Finance Research Letters 61 (2024) 104979

Contents lists available at ScienceDirect

Finance Research Letters

journal homepage: www.elsevier.com/locate/frl

Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication☆

Tobias Schimanski [a], Andrin Reding [b], Nico Reding [b], Julia Bingler [c], Mathias Kraus [d], Markus Leippold [e,*]

[a] University of Zurich, Switzerland
[b] University of St. Gallen, Switzerland
[c] University of Oxford and Council of Economic Policies (CEP), United Kingdom
[d] FAU Erlangen-Nürnberg, Germany
[e] University of Zurich and Swiss Finance Institute (SFI), Switzerland

ARTICLE INFO

JEL classification:
G2
G38
C8
M48

Keywords:
ESG analysis in financial markets
Natural language processing
BERT model

ABSTRACT

Environmental, social, and governance (ESG) criteria take a central role in fostering sustainable development in economies. This paper introduces a class of novel Natural Language Processing (NLP) models to assess corporate disclosures in the ESG subdomains. Using over 13.8 million texts from reports and news, specific E, S, and G models were pretrained. Additionally, three 2k datasets were developed to classify ESG-related texts. The models effectively explain variations in ESG ratings, showcasing a robust method for enhancing transparency and accuracy in evaluating corporate sustainability. This approach addresses the gap in precise, transparent ESG measurement, advancing sustainable development in economies.

1. Introduction

Since the inception of environmental, social, and governance (ESG) criteria in 2004 (UN, 2004), its prominence within the corporate landscape has continuously increased, and the integration of ESG considerations has fundamentally transformed business operations. This shift is evident through various indicators. The value of assets under management following the United Nations Principles for Responsible Investment swelled from US$21 trillion in 2010 to an impressive US$123 trillion in 2023.[1] The burgeoning number of ESG rating agencies, albeit with differing assessments, underscores the escalating significance of ESG quantification (Berg et al., 2022). Additionally, the evolution of reporting standards and mandatory disclosure requirements, exemplified by the European Union's directives, reflect this upward trajectory of ESG's importance (EU, 2023). However, this increasing disclosure of ESG information opens up avenues for deceptive greenwashing practices, which further necessitates transparent and comprehensive evaluations of corporate ESG representations.

Despite this growing importance, current ESG measurement methods, particularly ESG ratings, frequently yield inconsistent results, raising questions about the accurate integration of ESG efforts by companies (Berg et al., 2022). A promising solution to this inconsistency is scrutinizing corporate disclosures through Natural Language Processing (NLP) techniques. While there has been

☆ This paper has received funding from the Swiss National Science Foundation (SNSF) under the project 'How sustainable is sustainable finance? Impact evaluation and automated greenwashing detection' (Grant Agreement no. 100018_207800).

* Corresponding author.
E-mail address: markus.leippold@bf.uzh.ch (M. Leippold).

[1] See www.unpri.org/pri.

https://doi.org/10.1016/j.frl.2024.104979
Received 20 November 2023; Received in revised form 13 December 2023; Accepted 2 January 2024
Available online 9 January 2024

development in NLP methodologies within the climate change sphere (see, e.g. Webersinke et al., 2022; Sautner et al., 2023), the broader environmental sector, as well as the social and governance realms, have been partially or entirely overlooked.

Our paper seeks to bridge the identified gap by making several substantial contributions to quantifying ESG narratives using NLP techniques. Firstly, we augment the current body of literature by pre-training new language models that specifically cater to the environmental, social, and governance domains. To this end, we utilize a large corpus comprising over 13.8 million textual samples. Secondly, we create and introduce three expert-annotated datasets. Each of these datasets encompasses 2000 text samples across the three ESG pillars. By doing so, we facilitate training precise classification models for each subdomain. Furthermore, this enhancement facilitates the refinement of text classification when a given text does not fall exclusively within a single ESG category but intersects across two or three categories simultaneously. Current ESG classifiers struggle with multiclassification, leading to noisy inferences.[2] Thirdly, and of notable significance, we conduct a performance evaluation of these models by scrutinizing over 2500 annual reports from major European corporations dated between 2016 and 2021. By exploring the correlation between corporate communication patterns and their ESG ratings, we provide the rigorous accuracy assessment often absent in NLP applications published in the finance literature. Our commitment to transparency is underlined by the publication of both the datasets and the models for public use, enabling a broad spectrum of stakeholders to perform in-depth analyses of corporate disclosures.[3]

Our pre-trained ESG models achieve state-of-the-art performance in text classification within their respective domains and show substantial effectiveness in practical, real-world applications. After adjusting for a variety of control variables, our analysis reveals a robust and statistically significant positive relationship between the quality of ESG disclosures and the corresponding ESG ratings awarded to companies. The implications of our findings extend across the academic and commercial sectors. In an era where the volume of textual information is growing, the availability of precise and actionable ESG metrics becomes paramount. The models developed in this study are instrumental for investment professionals, ESG analysts, and corporate strategists, providing the means to sift through vast data sources effectively and derive valuable insights. In addition, companies may be incentivized to align their operations with ESG principles more closely by ensuring that corporate disclosures undergo a stringent evaluation. The academic community will also stand to gain, as the datasets and models established here lay a groundwork for subsequent ESG-centric scholarly pursuits.

Prior research has harnessed NLP to understand ESG integration within communications better. Earlier finance-centric NLP applications predominantly employed keyword-based approaches, which lacked contextual sensitivity (Cody et al., 2015; Sautner et al., 2023). Recent advancements have embraced machine learning models like BERT, which offer context-aware capabilities to overcome this shortcoming. Numerous BERT-based datasets have been introduced, addressing a spectrum of tasks such as climate content classification, topic detection, question-answering systems, and claim detection and verification (Webersinke et al., 2022; Kölbel et al., 2022; Bingler et al., 2022, 2023; Callaghan et al., 2021; Varini et al., 2021; Luccioni et al., 2020; Stammbach et al., 2022; Wang et al., 2021). Yet, research on NLP applications for the broader environmental, social, and governance dimensions of ESG remains scarce.

The lack of models tailored to the comprehensive ESG framework is due to domain-specific pretraining. This process, in which a model learns language nuances through semi-supervised methods, usually involves a large corpus of specialized texts. While these broad corpora endow a model with solid language understanding, they can stumble upon specialized, niche language terms (Araci, 2019). This challenge has prompted the generation of specialized text corpora for subdomains, resulting in models that excel in tasks like classification or claim verification within their respective areas (Rasmy et al., 2021; Chalkidis et al., 2020; Araci, 2019). Specifically, ClimateBERT has been further refined in the climate domain with a targeted corpus (Webersinke et al., 2022).

Hence, while ESG's relevance continues to burgeon, there remains a pronounced void in researching and developing NLP methods to assess ESG communication across all its pillars comprehensively. This study addresses this need by creating holistic models and datasets for the nuanced evaluation of the environmental, social, and governance aspects of ESG.

2. Pretraining environmental, social, and governance models

We create pretrained models for the environmental, social, and governance subdomains. We need specialized datasets to enhance the language model's understanding of the subdomains to perform this pretraining. We create these specialized datasets for our tasks following a two-step procedure. First, we compile a base dataset of relevant underlying sources. As relevant sources, we define corporate news, annual reports, and sustainability reports. This decision aims to strengthen the models' ability to specialize in corporate jargon. We split each source into its sentences (see Appendix A for an overview). Second, we apply a keyword search to find relevant text passages for all three subdomains of ESG (see Appendix B). Thus, we create a specific dataset to train each model for a particular task. The sentence characteristics of the base sources and the datasets can be viewed in Table 1.

After creating the datasets for each subdomain of ESG, we pre-train the RoBERTa (Liu et al., 2019) and DistilRoBERTa (Sanh et al., 2020) models. The main difference is in their model size. While RoBERTa possesses 125 million parameters, its smaller counterpart, DistilRoBERTa, consists of 85 million parameters. For each environmental, social, and governance domain, we further pretrain one RoBERTa and one DistilRoBERTa model. As a result, we obtained six models (see Appendix C for the pretraining details).

[2] As an example, the sentence 'Our company scores highly on all ESG criteria.' is classified by FinBERT-ESG (Huang et al., 2023) as 90% 'Social', 6% 'Governance,' 1% 'Environmental', and 3% as 'None' (see https://huggingface.co/yiyanghkust/finbert-esg).

[3] The models are accessible on the Huggingface platform at https://huggingface.co/ESGBERT. Furthermore, extended tutorials on the model usage can be found on Medium under this link.

Table 1
Pretraining datasets for each domain.

Domain	Num. of sentences	Avg. num. of words		
		Q1	Mean	Q3
Base Data	13,846,000	16	25.12	30
Environment	2,100,586	17	27.03	32
Social	1,787,198	18	28.67	34
Governance	3,011,546	19	30.52	36

Table 2
Classification results of five-fold cross-validation. The table reports the accuracy for the environment, social, and governance models, with standard deviations (std.) in brackets. For FinBERT-ESG (Huang et al., 2023), we tested with all E, S, and G sentences in our dataset.

Domain	Model	Accuracy (std.)
Env	DistilRoBERTa	0.940 (0.013)
	RoBERTa	0.950 (0.016)
	EnvDistilRoBERTa	0.950 (0.011)
	EnvRoBERTa	**0.957 (0.010)**
	FinBERT-ESG	0.8765
Soc	DistilRoBERTa	0.927 (0.007)
	RoBERTa	0.919 (0.008)
	SocDistilRoBERTa	0.932 (0.015)
	SocRoBERTa	**0.934 (0.014)**
	FinBERT-ESG	0.7585
Gov	DistilRoBERTa	0.885 (0.011)
	RoBERTa	0.887 (0.011)
	GovDistilRoBERTa	**0.897 (0.011)**
	GovRoBERTa	0.896 (0.011)
	FinBERT-ESG	0.8040

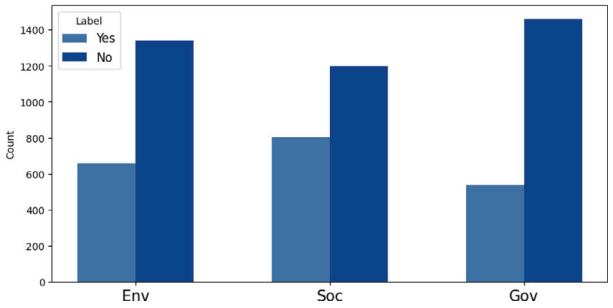

Fig. 1. Label distribution in the 2k expert-annotated datasets.

3. Creating environmental, social, and governance classifier models

Apart from pretraining models in the ESG domain, our goal is to create classifier models that detect ESG communication in textual sources. In this context, we proceed in two steps. In the first step, we create three datasets with 2000 text samples for the environmental, social, and governance domains (details on data creation can be found in the Appendix D)). To enable a common understanding among annotators, we create ESG labeling guidelines (see Appendix E). Then, three expert annotators from the author team label each text sample. We achieve a very high inter-annotator agreement of more than 86% on each task (for more details, see Appendix F). Fig. 1 outlines the resulting label distribution.

We used the created datasets in the second step to fine-tune and evaluate several classification models. We perform a five-fold cross-validation to evaluate the model performance with the created datasets. This allows us to test the model performance on the entirety of the dataset. As Table 2 shows, the further pre-trained models consistently outperform their base models. Even the smaller, further pretrained DistilRoBERTa models achieve on-par or superior performance compared to the larger base RoBERTa models. Overall, the results indicate a strong model performance with over 93% accuracy for the social and environmental domain and over 89% for the governance domain. These results remain consistent when using different sets of hyperparameters, solidifying the superiority of our further pre-trained models (see Appendix I).

4. Model comparison with FinBERT-ESG

The FinBERT-ESG model represents the only published ESG classifier (Huang et al., 2023) and serves as an additional benchmark for our experiments. The model takes a text and assigns the label "environment", "social", "governance", or "none". This differs from our approach, where we develop models for each pillar of ESG.

Unfortunately, Huang et al. (2023) neither publish labeling guidelines nor the datasets or annotator information. Therefore, we can only evaluate their model on our datasets. Thus, we classify every text sample of our three datasets with their model. Since Huang et al. (2023) cite MSCI when introducing their label categories, we assume that they share a similar understanding of the concept of ESG displayed in our labeling guidelines.

As Table 2 shows, we outperform the models on a large margin. This lower performance may arise for three reasons. First, the authors use one model to classify three categories of ESG and non-ESG. Thus, they cannot handle sentences that could be assigned with multiple labels. For instance, the sentence "In our company, we are convinced that all three aspects of ESG are of equal importance". is assigned the "social" label with the following label probabilities: Social: 0.474, None: 0.434, Governance: 0.076, Environmental: 0.015. Our single models would assign the E, S, and G labels to this example. Second, the different pretraining processes likely also drive the lower performance. Since our models specialize in E, S, and G communication, they are better suited for classification. Third, the authors use a 2k dataset to fine-tune one single model, while we use a 2k dataset for each model. Combined with using three distinct models, this likely drives the model's capabilities because of only focusing on one differentiation, e.g., E or not E. In turn, FinBERT-ESG needs to differentiate between four labels, E, S, G, and none.

5. ESG communication and ESG ratings

To find evidence for the validity of the models, we analyze whether the models' assessed communication patterns in companies' annual reports can explain variations in their respective ESG ratings. We hypothesize that companies are highly incentivized to disclose their ESG activities for various interconnected reasons. There is increasing societal and corresponding regulatory pressure to disclose ESG activities. On the one hand, there are protests for more climate action around the world[4] and increased consumer awareness of overall ESG implementation.[5] On the other hand, legislators recognize the need for action in the ESG domain by introducing new regulatory frameworks – particularly in Europe.[6] Therefore, we argue that ESG communication is associated with higher ESG ratings. At the same time, we acknowledge that sole general ESG communication can be prone to greenwashing. Therefore, this relationship might instead represent a general average than the truth for every individual company.

This study analyzes annual reports from the EuroStoxx600 index from 2017–2021. Therefore, we sampled over 2500 annual reports from 600 enterprises from 22 countries. To mitigate concerns about the divergence of individual ratings, we consider the ESG ratings of three major data providers: Bloomberg, Refinitiv Asset4, and RobecoSAM. Furthermore, we complement the ESG data with fundamental data from Compustat. To quantify the ESG communication of a company, we employ the best checkpoints of our fine-tuned DistilRoBERTa E, S, and G models on every sentence of the firm's annual report.[7] If a sentence qualifies for either subcategory of ESG, it is automatically an ESG sentence. Furthermore, a sentence with more than one label between E, S, and G is assigned with the $multilabel$ label.

To build a company score, we divide the number of ESG sentences by the number of all sentences in the annual report (ESG_com):

$$ESG\_com_{i,t} = \frac{\#ESGsentences_{i,t}}{\#allsentences_{i,t}},$$

where i indexes the company and t denotes the respective year. In addition to the E, S, G, and ESG communication score, we also calculate a $multilabel$ score. This signals sentences that are assigned to more than one label of E, S, and G. Then, to further investigate the relationship between ESG communication and ESG ratings, we propose the following model:

$$ESG\_rat_{i,t} = \alpha + \beta_{ESG\_com} * ESG\_com_{i,t} + \beta^{\top} X_{i,t} + \delta_i + \nu_t + \epsilon_{i,t},$$

where ESG_rat denotes a company's ESG (or E, S, G) rating. We both investigate the relationship between individual ratings and build a combined rating. We denote the company fundamentals by $X_{i,t}$ (see Table G.6 in Appendix G), and by δ_i and ν_t we denote industry and year-fixed effects, respectively, and $\epsilon_{i,t}$ is the error term.

The regression results support the assumption that ESG communication possesses explanatory power for combined ESG ratings. As Table 3 shows, all ESG communication coefficients indicate a strong and significant relationship. Simplified, a 1% increase in ESG (E) communication is associated with an increase in the ESG (E) rating by 0.6% (0.88%). These findings are largely consistent when regressions are performed with the single Refinitiv Asset4, Bloomberg, and RobecoSAM ratings as dependent variables (see Appendix H).

[4] See for example Fridays for Future.

[5] See for instance this PWC market study.

[6] Examples include the EU green deal.

[7] Since the results only marginally differ between DistilRoBERTa and RoBERTa, we decided to use the smaller, more energy-efficient and faster models. This means that we chose the pre-trained and fine-tuned DistilRoBERTa models.

Table 3
Regression results. Significance levels are denoted as follows: $p \leq 0.10$ by *, $p \leq 0.05$ by **, and $0.01 \leq p$ by ***.

	Dependent variable: Combined rating			
	ESG_rat	env_rat	soc_rat	gov_rat
ESG_com	0.6075***			
env_com		0.8758***		
soc_com			0.7710***	
gov_com				0.8967**
roa	−0.0003	−0.0004	−0.0009	0.0003
log(market cap)	−0.0090	−0.0043	−0.0033	−0.0210
mtbr	−0.0001	−0.0002	−0.0000	−0.0002
log(revenue)	0.0490***	0.0462***	0.0441***	0.0322**
current ratio	−0.0101	−0.0164**	−0.0048	−0.0042
Fixed Effects				
industry	Yes	Yes	Yes	Yes
year	Yes	Yes	Yes	Yes
Observations	1687	1687	1687	1687
R2	0.20341	0.22051	0.17035	0.06511

6. Conclusion

This paper demonstrates the development of robust pre-trained and fine-tuned models to detect ESG communication in textual disclosures. The publication of datasets and models will help a variety of stakeholders to rigorously and transparently assess companies' ESG communication. Although we include a large number of textual samples, the models are likely limited to written disclosures and may fail to show the same effectivity on transcripts of verbal communication. Furthermore, the broad nature of the governance label might entail generalizability problems. Finally, although the results of the regression analysis suggest that the models are indeed working in the intended manner, further investigations are needed to uncover patterns beyond correlations, providing exciting avenues for future research.

CRediT authorship contribution statement

Tobias Schimanski: Writing – review & editing, Writing – original draft, Software, Formal analysis, Data curation, Conceptualization. **Andrin Reding:** Writing – original draft, Software, Data curation, Conceptualization. **Nico Reding:** Writing – original draft, Software, Data curation, Conceptualization. **Julia Bingler:** Writing – review & editing, Writing – original draft, Conceptualization. **Mathias Kraus:** Writing – review & editing, Writing – original draft. **Markus Leippold:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization.

Data availability

All the models and data are freely available (links in the paper).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.frl.2024.104979.

References

Araci, D., 2019. Finbert: Financial sentiment analysis with pre-trained language models.
Berg, F., Kölbel, J.F., Rigobon, R., 2022. Aggregate confusion: The divergence of ESG ratings*. Rev. Finance 26 (6), 1315–1344.
Bingler, J.A., Kraus, M., Leippold, M., Webersinke, N., 2022. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. Finance Res. Lett. 102776.
Bingler, J.A., Kraus, M., Leippold, M., Webersinke, N., 2023. How Cheap Talk in Climate Disclosures Relates to Climate Initiatives, Corporate Emissions, and Reputation Risk. Swiss Finance Institute Research Paper (22-01).
Callaghan, M., Schleussner, C.-F., Nath, S., Lejeune, Q., Knutson, T.R., Reichstein, M., Hansen, G., Theokritoff, E., Andrijevic, M., Brecha, R.J., et al., 2021. Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. Nat. Clim. Change 11 (11), 966–972.
Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I., 2020. Legal-bert: The muppets straight out of law school.
Cody, E.M., Reagan, A.J., Mitchell, L., Dodds, P.S., Danforth, C.M., 2015. Climate change sentiment on twitter: An unsolicited public opinion poll. PLoS One 10 (8), 1–18. http://dx.doi.org/10.1371/journal.pone.0136092.
EU, 2023. Corporate sustainability reporting.
Huang, A.H., Wang, H., Yang, Y., 2023. Finbert: A large language model for extracting information from financial text*. Contemp. Account. Res. 40 (2), 806–841, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1911-3846.12832.
Kölbel, J.F., Leippold, M., Rillaerts, J., Wang, Q., 2022. Ask BERT: How regulatory disclosure of transition and physical climate risks affects the CDS term structure*. J. Financ. Econom. nbac027. http://dx.doi.org/10.1093/jjfinec/nbac027.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Luccioni, A., Baylor, E., Duchene, N., 2020. Analyzing sustainability reports using natural language processing.

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., Zhi, D., 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. npj Digit. Med. 4 (86).

Sanh, V., Debut, L., Chaumond, J., Wolf, T., 2020. Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Sautner, Z., Lent, L. v., Vilkov, G., Zhang, R., 2023. Firm-level climate change exposure. J. Finance 78 (3), 1449–1498, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13219.

Stammbach, D., Webersinke, N., Bingler, J.A., Kraus, M., Leippold, M., 2022. A dataset for detecting real-world environmental claims. arXiv preprint arXiv:2209.00507.

UN, 2004. Who Cares Wins: connecting Financial Markets to a Changing World. Technical Report, United Nations Global Compact.

Varini, F.S., Boyd-Graber, J., Ciaramita, M., Leippold, M., 2021. Climatext: A dataset for climate change topic detection.

Wang, G., Chillrud, L.G., McKeown, K., 2021. Evidence based automatic fact-checking for climate change misinformation. In: ICWSM Workshops. URL https://api.semanticscholar.org/CorpusID:237424411.

Webersinke, N., Kraus, M., Bingler, J., Leippold, M., 2022. Climatebert: A pretrained language model for climate-related text. In: Proceedings of AAAI 2022 Fall Symposium: The Role of AI in Responding to Climate Challenges.