Transformers-Based Approach for a Sustainability Term-Based Sentiment Analysis (STBSA)

Blaise W. Sandwidi
CEG Department
International Finance Corporation (IFC)
2121 Pennsylvania Avenue N.W.,
Washington, DC 20433 U.S.A
bsandwidi@ifc.org

Suneer P. Mukkolakal
ITS Department
The World Bank
1818 H Street, N.W.,
Washington, DC 20433 U.S.A
spallitharammalm@worldbankgroup.org

Abstract

Traditional sentiment analysis is a sentence-level or document-level task. However, a sentence or paragraph may contain multiple target terms with different sentiments, making sentiment prediction more challenging. Although pre-trained language models like BERT have been successful, incorporating dynamic semantic changes into aspect-based sentiment models remains difficult, especially for domain-specific sentiment analysis. To this end, in this paper, we propose a <u>T</u>erm-<u>B</u>ased <u>S</u>entiment <u>A</u>nalysis (TBSA), a novel method designed to learn **Environmental, Social, and Governance (ESG)** contexts based on a sustainability taxonomy for ESG aspect-oriented sentiment analysis. Notably, we introduce a technique enhancing the ESG term's attention, inspired by the success of attention-based neural networks in machine translation (Bahdanau et al., 2015) and Computer Vision (Bello et al., 2019). It enables the proposed model to focus on a small region of the sentences at each step and to reweigh the crucial terms for a better understanding of the ESG aspect-aware sentiment. Beyond the novelty in the model design, we propose a new dataset of 125,000+ ESG analyst-annotated data points for sustainability term-based sentiment classification, which derives from historical sustainability corpus data and expertise acquired by development finance institutions. Our extensive experiments combining the new method and the new dataset demonstrate the effectiveness of the Sustainability TBSA model with an accuracy of 91.30% (90% F1-score). Both internal and external business applications of our model show an evident potential for a significant positive impact toward furthering sustainable development goals (SDGs).

1 Introduction

In 2015, the United Nations (UN) adopted the 2030 Agenda and its 17 Sustainable Development Goals (SDGs; Nations (2015)), addressing global challenges including poverty, inequality, climate change, environmental degradation, peace, and justice. The Secretary General's Roadmap for financing this collective and transnational effort invites all stakeholders to consider environmental, social, and governance (ESG) issues. ESG matters have assumed relevance for investors, regulators, and industry participants, while ESG criteria are increasingly used to measure the impact of investment activities on sustainable development. However, ESG-integrated investing remains challenging, even for world-class asset managers, institutional investors, and pension funds, because of data gaps in coverage of emerging markets and a lack of analytical capacity. Further, these markets present the greatest opportunities for investors to achieve impacts through the SDGs because their development needs are the most significant.

At the same time, there is growing recognition of the fundamental role played by data, primarily structured data, in achieving the objectives set out in the SDGs (Griggs et al., 2013; Nilsson et al., 2016; Conforti et al., 2020; Vinuesa et al., 2020). Structured data and SDG metrics are essential to ensure the successful design of local projects but are often absent when required for insights into beneficiaries' needs and values (Conforti et al., 2020). Unstructured data can provide such insights. Natural language processing (NLP) techniques can process such qualitative data to provide relevant facts and figures to project developers. Expected benefits are time and cost reductions, higher operations efficiencies, due diligence improvements, and better sustainability impact assessments (Conforti et al., 2020; Sokolov et al., 2021; Ulibarri et al., 2019). Recent progress in masked language modeling such as Google BERT (bidirectional encoder representations from transformers, (Devlin et al., 2019), RoBERTa (robustly optimized BERT approach (Liu et al., 2019)), and DeBERTa (decoding-enhanced BERT, (He et al., 2021))—combined

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 157 - 170
December 7, 2022 ©2022 Association for Computational Linguistics

with cloud computing, is unlocking the potential for creating analytical capacity to assess unstructured data at scale and is facilitating SDG-aligned financing for emerging markets to address the $4.2 trillion USD annual shortfall in investments needed to meet the SDGs (OECD, 2020).

Despite these advances, NLP research and applications that contribute to sustainable development are absent (Conforti et al., 2020). This gap is attributed to the lack of high-quality sustainability data and the scarcity of relevant labeled data to train sustainability-domain language models. Our work proposes a sustainability-domain adaptation of transformer-based models to perform various NLP tasks, such as ESG term extraction and sentiment analysis. Such a sustainability domain-specific language model is a significant advance; pre-trained models and commercial sentiment analysis solutions perform poorly at predicting ESG sentiments because of differences in domain-specific vocabulary (these models are trained using datasets such as restaurant or movie reviews or tweets that are not relevant to sustainability analysis). Domain-specific models are also necessary to process sustainability reporting documents which are typically lengthy, complex, and use terms that do not carry emotional connotations, unlike movie or restaurant reviews. Hence the need to create a specific taxonomy for context-based ESG sentiment analysis (Ulibarri et al., 2019).

Development finance institutions have decades of archival sustainability data created from project due diligence and monitoring. We use examples of such data to create a unique ESG taxonomy and human-annotated dataset. Namely, we equip two pre-trained language models (RoBERTa and DeBERTa) to understand ESG context by fine-tuning and modifying the models into a sustainability term-based sentiment analysis (STBSA) model, thereby creating a new approach based on an ESG taxonomy of more than 1,200 terms. We then train the models with human-annotated data to predict the context of ESG terms in sentences and classify words by positive, negative, or neutral ESG sentiment. Significantly, our experiments find that the STBSA model (based on RoBERTa) performs with 91.30% accuracy (90% F1-score) and outperforms the current state-of-the-art baseline models for sentiment analysis tasks.

2 Related Work

Aspect-based Sentiment Analysis. In the beginning, work on sentiment analysis mainly focused on identifying the overall sentiment of a unit of text. The amount of text varied from an entire document (Pang et al., 2002; Turney, 2002) to merely paragraphs or sentences (Hu and Liu, 2004). However, only considering the overall sentiment fails to capture the sentiments over the aspects on which an entity can be reviewed or sentiment expressed toward different entities. To remedy this, two new tasks have been introduced: aspect-based sentiment analysis (ABSA) and targeted sentiment analysis. Aspect-based sentiment analysis assumes a single entity per unit of analysis and tries to identify sentiments towards different aspects of the entity (Lu et al., 2011; Lakkaraju et al., 2014; Alghunaim, 2015; Bagheri et al., 2013; Brody and Elhadad, 2010). However, it considers only one single entity in the given text.

Target-based or target sentiment analysis is another task that identifies polarity towards a target entity, as opposed to over an entire volume of text (Saeidi et al., 2016; Mitchell et al., 2013; Jiang et al., 2011; Dong et al., 2014; Vo and Zhang, 2015). Jiang et al. (2011) were the first to propose targeted sentiment analysis on Twitter. They demonstrated the importance of targets by showing that 40% of sentiment errors are due to not considering them in classification. However, this task only identifies the overall sentiment, and the existing corpora consist only of text with one single entity per unit of analysis. This task caters to more generic text by making fewer assumptions while extracting fine-grained information.

ESG-domain transformers-based models. In recent years, transformer-based models have become the default solution for NLP tasks such as search, machine translation, or sentiment analysis (Tunstall et al., 2022). Only a few studies apply language models to the sustainability area. ClimateBERT, proposed by Bingler et al. (2021), analyzes companies' climate risk using the Task Force on Climate-Related Financial Disclosures framework. Another application, developed by Ulibarri et al. (2019), is an artificial neural network classifier for modeling environmental impact statement documents from the US Environmental Protection Agency. Finally, Nugent et al. (2020) demonstrate that fine-tuning BERT using large amounts of business and financial news data from the Reuters News

Archive led to better results with classification tasks such as detecting ESG controversies.

Terms-based sentiment analysis. Term-based sentiment analysis is particularly valuable in domain-specific text, which very much resembles how a human domain expert comprehends this text content. Domain-specific text such as sustainability reporting documents are very complex, often ambiguous, and may have multiple target terms in a single sentence. Moreover, the same terms may have different meanings or polarity depending on the context in which they appear (Ulibarri et al., 2019), demanding a different approach. Zhang et al. (2022) show that previous methods for aspect-based sentiment models are unable to achieve the same performance as human-level sentiment understanding. Additionally, Bahdanau et al. (2015) argue that basic encoder-decoder architecture with a fixed-length vector is a bottleneck in improving those models' performance. Inspired by the above research, both aspect-based sentiment and transformers-based architectures, we proposed a novel architecture that addresses the issue of long and complex sentences by expending the ABSA to emphasize parts of a source sentence that are relevant to predicting ESG sentiment.

3 Methodology

Most aspect-based sentiment analysis methodologies comprise multi-grained NLP tasks and consist of two major subtasks: target term extraction and sentiment classification (Yang et al., 2021). Accordingly, this section introduces our approach for ESG terms selection and extraction and presents the model design for conducting ESG sentiment classification.

3.1 ESG Taxonomy Development and Extraction

ESG taxonomy. This work uses an ESG risk taxonomy or collection of ESG terms based on the International Finance Corporation's (IFC) Environmental and Social Performance Standards and Corporate Governance Methodology.[1] The eight Environmental and Social Performance Standards and the six Corporate Governance Methodology parameters provide the highest level of aggregation of the taxonomy. The lowest level comprises 1,200 unique ESG risk terms (with more than 4,750 variations, including acronyms, abbreviations, and spelling variants). This taxonomy organizes information by IFC performance Standards, ESG subthemes, and topics and is compatible with sustainability disclosure standards such as the UN SDGs, the Global Reporting Initiative (GRI), and the Sustainability Accounting Standards Board (SASB) framework. Details on the whole structure of the taxonomy can be viewed in Appendix A.

ESG terms selection. Three rules govern the creation of the ESG term taxonomy. First, the relevance of the term within the text to ESG context, such as "endangered species," "child labor," "water pollution," "climate change," "biodiversity impacts," or "gender-based violence." Second, avoidance of broader concepts and stop words. For example, rather than use words like "water," we use specific composites such as "potable water," "water pollution," and "drinking water." Third, the use of nouns rather than adjectives as adjectives may qualify a wide variety of nouns, are often unspecific, and can increase instances of false positives. In addition to these rules, we use unsupervised machine learning techniques to add new risk terms and incorporate emerging ESG topics.

3.2 Sustainability-Domain Model Architecture

Problem statement and ESG sentiment definition. A **positive** ESG sentiment is a statement that expresses the perception of a company's or project's positive impact(s) on society or the absence of ESG risk. For instance, a statement such as "The company managed to significantly limit the risk of child labor in the supply chain" is considered positive in line with IFC's ESG standards. In contrast, a negative ESG sentiment is a statement that indicates a lack of compliance with IFC's ESG standards or the occurrence of an ESG risk event. For instance: "Evidence has surfaced of a

[1] IFC's Performance Standards on Environmental and Social Sustainability are a global benchmark for sustainability practices. To date, 130 financial institutions in 38 countries have adopted the Equator Principles, based on these standards. Leading development institutions—including the European Bank for Reconstruction and Development and the Asian Development Bank—adopted practices rooted in these standards. Between 2006 and 2016, an estimated US$4.5 trillion in investments across emerging markets adhered to IFC's standards or to principles inspired by them (Corporation, 2016). In 2011, IFC was the first development financial institution (DFI) to require corporate governance analysis for every investment transaction as part of its due diligence process. IFC's Corporate Governance Methodology evaluates the corporate governance risks and opportunities of client companies. It was distilled into the Corporate Governance Development Framework used by 34 DFIs in their investment processes

widespread use of child labor in the cocoa sector in emerging markets". **Neutral** ESG sentiments are factual statements that either refer to an ESG context but do not express positive or negative sentiments or are irrelevant in the ESG context. ESG terms used for labeling purposes do not per se imply positive or negative sentiments, even if a word may be considered positive (e.g., training) or negative (e.g., penalties and fines). Only the context in which these terms are used matters. Therefore, while the term "child labor" may be linked with a negative sentiment, stating **the absence of child labor** expresses a positive ESG sentiment. Finally, the sentence's structure can be complex, with multiple target terms. For instance: "The world's largest chocolate manufacturers provided support in addressing large-scale deforestation in the cocoa sector, but there is still evidence of child labor in the supply chain." When considering "deforestation" and "child labor", a traditional sentiment classification will fail to identify the correct sentiments. Hence the need to develop an approach which can handle the complexity and potential ambiguity of words and sentences expressing ESG sentiments.

The new approach. To meet this challenge, we propose to extend previous aspect-based sentiment works (Tang et al., 2016; Zhang et al., 2022) by enabling the transformer-based model to automatically and explicitly emphasize parts of a source sentence that are relevant to predicting a target word polarity. We call this novel architecture **ESG terms attention augmentation**. It is inspired by the success of attention-based neural networks in machine translation (Bahdanau et al., 2015) and Computer Vision (Bello et al., 2019). Its design and functioning are described in detail below.

A sentence-aspect pair (S, A_t) is given. The sentence is represented as $S = \{w_1^s, w_2^s, w_3^s, ..., w_n^s\}$ which consists of series of n words. The ESG aspect, also called a risk term is denoted as $A_t = \{w_1^a, w_2^a, w_3^a, .., w_t^a\}$ which is a part of S. A sentence S may consist of one or more ESG risk terms. STBSA aims to build a sentiment classifier that can precisely predict the ESG sentiment of sentence S for a specific ESG risk term, including multiple target terms with different sentiments. The overall architecture of the STBSA model, adapted from Zhang et al. (2022), is illustrated in Figure 1.

ESG terms attention augmentation. Because a sentence may contain multiple target terms that describe different sentiments that are difficult to predict using BERT or RoBERTa, we propose an innovative approach to achieve STBSA via transformer-based models. (Sun et al., 2019) and (Zhang et al., 2022) show improvements to the attention mechanic for sentiment analysis tasks based on transformer models by constructing an auxiliary sentence in addition to the original sentence. Similarly, we annotate and copy target words from sentences during pre-processing and create two copies of such terms in the sentence—one at the beginning and one in its original position. This modification of the sentence structure has two advantages: First, since the text input is changed, the outputs of the transformer-based model differ. Second, since an additional target term appears at the beginning of the sentence, its frequency increases and gains more attention in the model.

Human expert annotations. We designed a rigorous process to prepare a human-annotated training dataset with the labeling rules described in annotator guidelines. Three criteria are used to select the ESG documents to annotate: Relevance, Reliability, and Vintage. Content relevance is determined by the potential of text to support decisions, such as company sustainability reports and ESG-related news reports. Reliability refers to a qualified source of data and analysis prepared or reviewed by ESG experts. Data vintage is ascertained by using current sources, with a preference for the most recent data. The training dataset comprises three ESG sentiment types – positive, negative, and neutral – which are manually assigned to each sentence based on the targeted term.

Model fine-tuning procedure. We embrace a data-centric artificial intelligence (AI) strategy by proposing a sustainability-domain algorithm based on high-quality labeled data provided by human experts. Our model uses a transfer learning technique, used with success in computer vision, to train a convolutional neural network on one task and then adapt it to a new task (Tunstall et al., 2022). The fine-tuned model comprises the model body (initially trained for masked word predictions) and the custom classification head. During transfer learning, the body weights from general-purpose language models (the RoBERTa and DeBERTa corpus) are used for initialization, a starting point to create the sustainability domain-specific model based on the custom ESG taxonomy and human-annotated ESG data.

Hyperparameter settings. Meta AI and Mi-

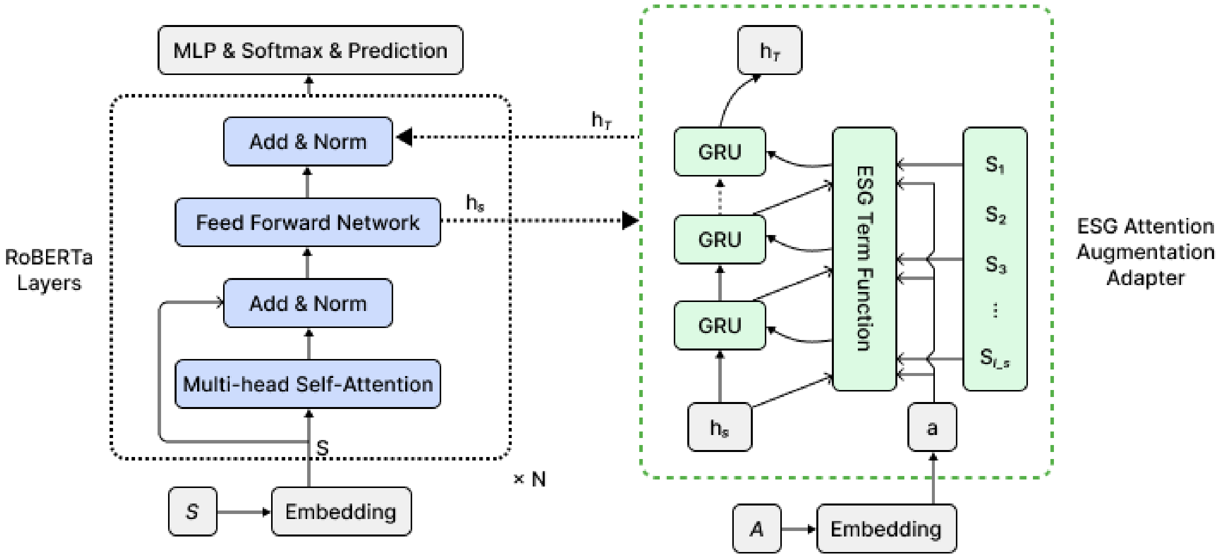

Figure 1: shows the STBSA framework adapted from Zhang et al. (2022). The blue blocks are the pre-trained RoBERTa model, which is frozen during the fine-tuning steps. The right green blocks represent the "ESG term attention augmentation" modifications performed during the fine-tuning, on top of the RoBERTa layers and with ESG-expert annotated data.

crosoft released the pre-trained RoBERTa and DeBERTa models on Hugging Face. [2] Our best performing fine-tuned RoBERTa is composed of the pre-trained RoBERTa layers and a custom classification head, consisting of two hidden layers (of 786 and 56 dimensions) and a softmax output layer (of 3 dimensions). The best and most stable model was found with 8 epochs, 0.1 dropout rate, 32 as batch size, 5e-6 as learning rate, 42 seed values, and 800 of the model's warm-up steps. We used the warm-up optimization strategy (He et al., 2016) by training the model with a varying learning rate along with all the training steps. A linear scheduler initialized the learning rate with a value near zero. After 800 training steps, the learning rate reached a preset peak value (5e-6) and slowly decreased.

3.3 Machine Learning Operations and Bias Management Process

Experiment context. As NLP models have shown a good level of accuracy in classifying general English language sentiments, we were challenged by the black-box nature of the neural models and inherent bias that training data poses. This motivated us to start developing a Proof of Concept (POC), led by the World Bank Group Technology and Innovation Lab, which successfully validated the use of LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017), and Fairlearn (Bird et al., 2020) in understanding the model behavior and fine-tuning the model to avoid bad bias.

Machine learning operations (MLOps). Training models to achieve acceptable accuracy and F1-scores requires robust processes to monitor data drift and retrain models to perform consistently on new input data. Such methods must include approaches to understand model biases and explain performance. Our research advances the use of Explainable AI frameworks and techniques to improve understanding of model performance. A mature MLOps and data management process is the cornerstone of training a trustworthy and fair model (Schwartz et al., 2022). Our experiments applied the MLOps process described in Figure 2. This approach has four domains: Domain Data, Data Science, Trust Analysis, and Consumption. All four parts maintain feedback loops to each other to achieve the overall objective of increasing the quality of ML inferences.

Figure 2 describes the process, which starts with domain data experts collecting, cleaning, and analyzing input data. Labeled data is quality assured by evaluating inter-annotator agreements. This approach prevents individual labeler bias from impacting the model. Next, the data science stage focuses on training and testing the model with labeled data. Section 4.2 describes model selection and performance metrics. Following this, the Trust Analysis step centers on model evaluation. This step

[2] RoBERTa base: https://huggingface.co/roberta-base; DeBERTa base: https://huggingface.co/microsoft/deberta-base.

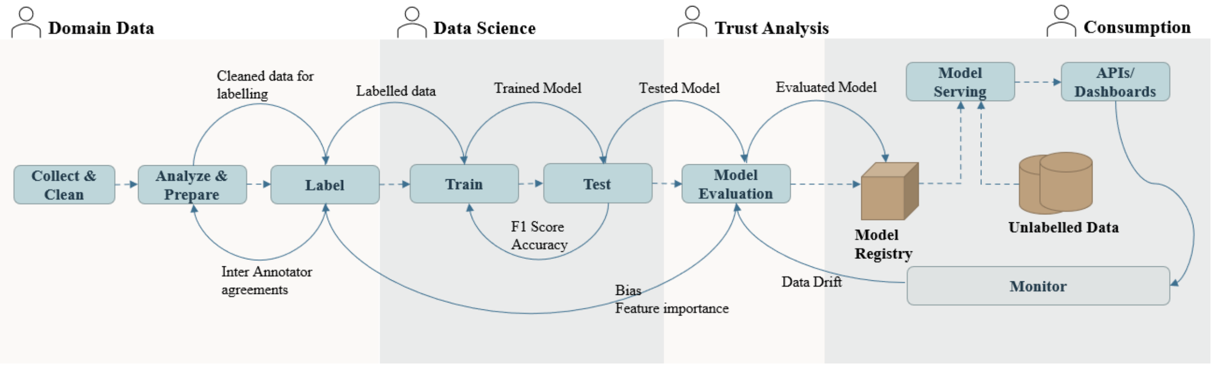

Figure 2: Phases of MLOps

determines if the model has any unforeseen biases that may skew the results. We experimented with LIME (local interpretable model-agnostic explanations, Ribeiro et al. (2016)) and SHAP (Shapley additive explanations, (Lundberg and Lee, 2017)) to understand how the model makes predictions. Model sensitivity analysis and feedback are provided to domain experts and data scientists to adjust the labeled data and model architecture.

Lastly, models are published in the model registry for the final step, Consumption. The model serving component uses the most recent version of the model from the model registry and predicts outcomes on API or Batch requests. Subsequently, model monitoring provides feedback at the model evaluation stage to assess data drift. The key theme of this proposal is that any production-grade AI/ML system must be a multi-stakeholder and interdisciplinary undertaking. An MLOps model brings forth these experts systematically and collectively works to make the model's prediction more relevant to the business problem that the model is trying to address.

4 Experiment

4.1 Experiment Settings

Proposed Dataset. Using rules outlined in ESG sentiment annotation guidelines, six ESG analysts worked over 1.5 years to refine the ESG taxonomy and produce labeled data for model training. The final training dataset comprised 126,480 sentences taken from ESG news, IFC internal project documentation, project evaluations by the World Bank Group Independent Evaluation Group, IFC Compliance Advisor Ombudsman project assessment reports, and publicly available information, including IFC ESG project disclosures and public disclosures by listed companies including annual and sustainability reports. The labeled dataset is presented in Table 1.

Quality assurance of labeled data. ESG sentiment annotation guidelines and inter-annotator agreement metrics ensure the creation of high-quality training data. Only sentences with consensus from at least two labelers are eligible as training data to mitigate the risk of conflicting labels. Consistency of labeling among annotators or inter-annotator agreement is tracked using Cohen's kappa coefficient, which measures the reliability of agreement between two labelers, considering the possibility that agreements could occur by chance (Cohen, 1960). In addition to Cohen's kappa, the percentage of inter-annotator agreement is used as a secondary quality indicator. These annotator agreement metrics improve the consistency of training data and manage inevitable differences between annotators (Pustejovsky and Stubbs, 2012; Bobicev and Sokolova, 2017). The average Cohen's kappa value was 0.75, indicating substantial agreement among labelers.

Train, validation, and test datasets. The final labeled set of 126,480 sentences comprised 37,054 (29%) positive, 27,579 (22%) negative, and 61,847 (49%) neutral labels. We randomly split this set into 107,540 sentences (85%) designated for model training and validation and 18,940 sentences (15%) for model evaluation. The subsets' class distribution is similar to the final set labeled above.

4.2 Experiment Results

Pre-trained model performance. Table 2 shows accuracy and F1-scores for the pre-trained RoBERTa-base and DeBERTa-base models on the test set. As expected, pre-trained models poorly predict ESG sentiments without domain-specific

ESG document type	Sentence count	Positive labels	Negative labels	Neutral labels
ESG news report	35,560	33.38%	26.34%	40.26%
IFC internal project documents	29,796	34.80%	18.36%	46.83%
Public company disclosures	19,213	26.96%	10.44%	62.60%
Public DFI project disclosures	31,900	29.35%	15.86%	54.79%
Independent project evaluations	10,011	2.70%	56.65%	40.64%
Total	**126,480 (100%)**	**37,054 (29%)**	**27,579 (22%)**	**61,847 (49%)**

Table 1: ESG sentiment labeled dataset

training. Most predictions are neutral. Pre-trained models can assess context information in ESG text but are less successful at predicting positive and negative ESG sentiments as these models are not trained on these types of labels.

Baseline model performance. For a further baseline comparison, we used the Fin-BERT model (Araci, 2019) as a benchmark to compare the performance of our model. Three arguments justify this choice: the domain proximity of financial and sustainability reporting (Nugent et al., 2020; IIRC, 2011); the FinBERT model's availability and usage metrics on open-source platforms, notably on Hugging Face; and, most importantly, its use of similar sentiment classes (positive, negative, neutral). FinBERT shows 69 % accuracy and 54% F1-score on the test data. Compared with the pre-trained RoBERTa-base and DeBERTa-base models, Fin-BERT demonstrates better performance, particularly for negative and positive sentiment predictions.

ESG fine-tuned model performance. The four last lines of Table 2 show the accuracy and F1-score of fine-tuned models. Compared to the FinBERT baseline, we observe a significant increase in accuracy from 69% to 88% and F1-score from 54% to 84% for the RoBERTa-base model fine-tuned for ESG. The fine-tuned DeBERTa and FinBERT models show similar levels of accuracy and F1-score. These results demonstrate that after ESG-domain training, the models demonstrate improved performance. After additional modifications to input data to emphasize ESG terms (attention augmentation), we reached 91.30% accuracy and 90.2% F1-score with RoBERTa. Detailed metrics, including Precision and Recall of the STBSA model, are presented in Appendix C.

Adjusting for imbalanced training data. ESG sentiment classes are not distributed equally. This data structure is expected in the ESG domain because most ESG terms occur in neutral contexts. To address this imbalanced classification issue (Hovy and Prabhumoye, 2021), we under-sampled the neutral class to obtain a new data structure with 37,054 positive labels (36%), 27,579 negative labels (27%), and 37,000 neutral labels (36%). The experiment based on this data structure shows both accuracy and F1-score of 91%. These adjustments do not lead to a substantial performance gain and result in a significant loss of labeled data (20%). As a result, we decided to continue experimenting with the complete labeled data set.

4.3 Real-world deployment of the STBSA by IFC (World Bank Group)

Our STBSA model has been deployed in an IFC internal machine-learning platform called MALENA or Machine Learning ESG Analyst. The platform's primary use is support for ESG due diligence and impact assessment of IFC projects. As of September 2022, the model successfully analyzed more than 112,000 ESG-related text documents, including documents proprietary to IFC and public records disclosed through the IFC Project Information and Data Portal. The model identified more than 14 million ESG risk terms, with 3,318,476 detected in a positive context, 1,141,755 in a negative context, and 10,359,769 in a neutral context. ESG sentiment profiles for close to 8,533 companies in 175 countries, seven regions, and 33 investment sectors are derived from model inferences. An active learning mechanism allows expert IFC users to provide feedback on model predictions, leading to improvements in model performance.

5 Positive impact

5.1 Strengthen ESG due diligence and Impact Assessment

The MALENA platform offers a unique solution to sustainability-domain stakeholders (investors, regulators, project proponents, etc.) to better conduct ESG due diligence. It enables the use of NLP to

Models	**Accuracy(%)**	**F1-score (%)**
Pertained models		
RoBERTa-base (Liu et al., 2019)	68.00	27.00
DeBERTa-base (He et al., 2021)	17.00	10.00
Baseline model		
FinBERT (Araci, 2019)	69.23	54.07
ESG-fine-tune models		
RoBERTa-base + ESG-fine-tuning	88.00	84.00
DeBERTa-base + ESG-fine-tuning	87.00	82.00
FinBERT + ESG-fine-tuning	87.44	87.31
RoBERTa-base + ESG-fine-tuning+ **Attention Augmentation** **=proposed-STBSA**	91.30 %	90.20 %

Table 2: Experiment results. Table 2 shows the model's accuracy and F1-score for pretrained RoBERTa and DeBERTa, for the baseline model (FinBERT), and for our ESG fine-tuned models. Accuracy and F1-score are calculated based on the randomly selected 18,940 sentences, including 5,572 positive, 4,121 negative, and 9,247 neutral labels. The STBSA model Error analysis is presented in Appendix D - Table 4.

identify and manage ESG risks during project appraisal, to support early-stage Environmental and Social Impact Assessment (ESIA) review, and to monitor the evolution of climate coverage in the media in order to dynamically hedge climate change risk. For instance, a recent experiment conducted by Curmally et al. (2022) on a sample of 530 IFC projects demonstrated that project sentiment scores (derived from our STBSA on projects' early-stage assessment documents, namely ESIAs) perform efficiently as proxies for project risk assessments and to predict E&S performances. Such information is crucial for allocating resources and technical expertise, determining legal requirements, and creating extensive and thorough environmental and social action and remediation plans. Additionally, our model offers a new comprehensive framework and an efficient tool to measure with increased accuracy the positive impact of investments in sustainable activities, both in emerging and developed markets. As we approach 2030, an accurate sentiment profile can be used as a proxy to assess how and to what extent projects or investment benefit local communities and indigenous people, respect the natural environment and contribute to the SDGs.

5.2 Redirect financing to green investments

Investors can play an essential role in redirecting finance to emerging markets by aligning investment strategies with the SDGs. However, gaps in sustainability data and analytical capacity are significant blockers (IFC and Amundi, 2021). Research finds that unstructured data (news articles, annual, integrated, impact and sustainability reports,etc.) is underused in analyzing investment performance (Varco, 2016). Our model has a significant implication in helping investors evaluate to what extent their activities are aligning with and contributing to the SDGs. The proposed ESG taxonomy can be leveraged as a framework to detect investment opportunities in corporate disclosures, and check project, or portfolio SDG-alignment. Facilitating SDG-aligned financing for emerging markets has the potential to address the $4.2 trillion USD anual shortfall in investments needed to meet the SDGs (OECD, 2020). Further, our STBSA model allows rapid assessments of Task Force on Climate-Related Financial Disclosures (TCFD) documents and other corporate disclosures. Analysis of such texts can help align portfolios with the Paris Agreement on Climate Change(Kölbel et al., 2022) and redirect financing to green and climate-fostering investment (Rolnick et al., 2019). IFC intends to make our STBSA model, as well as MALENA's insights and analytical capabilities, available to institutional investors and asset managers to identify ESG risks better and construct SDG-aligned investment portfolios.

5.3 Offer a Climate Analytics Solution as a global public good

AI-based platforms like MALENA can play a transformative role in addressing the gaps in sustainability data and limited analytical capacity. By reviewing public unstructured text disclosures, they can also address gaps in emerging-market coverage. Our model handles capacity constraints associated with reviewing large amounts of text by conducting this first level of analysis at scale Stede and Patz (2021). Further, by structuring the review of these disclosures using IFC's longstanding, market-tested ESG taxonomy (based on IFC's

ESG standards and aligned with the SDGs), IFC offers its ESG expertise at a level only accessible with. Widespread use of the public good version of MALENA will democratize access to ESG capacity globally, given the significant overlap with IFC's target markets. The demonstration effect of creating bespoke AI solutions to address development problems is already contributing to a vibrant AI for SDGs ecosystem in the development finance community as several risk guarantee agencies, development banks, and export credit agencies are interested in learning from IFC's experience using AI.

6 Discussion and Conclusion

In this paper, we proposed a novel approach to realize a term-based sentiment analysis built on a unique ESG taxonomy to address the limitations of the aspect-based sentiment analysis models and off-the-shelf sentiment analyzers for sustainability-domain applications. Furthermore, using historical sustainability corpus data and expertise from a development finance institution (IFC), we produced an unprecedented human-annotated dataset of 125,000+ sentences for ESG sentiment classification. The subsequent experiments demonstrated the effectiveness of this model with an accuracy of 91.3% and a 90% F1-score, outperforming the current state-of-the-art baseline models by over 20 points (Araci, 2019). Our STBSA model addresses three challenges. First, it offers a new model design with capabilities to handle multiple target terms and different sentiments by leveraging an ESG domain-specific taxonomy with more than 1,200 ESG risk terms. Recent studies underscored the difficulties of developing sustainability domain-specific taxonomies (Nugent et al., 2020; Ulibarri et al., 2019; Lennox et al., 2019), which are blockers to building more efficient and better-performing models. Second, it proposes an unprecedented sustainability domain NLP model, which yields a far higher performance (91.3% accuracy, 90% F1-score) than baseline models such as FinBERT (Araci, 2019) or similar studies such as the ones presented by (Ulibarri et al., 2019) or (Bingler et al., 2021) with 70% and 75% accuracy respectively. Our model fills a critical research gap in the NLP literature. Third, for investors in emerging markets, it offers the potential to enhance ESG due diligence and impact assessments resulting in a positive impact for green investments and contributing to achieving the UN SDGs.

These findings, while promising, have limitations and create opportunities for future research. First, the model can only understand and predict ESG sentiment in English (about 75% of the corpus). There are obvious benefits to expanding its understanding to additional languages such as French, Mandarin, Portuguese and Spanish. Second, as our STBSA model is derived from "black box" systems, the explainability and transparency framework proposed in this paper needs to be fully implemented to enable users to understand its design, operation, and biases, and to trust its predictions. This paper emphasizes data-driven AI and keeping humans in the loop and proposes a new multi-stakeholder framework for operationalizing AI systems. It is essential to ensure that complex and computationally heavy models, such as illustrated in this paper, do not penalize developing countries with limited data, leading to model biases (Conforti et al., 2020). This awareness may help mitigate underlying word embeddings biases of pre-trained language models associated with specific demographics such as gender, ethnic minorities, and local communities (Hovy and Prabhumoye, 2021). This paper provides a first but decisive step toward further research at the intersection of NLP and ESG. We intend to partially release the model and ESG-annotated data as a public good to enable a strong baseline for sustainability domain research, given its major value for the research community either to replicate our approach or to stimulate further research. We hope the results and dataset inspire the NLP and sustainability research communities to actively explore how advanced language modeling can be applied to ESG and impact data to support creating solutions furthering the SDGs.

Acknowledgements

We appreciate the feedback from ESG Specialists and Data Scientists. The views and conclusions of this paper are those of the authors and do not reflect the official policy or position of IFC.

References

Abdulaziz Alghunaim. 2015. *A vector space approach for aspect-based sentiment analysis*. PhD dissertation, Massachusetts Institute of Technology.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *ArXiv*, abs/1908.10063.

Ayoub Bagheri, Mohammad Hossein Saraee, and Franciska de Jong. 2013. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowl. Based Syst.*, 52:201–213.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. 2019. Attention augmented convolutional networks. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3285–3294.

Julia Anna Bingler, Mathias Kraus, Markus Leippold, and Nicolas Webersinke. 2021. Cheap talk and cherry-picking: What climatebert has to say on corporate climate risk disclosures. *Corporate Finance: Governance*.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Victoria Bobicev and Marina Sokolova. 2017. Inter-annotator agreement in sentiment analysis: Machine learning perspective. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 97–102, Varna, Bulgaria. INCOMA Ltd.

Samuel Brody and Noemie Elhadad. 2010. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 804–812, Los Angeles, California. Association for Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Costanza Conforti, Stephanie Hirmer, Dai Morgan, Marco Basaldella, and Yau Ben Or. 2020. Natural language processing for achieving sustainable development: the case of neural labelling to enhance community profiling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8427–8444, Online. Association for Computational Linguistics.

International Finance Corporation. 2016. Sustainability is opportunity: How ifc has changed finance.

Atiyah Curmally, Blaise W. Sandwidi, and Aditi Jagtiani. 2022. *Chapter 9: Artificial intelligence solutions for environmental and social impact assessments*, pages 163 – 177. Edward Elgar Publishing, Cheltenham, UK.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

David John Griggs, Mark Stafford-Smith, Owen Gaffney, Johan Rockström, Marcus C. Öhman, Priya Shyamsundar, Will Steffen, Gisbert Glaser, Norichika Kanie, and Ian R. Noble. 2013. Policy: Sustainable development goals for people and planet. *Nature*, 495:305–307.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Eduard H. Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

IFC and Amundi. 2021. Artificial intelligence solutions to support environmental, social, and governance integration in emerging markets.

IIRC. 2011. Towards integrated reporting: Communicating value in the 21st century. *International Integrated Reporting Council*.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies, pages 151–160, Portland, Oregon, USA. Association for Computational Linguistics.

Julian F Kölbel, Markus Leippold, Jordy Rillaerts, and Qian Wang. 2022. Ask BERT: How Regulatory Disclosure of Transition and Physical Climate Risks Affects the CDS Term Structure. *Journal of Financial Econometrics*. Nbac027.

Himabindu Lakkaraju, Richard Socher, and Chris Manning. 2014. Aspect specific sentiment analysis using hierarchical deep learning. In *Annual Conference on Neural Information Processing Systems (NIPS), Workshop on Deep Learning and Representation Learning, 2014*, pages 1–9, Montreal, Canada.

Robert J. Lennox, Diogo Veríssimo, William M. Twardek, Colin R. Davis, and Ivan Jarić. 2019. Sentiment analysis as a measure of conservation culture in scientific literature. *Conservation Biology*, 34.

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. *CoRR*, abs/1907.11692.

Bin Lu, Myle Ott, Claire Cardie, and Benjamin Ka-Yin T'sou. 2011. Multi-aspect sentiment analysis with topic models. *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 81–88.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *ArXiv*, abs/1705.07874.

Margaret Mitchell, Jacqui Aguilar, Theresa Wilson, and Benjamin Van Durme. 2013. Open domain targeted sentiment. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1654, Seattle, Washington, USA. Association for Computational Linguistics.

United Nations. 2015. Transforming our world: The 2030 agenda for sustainable development.

Måns Nilsson, David John Griggs, and Martin Visbeck. 2016. Policy: Map the interactions between sustainable development goals. *Nature*, 534 7607:320–2.

Timothy Nugent, Nicole Stelea, and Jochen L. Leidner. 2020. Detecting esg topics using domain-specific language models and data augmentation approaches. *ArXiv*, abs/2010.08319.

OECD. 2020. *Global Outlook on Financing for Sustainable Development 2021*. Organisation for Economic Co-operation and Development (OECD), Paris.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning - a Guide to Corpus-Building for Applications*. O'Reilly.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, Alexandra Sasha Luccioni, Tegan Maharaj, Evan D. Sherwin, Surya Karthik Mukkavilli, Konrad Paul Kording, Carla P. Gomes, Andrew Y. Ng, Demis Hassabis, John C. Platt, Felix Creutzig, Jennifer T. Chayes, and Yoshua Bengio. 2019. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55:1 – 96.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods. In *COLING*.

Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. 2022. Towards a standard for identifying and managing bias in artificial intelligence.

Alik Sokolov, Jonathan Mostovoy, Jack Ding, and Luis Seco. 2021. Building machine learning systems for automated esg scoring. *The Journal of Impact and ESG Investing*.

Manfred Stede and Ronny Patz. 2021. The climate change debate and natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 8–18, Online. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. In *NAACL*.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2022. *Natural Language Processing with Transformers: Building Language Applications with Hugging Face*. O'Reilly Media, Incorporated.

Peter Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual*

Meeting of the Association for Computational Linguistics, pages 417–424, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Nicola Ulibarri, Tyler A. Scott, and Omar Perez-Figueroa. 2019. How does stakeholder involvement affect environmental impact assessment? *Environmental Impact Assessment Review*, 79:106309.

Chris Varco. 2016. The value of esg data: Early evidence for emerging markets equities.

Ricardo Vinuesa, Hossein Azizpour, Iolanda Leite, Madeline Balaam, Virginia Dignum, Sami Domisch, Anna Felländer, Simone Daniela Langhans, Max Tegmark, and Francesco Fuso Nerini. 2020. The role of artificial intelligence in achieving the sustainable development goals. *Nature Communications*, 11.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *IJCAI*.

Heng Yang, Biqing Zeng, Jianhao Yang, Youwei Song, and Ruyang Xu. 2021. A multi-task learning model for chinese-oriented aspect polarity classification and aspect term extraction. *Neurocomputing*, 419:344–356.

Kai Zhang, Kun Zhang, Mengdi Zhang, Hongke Zhao, Qi Liu, Wei Wu, and Enhong Chen. 2022. Incorporating dynamic semantics into pre-trained language model for aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3599–3610, Dublin, Ireland. Association for Computational Linguistics.

Appendix

A Structure of the ESG taxonomy

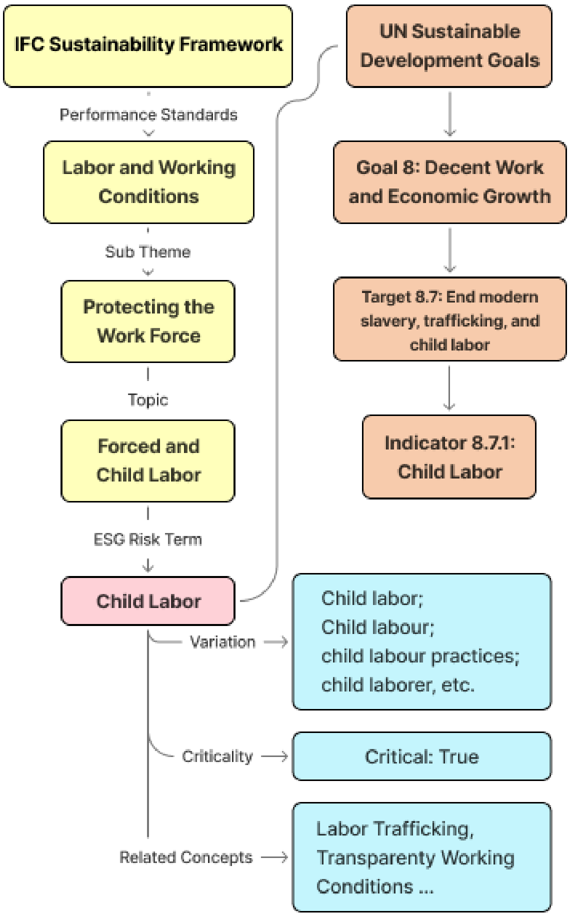

Figure 3: This figure shows the different levels of the ESG taxonomy used to train our STBSA for one ESG risk term, here "Child Labor". This structure includes the IFC Sustainability Framework (top level), the IFC Performance Standards and Corporate Governance Methodology, a Subtheme, a Topic, a target ESG Risk Term (here "Child Labor"), and its variations and related terms. This figure also provided indications on how the target ESG term “Child Labor” is mapped to the United Nations Sustainable Development Goals (SDGs), notably to SDG 8 (Decent Work and Economic Growth), to the Target 8.7 (End modern slavery, trafficking, and child labor) and to the indicator 8.7.1 (Child Labor).

B Ethical and Societal Implications

AI Models that are trained to achieve higher levels of statistical accuracy. While that is important, this research's focus on MLOps, the Trust Analysis framework acknowledges the existence of bad bias in the data and strives to reduce Ethical and societal impact. Without a strong MLOps and Trust analysis framework, machine learning models have the potential to yield statistically high performance but are ethically poor. This paper presents humans in the loop to ensure trained models do not exhibit bad bias. The proposed framework is explained in section 3.3.

C Detailed Metrics for the Sustainability Term-Based Sentiment Analysis (STBSA) Model

Appendix C presents the model Precision and Recall for each sentiment class: Positive, Neutral, and Negative (see Table 3 - Panel A). Additionally, the appendix shows the STBAS model performance over three different aspects, namely Environmental and Social, Corporate Governance, and Climate Change. This subdivision intends to identify any underperformance of the model and determine if there are systemic biases related to a particular aspect of the three pillars composing the Environmental, Social, and Governance domains (see Table 3 - Panel B).

D The Sustainability Term-Based Sentiment Analysis (STBSA) Model Error Analysis

Appendix D Table 4 displays three review examples and their prediction results by the RoBERTa-base model, FinBERT, and our STBSA. As we can see from the "RoBERTa-base" column when there are multiple target terms, the vanilla RoBERTa makes the wrong classification; this model is not trained to classify sustainability term-based sentiment analysis. Fin-BERT, to some extent, is able to predict certain ESG sentiments correctly but fails the sentence with multiple ESG terms with different sentiments.

Panel A: Sentiment Class	**Samples**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Positive	5,572	87.20	92.70	89.80	
Neutral	9,247	92.60	88.30	90.40	
Negative	4,121	89.40	91.00	90.20	
Micro-Avg	18,940	90.20	90.20	90.20	
Macro-Avg	18,940	89.70	90.70	90.10	
					91.30

Panel B: Label Type	**Samples**	**Accuracy**	**F1-Score**
Environmental and Social	14,413	90.50	90.5
Corporate Governance	1,165	89.00	88.4
Climate Change	3,362	89.10	88.50

Table 3: The panel A of this table presents the model Precision and Recall for each class sentiment class (Positive, Neutral, and Negative) Accuracy and F1-score are calculated based on the randomly selected 18,940 sentences, including 5,572 positive, 4,121 negative, and 9,247 neutral labels. The panel B shows the detailed metrics for the Sustainability Term-Based Sentiment Analysis (STBSA) Model. The model Accuracy and F1-score are calculated based on the randomly selected 18,940 sentences, including 14,413 environmental and social labels, 1,165 corporate governance labels, and 3,362 climate change-related labels

Case Examples: The label in brackets represents the ground truth provided by ESG analysts	RoBERTa-base	FinBERT	STBSA
ESG terms: "communities" (**Pos**), "displacement" (**Neg**), "armed conflict" (**Neg**) **Sentence**: We intend to maintain our support for extending the benefits and services of the state to **communities** that have been historically marginalized and communities that have been significantly impacted by the **displacement** and the violence of the **armed conflict**.	**Pos/ Neg/ Neg** ✗ ✗ ✗	**Pos/ Neg/ Neg** ✓ ✗ ✗	**Pos/ Neg/ Neg** ✓ ✓ ✓
ESG terms: "deforestation" (**Pos**), "child labor" (**Neg**) **Sentence**: World's largest chocolate manufacturers provided support in addressing large-scale **deforestation** in the cocoa sector, but there is still evidence use of **child labor** in the supply chain.	**Pos/ Neg** ✗ ✗	**Pos/ Neg** ✓ ✗	**Pos/ Neg** ✓ ✓
ESG terms: "Sustainability" (Neu), "climate change" (**Neg**) **Sentence**: The Head of the Communication and **Sustainability** Office agreed, saying that the **climate change** is one of the greatest threats to life on earth with alarming and long-term effects.	**Neu/ Neg** ✓ ✗	**Neu/ Neg** ✓ ✗	**Neu/ Neg** ✓ ✓

Table 4: Error analysis of three sentences with multiple target ESG terms. The colored words in parentheses represent the ground truth provided by IFC's ESG analysts. The symbol ✓ means the predicted sentiment is correct, and the symbol ✗ means the predicted sentiment is wrong