0.392

fine-grained ESG issue identification task. The shared task participants' experiences also demonstrate that the proposed DynamicESG dataset is already usable, and users can reconstruct the dataset effortlessly with our provided code.

5 DISCUSSION

5.1 ESG Rating

As mentioned in Section 1, a key issue with current ESG ratings is their annual update frequency, which is inadequate given the daily fluctuations of the financial market. In response to this, we focus on news articles as data resources in our DynamicESG dataset, following the MSCI ESG rating scheme for annotations. This section delves into using the DynamicESG dataset to assess potential changes in ESG ratings by analyzing daily news.

Firstly, it is intuitive that news identified as an ESG opportunity (risk) will have a positive (negative) influence on the ESG rating. Based on the experimental results in Section 4.1, it can be argued that models can perform very well in this task. This implies that sentiment analysis from the ESG perspective is almost resolved with the proposed DynamicESG dataset. Secondly, as discussed in Section 3.2 and based on the MSCI ESG rating guidelines, events with a short-term impact duration will have a larger influence on the ESG rating than those with long-term impact. Therefore, to estimate the magnitude of an event's influence on a company's ESG rating, the impact duration needs to be inferred. This novel task introduced in the proposed DynamicESG shows that it remains a challenge for both financial background annotators and machine learning models.

However, to derive an ESG score, more fine-grained weighting information is needed. As such, we provide fine-grained annotations for ESG key issues in the third step, demonstrating through experiments in Section 4 that this is the most challenging task in the proposed dataset. For example, only three out of thirteen key environmental issues are considered when scoring the ESG performance of companies in the oil and gas drilling industry. This underscores the importance of identifying the key issues. With the key issue of a news article, we can derive the weighting (a precise number) of its influence on the ESG rating.

In conclusion, the proposed DynamicESG presents the first dataset for adhering to MSCI ESG guidelines to analyze new events for a company. This dataset lays the groundwork for future research to update the fine-grained influence of daily news events on a company's ESG scores.

5.2 Limitations

There are certain limitations that merit acknowledgment. First, our dataset presently incorporates two widely adopted ESG standards, MSCI and SASB. However, given the diversity of existing ESG frameworks, additional guidelines could be included in the dataset to enhance its comprehensiveness. Future endeavours could thus focus on expanding the labeling scheme to encompass more guidelines, thereby increasing the dataset's versatility. Second, our current dataset does not distinguish between different scales at which ESG events occur. For instance, a company-specific event may exert a different impact on the ESG rating compared to a global or industry-wide event. Future iterations of the dataset could factor in these event scales, providing a more nuanced understanding of ESG impacts. Third, it is essential to acknowledge that the news articles within our dataset may occasionally present an incomplete or biased perspective of events. This can arise from factors such as the political inclinations of the publication, selective reporting, or sensationalism. In future endeavors involving our dataset, it is imperative to account for these inherent biases within news articles, thereby ensuring a more balanced and accurate analysis.

6 CONCLUSION

The DynamicESG dataset offers a unique approach to ESG rating by providing daily analyses from news articles, adhering to MSCI and SASB standards. While the dataset highlights the challenges of weighting ESG issues and assessing impact durations, it emphasizes the need for dynamic responses in the rapidly changing financial market. We anticipate that DynamicESG will pave the way for advanced ESG rating research and innovations.

ACKNOWLEDGMENTS

This research is supported by National Science and Technology Council, Taiwan, under grants 110-2221-E-002-128-MY3, 110-2634-F-002-050-, and 111-2634-F-002-023-. The work of Chung-Chi Chen was supported in part by JSPS KAKENHI Grant Number 23K16956 and a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

[2] Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-Lingual ESG Issue Identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin).*

[3] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[4] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[5] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020.* Association for Computational Linguistics, Online, 657–668. https://doi.org/10.18653/v1/2020.findings-emnlp.58

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).* Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[8] Guido Giese, Zoltán Nagy, and Linda-Eling Lee. 2021. Deconstructing ESG ratings performance: Risk and return for E, S, and G by time horizon, sector, and weighting. *The Journal of Portfolio Management* 47, 3 (2021), 94–111.

[9] Parker Glenn, Alolika Gon, Nikhil Kohli, Sihan Zha, Parag Pravin Dakle, and Preethi Raghavan. 2023. Jetsons at the FinNLP-2023: Using Synthetic Data and Transfer Learning for Multilingual ESG Issue Classification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin).*

[10] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[11] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[12] Ook Lee, Hanseon Joo, Hayoung Choi, and Minjong Cheon. 2022. Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms. *Sustainability* 14, 14 (2022), 8745.

[13] Elvys Linhares Pontes, Mohamed Benjannet, and Lam Kim Ming. 2023. Leveraging BERT Language Models for Multi-Lingual ESG Issue Identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin).*

[14] 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* (2019).

[15] Srishti Mehra, Robert Louka, and Yixun Zhang. 2022. ESGBERT: Language Model to Help with Classification Tasks Related to Companies Environmental, Social, and Governance Practices. *arXiv preprint arXiv:2203.16788* (2022).

[16] Natraj Raman, Grace Bang, and Armineh Nourbakhsh. 2020. Mapping ESG trends by distant supervision of neural language models. *Machine Learning and Knowledge Extraction* 2, 4 (2020), 453–468.

[17] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).* Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[18] Weiwei Wang, Wenyang Wei, Qingyuan Song, and Yansong Wang. 2023. Leveraging Contrastive Learning with BERT for ESG Issue Identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing (FinNLP) and the Second Multimodal AI For Financial Forecasting (Muffin).*