Funk, Christoph; Tönjes, Elena; Haas, Christian

Working Paper

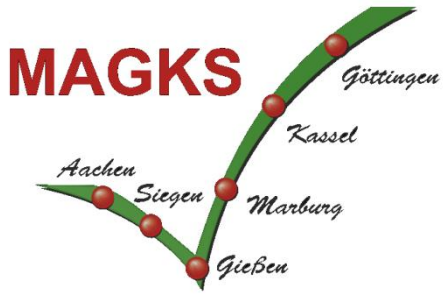
Exploring the predictive capacity of ESG sentiment on official ratings: A few-shot learning perspective

MAGKS Joint Discussion Paper Series in Economics, No. 12-2024

Provided in Cooperation with:
Faculty of Business Administration and Economics, University of Marburg

Suggested Citation: Funk, Christoph; Tönjes, Elena; Haas, Christian (2024) : Exploring the predictive capacity of ESG sentiment on official ratings: A few-shot learning perspective, MAGKS Joint Discussion Paper Series in Economics, No. 12-2024, Philipps-University Marburg, School of Business and Economics, Marburg

This Version is available at:
https://hdl.handle.net/10419/301239



Joint Discussion Paper Series in Economics

by the Universities of

Aachen · Gießen · Göttingen
Kassel · Marburg · Siegen

ISSN 1867-3678

No. 12-2024

Christoph Funk, Elena Tönjes, and Christian Haas

Exploring the Predictive Capacity of ESG Sentiment on Official Ratings: A Few-Shot Learning Perspective

This paper can be downloaded from

https://www.uni-marburg.de/en/fb02/research-groups/economics/macroeconomics/research/magks-joint-discussion-papers-in-economics

Coordination: Bernd Hayo • Philipps-University Marburg
School of Business and Economics • Universitätsstraße 24, D-35032 Marburg
Tel: +49-6421-2823091, Fax: +49-6421-2823088, e-mail: hayo@wiwi.uni-marburg.de

Exploring the Predictive Capacity of ESG Sentiment on Official Ratings: A Few-Shot Learning Perspective

Christoph Funk [1] **Elena Tönjes** [2] **Christian Haas** [3]

[1]Justus Liebig University Giessen
Centre for International Development and Environmental Research (ZEU)
Senckenbergstrasse 3
35390 Giessen, Germany
[2]Justus Liebig University Giessen
Faculty of Economics and Business Studies
Licher Strasse 64
35394 Giessen, Germany
[3]Frankfurt School of Finance & Management
Adickesallee 32-34
60322 Frankfurt am Main, Germany
Christoph.Funks@wi.jlug.de, Elena.Tönjes@wi.jlug.de, c.haas@fs.de

Abstract

Environmental, social, and governance (ESG) criteria are increasingly central to corporate reporting. This study applies natural language processing (NLP) techniques, specifically a RoBERTa-based few-shot model, to conduct aspect-based sentiment analysis (ABSA). Our analysis targets ESG-related entities and their sentiments within EUROSTOXX 50 company reports to assess their impact on ESG ratings. The ratings data are sourced from established providers, including Refinitiv, S&P, and Bloomberg. Furthermore, to explore the potential reciprocal influences on these variables, we employ a vector auto-regressive (VAR) model, which facilitates the modeling of bidirectional interactions. This combination of advanced NLP methods and comprehensive data integration aims to provide detailed insights into the dynamics between company disclosures and rating providers' ESG scores. The results of our study indicate that in general there is no discernible relationship between the ESG sentiment as reflected in company reports on the EUROSTOXX50 and the ESG ratings provided by the rating agencies. Nevertheless, our tool can provide an alternative, fine-grained measure of companies' own views on ESG-related matters.

1 Introduction

Environmental, social, and governance (ESG) factors have become increasingly prominent in corporate reporting since their introduction in 2004 (UN 2004). This trend highlights the increasing importance of ESG considerations for both companies and stakeholders, as evidenced by the development of reporting standards such as those proposed by the Task Force on Climate-related Financial Disclosures (TCFD) (TCFD 2017). The manner in which companies report on ESG issues can significantly impact perceptions, observable in metrics like stock prices and, most notably, ESG ratings. The improvement of companies' ESG standards is often evaluated through official ratings from agencies like Standard & Poor's (S&P). However, the reliability and consistency of these ESG ratings have been questioned, paving the way for alternative, text-based indicators. Company reports contain extensive information about their ESG measurements and perspectives (Berg, Koelbel, and Rigobon 2022). Natural language processing (NLP) offers a powerful solution for efficiently analyzing large volumes from these reports without the need for manual review (Schimanski et al. 2024).

As NLP has gained traction as a tool in recent years, a significant body of literature has emerged on ESG-related topics in various sources, including corporate reports, academic papers, or news data. To date, the most common methods for text classification involve fine-tuned transformer models for classification, generative prompt-based models such as GPT 3.5 and unsupervised methods like Latent Dirichlet Allocation (LDA). These models are generally employed to estimate the extent of ESG reporting and to relate it to other measures, such as ESG ratings or stock returns. A common text format utilized in this context is news data. For instance, Fischbach et al. (2023) employ NLP techniques to identify ESG-related news headlines. Subsequently, a BERT model, designated as the ESG-miner, is trained to identify company headlines and categorize them as ESG-relevant or not. An ESG score is then calculated based on the sentiment of the related headlines.

Moreover, the advent of OpenAI's ChatGPT has led to a surge in interest in prompt-based generative models. For instance, Jain et al. (2023) employ GPT-3.5 as an ESG classifier. The authors demonstrate a 20% correlation between company stock returns and ESG news, suggesting that that their query-based ESG classifier can accurately identify ESG factors. This capability can assist investors making more informed decisions. Moreover, the authors of Föhr

et al. (2023) examine whether ChatGPT can be used as an auditing tool for sustainability reports, with the objective of assessing their compliance with the EU taxonomy.

Another frequently employed NLP technique in this context is topic modeling. Goloshchapova et al. (2019) applied LDA to the Corporate Social Responsibility (CSR) reports of several companies listed on major stock market indices in 15 industrialized countries. The findings indicate that certain topics, such as 'employee safety,' are more frequently addressed by companies in the UK and Europe. However, they also identify sectoral biases, with certain sectors focusing more on certain issues than others. In their study, Lee et al. (2023) employed BERTopic as a topic model to gain insight into ESG discourse. In contrast to the focus on corporate reporting, the authors employ news data from LexisNexis and academic papers from the Web of Science to identify differences in ESG discourse between these two sources.

Our study focuses on reports published by the companies themselves. These reports are generally broader in nature, providing an overview of the company's performance or events, such as annual reports. Alternatively, they may be more specific in their focus, addressing ESG-related issues in greater depth, such as dedicated ESG reports.

One format of ESG-related reporting is the TCFD report. In 2017, the TCFD published guidelines that include specific questions which can be addressed either in a dedicated TCFD report or within a company's broader report. These guidelines are recommendations and therefore not mandatory. Companies that support the TCFD guidelines are not required to address all of the issues identified by the TCFD. Consequently, numerous studies have employed NLP techniques to assess the extent to which companies adhere to the TCFD guidelines and broader ESG-related issues. For instance, Luccioni, Baylor, and Duchene (2020) sought to identify sections in corporate reports that address climate-related issues by training a RoBERTa Question-Answering Model. This model leverages the 14 TCFD questions and the corresponding text sections that answer these questions as training data. The model was then employed to ascertain whether there were any differences in the extent to which companies in different sectors addressed the 14 TCFD questions. Additionally, Bingler et al. (2022) utilized a BERT model to investigate whether companies might engage in selective reporting with regard to the 14 TCFD questions. The findings indicate that companies tend to omit information on strategy, metrics, and targets, suggesting a selective reporting strategy that prioritizes non-material risks while neglecting material risks. This behavior suggests that companies are engaging in cherry-picking when it comes to TCFD reporting. Additionally, Auzepy et al. (2023) employ a zero-shot model to analyze TCFD reports. The researchers developed fine-grained labels that align with the TCFD recommendations. Their findings revealed an increase in climate-related disclosures, although they also identified instances of selective reporting, indicating that some recommended topics may not have been fully addressed.

In their study, Friederich et al. (2021) trained a RoBERTa model and applied it to the company reports of 337 firms over a 20-year period. The model identified an overall increase in risk disclosure, with a particularly dynamic growth observed in transitional risks compared to physical risks. This conclusion was based on the observation that the mentions of risks, especially transitional risks, exhibited a pronounced increase around the year 2015.

This study builds upon the research presented in Schimanski et al. (2024), which examines the relationship between ESG ratings and reporting. The authors employ fine-tuned RoBERTa and DistilRoBERTa models to determine the relative amount of ESG reporting in corporate documents. To achieve this, one model is trained for each aspect of ESG on 2,000 sentences. Moreover, the authors of Schimanski et al. (2024) employ a fixed-effects panel data model for their time series analysis.

The objective of this study is to enhance the detection of ESG reporting through the use of an Aspect-Based Sentiment Analysis (ABSA) Few-Shot model. This approach not only requires less training data but also outperforms existing models on the same datasets. It offers a more detailed analysis by examining ESG subcategories and their respective sentiments.

In contrast to (Schimanski et al. 2024), we challenge the assumption of a unidirectional effect—from reporting to ratings—by utilizing a panel Vector AutoRegressive (pVAR) model for our time series analysis. This allows for the examination of reciprocal effects between variables, providing a more comprehensive understanding of the dynamic relationship between ESG reporting and ratings.

In addition, instead of merely examining the quantity of ESG reporting, we utilize a qualitative measure of sentiment toward ESG issues. This approach is more plausible given the limitations of a quantitative measure. For instance, if ESG reporting increases on sustainability, it is possible that only negative reporting increases. However, an increase in negative reporting should result in a decrease in the ESG score, rather than an increase. As mentioned by Schimanski et al. (2024), the relationship between reporting quantity and ESG scores is found to be positive, regardless of whether the reporting is negative or positive in tone. In contrast, an analysis of the tone of reporting indicates that a more negative sentiment should have a negative effect on ESG scores, while a more positive tone should have a positive effect on ESG scores. This aligns with the approach taken in this study, where we analyze the relationship between reporting sentiment and ESG scores.

Moreover, our ESG entity-based approach is more granular and provides a tool that outputs specific ESG entities with the corresponding sentiment. Stakeholders can use this

tool to gain insights from company reports without the need to manually read them and form impressions of which ESG entities might be particularly negatively or positively annotated for a specific company in a given year. Consequently, stakeholders are provided with a tool that offers insights into the specific ESG entities discussed in a report and their respective positive or negative sentiment. This contrasts with previous approaches, which only allowed for the analysis of the extent to which a report addressed the pillars of E, S, or G.

In essence, our study diverges from previous research in several key respects. We adopt a detailed approach by identifying ESG-related entities in corporate reports and categorizing them into subcategories, thereby moving beyond the analysis of the three main pillars of ESG reporting (environmental, social, and governance). Rather than focusing on the quantity of reporting, we examine the tone of ESG reporting, creating sentiment timelines for each ESG subcategory. This analysis covers reports from EUROSTOXX 50 companies from 1999 until 2023 across all report types. Furthermore, we investigate the bidirectional relationship between reporting and ratings from three major rating agencies (Refinitiv, Bloomberg, and S&P) through a panel vector auto-regressive (pVAR) model, recognizing that ratings can influence subsequent reporting. The results of our time series analysis indicate that there is no discernible relationship between our sentiment scores and the ESG scores provided by the rating agencies. This could be attributed to the questionable quality of ESG ratings (Berg, Koelbel, and Rigobon 2022; Schimanski et al. 2024), an insufficiently large data set, or a model misspecification. Nevertheless, our sentiment scores remain a valuable tool for gaining insights into the perspectives of companies on ESG issues.

Our contribution is four-fold: First, we show that our few-shot model works better on a much smaller training sample than larger models. Thus, we save human and computational resources due to less annotation and training time. Second, our model is more fine-grained than existing models and provides detailed insights into firms' views on ESG issues. Third, we demonstrate that ESG ratings issued by rating agencies fail to accurately reflect the sentiment of firms with regard to ESG issues. Furthermore, our findings suggest that other variables exert a more significant influence on the composition of ESG ratings, while the quality and consistency of ESG ratings are open to question. Fourth, we create a measure to extract the information on ESG issues contained in ESG reports, which can be used to analyze the reports on an ESG entity basis.

The remainder of this paper is organized as follows: In Section 2, we describe the data used to generate the sentiment indices and the ESG scores employed. In Section 3, we provide a brief overview of the methodologies employed for parsing the text data, the key functionality of the SetFit few-shot model, aspect-based sentiment analysis, and the integration of the two. Section 4 describes the tr aining process and the performance of the model. Section 5 presents the results of our ABSA model and pVAR, while Section 6 provides a conclusion.

2 Data and Methodology

ESG ratings

For our analysis, we used ESG scores from three different rating agencies, namely Refinitiv, Bloomberg and S&P. We collected the ESG scores from their respective platforms, with the data downloaded in May 2024. The timeframe for the Bloomberg ratings is from 2015 to 2022, providing us with 8 years of data. For S&P, we have ESG scores from 2014 to 2023, and for Refinitiv, from 2002 to 2023. However, we do not have the full timeframe for all companies in our dataset. A detailed overview of the dataset is provided in Tables 6, 7, and 8 in the appendix. It is important to note that for some companies, we have data for the entire time frame. However, in some cases, the ratings for certain reports are missing for certain years. From all rating agencies, we use annualized data for each ESG pillar E, S, and G. The scores range from 0 to 10 for Bloomberg and from 0 to 100 for Refinitiv and S&P. Additionally, for Refinitiv, we have data for subcategories within each pillar, but this data is only available from 2018 to 2022. To facilitate the results of our pVAR models, we have min-max scaled the rating scores between 0 and 1, as our sentiment scores range from -1 to 1.

Descriptive Statistics of the Dataset

In our study, we utilized reporting data from EuroSTOXX 50 companies. We downloaded all available reports from the Refinitiv Database for the period between January 1, 1999, and June 14, 2023. We identified and downloaded a total of 2,072 reports in PDF format across various reporting types. Of these, 14 were either empty or entirely corrupted. Additionally, 18 documents were not in English and were excluded as well. Consequently, 2,038 were machine-readable and free from coding errors, which were used for our further analysis. Table 1 summarizes the key statistics of the companies in our dataset, including the total number of reports, average number of reports per year, number of different report types, and the number of years with at least one report. Of the 50 companies in the EuroSTOXX 50, the total number of reports ranges from 9 for Adyen NV to 80 for Intesa Sanpaolo SpA. While the average number of reports per year ranges between 2 and 3 for most companies, each company has at least 2 different report types available. Although most companies have reports spanning ten or more years, two companies have reports for only 4 years (Prosus NV) and 5 years (Adyen NV).

Figure 1 provides a chart of the number of reports over the years. The chart illustrates the distribution of different report

types over time, highlighting the increase in reporting activity and diversity in recent years. We distinguish between eight different reporting types: Corporate Governance, ESG, Environment, Health and Safety Reports, Full Year, GRI Report, Remuneration Committee, Social Reports, and Sustainability Committee. One can clearly see that for the period between 1999 and 2010, there were not many different reporting types available in our dataset. Most reports from this period are annual reports, which we denote as Full Year. Notably, in 2005, there is a significant lack of reporting, with only 11 of the 50 companies having reports available in the Refinitiv database. Starting in 2011, there is a clear increase in ESG and Remuneration Committee reports, while Corporate Governance reports are only available between 2011 and 2014.

3 Methodology

Extracting text from PDFs

All of the reports utilized in our analysis are in PDF format. The conversion of textual information present within PDF documents into a format suitable for further NLP analysis is a more intricate process than that required for textual data stored in CSV or TXT files. In this paper, we apply a layout-parsing model, which is able to detect and extract actual text from PDF documents. In the context of our analysis, the text of the reports is included, while that of tables and graphs is deliberately omitted. This not only improves the quality of our data but also allows us to save computation time (Auzepy et al. 2023).

Our parsing model is based on Visual-Layout (VILA) groups introduced by Shen et al. (2021). VILA is able to convert textual data into groups of tokens (text lines or blocks) and to assign a layout tag to these tokens. There are several variants of VILA, called H-VILA (Visual Layout-guided Hierarchical Model) and I-VILA (Injecting Visual Layout Indicators). After several trials, we choose the H-VILA block variant trained on grotoap2 using the layoutLM model (Xu et al. 2020) as a base model since it delivered the best extraction and tokenization results. The output consists of the extracted text as groups of tokens together with the corresponding layout tags. Depending on the training set, the layout tags can be figures, body content, abstract and title. For our analysis, we keep the parts tagged as body content and abstract (Auzepy et al. 2023).

To implement this, we used several tools and models. We initialized the PDFExtractor with "pdfplumber" for extracting text and images from PDFs. The EfficientDetLayoutModel from layoutparser was used for detecting the layout of the documents. We then employed the HierarchicalPDFPredictor from the VILA library, specifically the "allenai/hvila-block-layoutlm-finetuned-grotoap2" model. However, in some instances, these extraction methods failed, so we used a fallback option to retrieve as much textual information as possible. First, in a few cases, coding errors appeared. In these instances, we simply skipped the problematic byte and moved on to the next one. For more complex cases, we integrated an OCR agent using Tesseract. This involved converting the PDF documents to images using the pdf2image library, after which OCR was performed. The disadvantage of this approach is that detecting text blocks and layout in each page of the PDF using the VILA model was no longer possible. Despite this limitation, our comprehensive approach ensures that we accurately and efficiently extract the relevant textual data from PDF reports, thereby improving the overall quality and reliability of our analysis (Auzepy et al. 2023).

Few-shot SetFit Model

In recent years, few-shot models have gained popularity due to their improving performance. Unlike standard models, which require thousands of data samples for training, few-shot and zero-shot models are advantageous as they require minimal to no labeled data, making them cost-effective and time-efficient. In the NLP context, zero-shot models do not require any labeled data for prediction, relying solely on semantic understanding. Conversely, few-shot models need only a small set of labeled data to outperform some standard models that depend on thousands of labeled training sentences.

Our analysis employs SetFit (Sentence Transformer Fine-tuning) (Tunstall et al. 2022), an efficient, prompt-free few-shot model using sentence transformers (ST) available on Hugging Face (https://huggingface.co/docs/setfit/index). The authors demonstrate that with merely 8 labeled sentences per class, SetFit surpasses the performance of a standard fine-tuned RoBERTa large model trained on a full set of three thousand examples (Tunstall et al. 2022). We further validate these findings, showing that with significantly fewer labeled sentences, SetFit achieves superior performance on the same datasets compared to the ESG RoBERTa and DistilRoBERTa models, each trained with two thousand sentences (Schimanski et al. 2024).

SetFit offers several advantages over existing few-shot models. When using RoBERTa as its base, SetFit outperforms smaller prompt-based models like GPT-3 and PET, although it does not surpass T-FEW (Liu et al. 2022). However, it is worth noting that SetFit is thirty times smaller than T-FEW, making it more compact while still delivering commendable performance without relying on prompts. This aspect is crucial as dependence on prompts, as seen with models like GPT-3, can lead to sensitive and unstable outcomes due to minor variations in wording (Tunstall et al. 2022).

The training of SetFit involves two key steps. Initially, the sentence transformers are trained in a Siamese manner on sentence pairs. Subsequently, a classifier head is trained using the encoded data from the first step. This bifurcation

Table 1: Company Summary with Report Types

Company Name	Total Reports	Average Reports per Year	Number of Report Types	Years with Reports
ASML Holding NV	58	2.4	4	24
AXA SA	46	1.9	4	24
Adidas AG	46	2.0	5	23
Adyen NV	9	1.8	2	5
Airbus SE	59	2.8	4	21
Allianz SE	65	2.7	5	24
Anheuser-Busch Inbev SA	54	2.3	2	23
BASF SE	36	1.5	5	24
BNP Paribas SA	40	2.4	4	17
Banco Bilbao Vizcaya Argentaria SA	30	2.3	4	13
Banco Santander SA	31	2.6	6	12
Bayer AG	38	1.6	4	24
Bayerische Motoren Werke AG	44	1.9	5	23
CRH PLC	45	2.0	4	23
Danone SA	32	1.8	4	18
Deutsche Boerse AG	36	1.6	4	22
Deutsche Post AG	34	1.5	3	23
Deutsche Telekom AG	45	2.0	6	23
Enel SpA	58	2.5	5	23
Eni SpA	54	2.3	7	23
EssilorLuxottica SA	43	2.1	4	20
Flutter Entertainment PLC	29	1.3	3	22
Hermes International SCA	21	1.1	2	20
ING Groep NV	37	1.6	3	23
Iberdrola SA	37	2.8	4	13
Industria de Diseno Textil SA	37	3.1	5	12
Infineon Technologies AG	41	1.8	3	23
Intesa Sanpaolo SpA	80	3.6	6	22
Kering SA	42	2.2	6	19
Koninklijke Ahold Delhaize NV	39	1.6	4	24
L'Air	38	1.9	2	20
L'Oreal SA	37	1.6	4	23
LVMH Moet Hennessy Louis Vuitton SE	53	2.9	3	18
Mercedes Benz Group AG	46	1.8	3	25
Muenchener Rueck	43	2.0	4	22
Nokia Oyj	53	2.2	5	24
Nordea Bank Abp	48	2.0	4	24
Pernod Ricard SA	28	1.9	2	15
Prosus NV	12	3.0	3	4
SAP SE	34	1.4	3	24
Safran SA	25	1.3	3	19
Sanofi SA	39	2.2	5	18
Schneider Electric SE	35	1.9	2	18
Siemens AG	43	1.8	4	24
Stellantis NV	27	2.5	5	11
TotalEnergies SE	36	1.8	5	20
UniCredit SpA	73	3.8	6	19
Vinci SA	31	1.9	2	16
Volkswagen AG	41	1.8	4	23
Vonovia SE	30	2.7	4	11

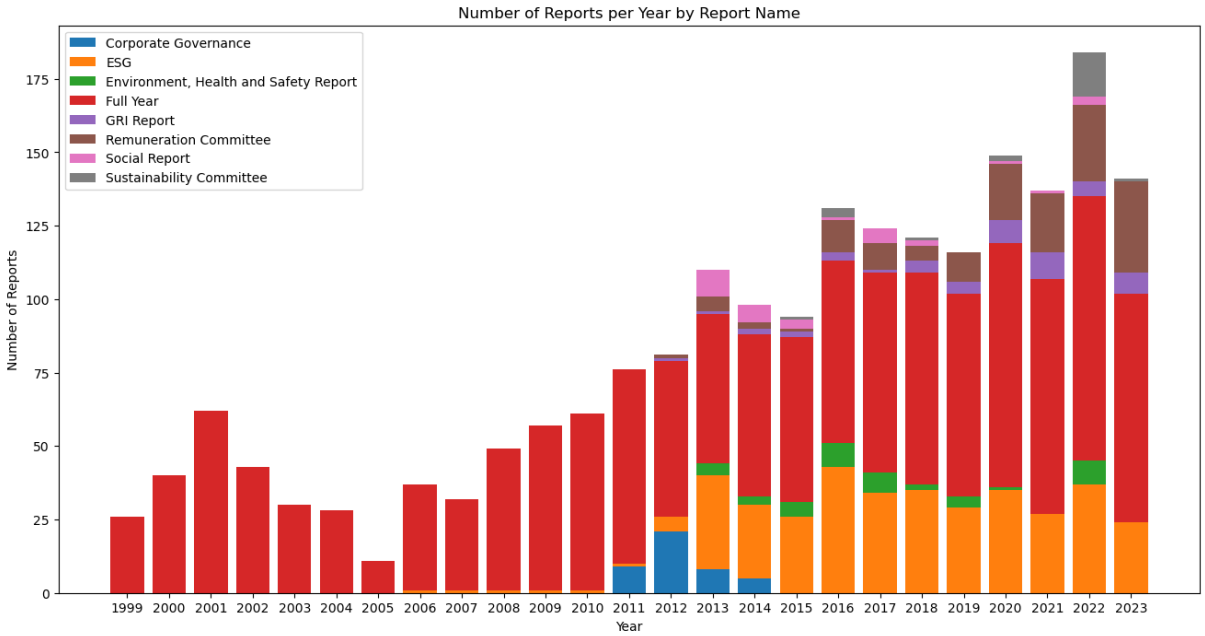


Figure 1: Number of Reports per Year by Report Name.

allows for a distinct separation between the ST fine-tuning phase and the classification head training phase, streamlining the process.

In the first stage, a contrastive training approach commonly employed in image similarity detection (Koch, Zemel, and Salakhutdinov 2015) is adopted to address the challenge of limited training data in few-shot scenarios (Tunstall et al. 2022). This approach utilizes a small set of K labeled examples, denoted as $D = \{(x_i, y_i)\}$, where x_i represents sentences and y_i their corresponding class labels. For each class label $c \in C$, a set of R positive triplets, $T_p^c = \{(x_i, x_j, 1)\}$, is generated, where x_i and x_j are randomly selected sentence pairs from the same class ($y_i = y_j = c$). Similarly, a set of R negative triplets, $T_n^c = \{(x_i, x_j, 0)\}$, is formed, where x_i are sentences from class c and x_j are sentences from different classes ($y_i = c, y_j \neq c$). The contrastive fine-tuning dataset T is then assembled by merging these positive and negative triplets across all classes: $T = \{(T_p^0, T_n^0), (T_p^1, T_n^1), \ldots, (T_p^{|C|}, T_n^{|C|})\}$, where $|C|$ denotes the number of class labels, $|T| = 2R|C|$ represents the total number of pairs in T, and R is a hyperparameter set to 20 in all evaluations as per Tunstall et al. (2022).

This contrastive training strategy effectively enlarges the training dataset in few-shot scenarios. Given a small number of labeled examples K for a binary classification task, the potential size of the contrastive fine-tuning set T is derived from the total number of unique sentence pairs that can be generated, amounting to $K(K-1)/2$, which significantly exceeds the original count of K samples (Tunstall et al. 2022).

In the subsequent stage, the fine-tuned sentence transformer (ST) encodes the original labeled training data $\{x_i\}$, producing a single sentence embedding per sample, denoted as $Emb^{x_i} = ST(x_i)$, where $ST()$ symbolizes the fine-tuned STs from the first step. These embeddings, alongside their corresponding class labels, form the training set for the classification head, $T^{CH} = \{(Emb^{x_i}, y_i)\}$, where $|T^{CH}| = |D|$ build the training set for the text classification step. A logistic regression model serves as the classification head throughout this model (Tunstall et al. 2022).

To perform inference with the trained model, the pre-fine-tuned sentence transformer (ST) first encodes an unseen input sentence, denoted as x_i, generating a sentence embedding. Following this, the classification head, which was trained in the preceding step, determines the class prediction for the input sentence based on its embedding. This process is formally represented as $x_i^{pred} = CH(ST(x_i))$, where CH represents the function used by the classification head to predict the class (Tunstall et al. 2022).

The SetFit model described above performs a standard text classification task. However, in this paper, we will use the ABSASetFit model, which classifies both an entity and the corresponding sentiment in a sentence. The modifications to the standard SetFit model will be explained in the following two sections.

Aspect-based Sentiment Analysis

In our analysis, we employed ABSA to extract ESG-related entities and their corresponding polarity. This methodology, augmented with insights from various studies (Zhang, Wang, and Liu 2018; Saeidi et al. 2016; Jo and Oh 2011; Pontiki et al. 2015, 2016) and methodologies from (Tunstall et al. 2022), helps us generate ESG criteria-specific sentiment time series. Furthermore, these time series are employed in our analysis of their relationship with ESG ratings. Compared to vanilla sentiment analysis, ABSA can extract a text's sentiment regarding a specific entity, such as a person, location, company, and more (Liu 2020).

ABSA is typically employed by businesses to gain insights into customer sentiment regarding specific aspects of products or services. Nevertheless, this form of enhanced sentiment analysis can also be beneficial for other domains, as any entity can be extracted from texts. In this study, we focus on entities related to ESG issues and the tones in which they are discussed—namely, positive, negative, or neutral.

Combining SetFit and ABSA

The SetFit model is tailored for ABSA tasks in a specialized variant, SetFitABSA, which is accessible via Hugging Face (https://huggingface.co/docs/setfit/how_to/absa). As ABSA models in particular demand a substantial quantity of labeled data, requiring annotators to identify both the entity in question and its sentiment within the training sentences, this process is particularly labor-intensive; therefore, a few-shot ABSA model like SetFitABSA can substantially reduce the effort and time required for annotation (Laperdon et al. 2023). In particular, ABSA models that are particularly traditional in nature require a substantial volume of labelled data, necessitating that annotators identify not only the entity in question but also its sentiment polarity within the training sentences. The labeling task is notably labor-intensive, and thus a few-shot ABSA model, such as SetFitABSA, can substantially reduce the effort and time required for annotation (Laperdon et al. 2023).

The core architecture of the SetFit model is retained, as outlined in section 3. In the ABSA framework, the few-shot model is deployed in two of three stages. Concisely, the process unfolds as follows:

1. Aspect Candidate Extraction: In the first stage, the candidate aspect or entity is extracted from the sentence. The 'SpaCy' library is used to tokenize the sentences and extract all nouns or noun compounds. Not all extracted nouns are actual aspects; therefore, they are referred to as aspect candidates (Laperdon et al. 2023).
2. Aspect Classification: In the second stage, a SetFit model determines whether the extracted candidate qualifies as an aspect. Training samples containing examples of aspect/non-aspect labels are needed for this step. Aspect candidates are merged with the entire training sentence to create a training instance following this template: `aspect_candidate:training_sentence` (Laperdon et al. 2023). For example, given the sentence "Waiters aren't friendly but the cream pasta is out of this world," assuming the nouns 'Waiters' and 'cream pasta' are aspects and 'world' is not an aspect, the templates would be:
   'Waiters: Waiters aren't friendly but the cream pasta is out of this world." with the label 1,
   "cream pasta: Waiters aren't friendly but the cream pasta is out of this world." with the label 1, and
   "world: Waiters aren't friendly but the cream pasta is out of this world." with the label 0.
   By training on such sentences, the model learns which aspect candidates are aspects or non-aspects (Laperdon et al. 2023).
3. Sentiment Classification: In the final stage, another instance of the SetFit model classifies the sentiment associated with the aspect. Training is similar to the aspect classification stage, but instead of a binary label, the label is one of three possible polarities: 'POS' for positive, 'NEG' for negative, and 'NEU' for neutral. Here, non-aspects are not included since only aspects are associated with polarities (Laperdon et al. 2023).

This streamlined approach to ABSA using SetFitABSA represents a significant advancement, leveraging the few-shot learning capabilities of SetFit to efficiently process and analyze sentiment with minimal labeled data. The authors claim that SetFitABSA, when applied to the SemEval14 ABSA Datasets 'Laptop14' and 'Restaurant14', SetFitABSA performs with a low number of training samples remarkably better than T5 (Raffel et al. 2023), despite being two times smaller, and better than GPT2-medium (Radford et al. 2019), even though being three times smaller. SetFitABSA even perfoms better than the 64 times bigger Llama2 (Touvron et al. 2023) when compared on equal training sample size (Laperdon et al. 2023).

4 Model Training

General Description of Model Training

In this study, we leverage our model to extract and analyze ESG aspects, along with their corresponding sentiment, from sentences within corporate reports. This approach offers a refinement over the methodology described by Schimanski et al. (2024), who categorize content strictly under the broad labels of 'E' (Environmental), 'S' (Social), and 'G' (Governance). Unlike these models, our technique seeks to uncover more nuanced ESG entities, subsequently mapping these to predefined ESG subcategories for a more granular analysis.

To construct our training sample, we employed sentences from the datasets provided by Schimanski et al. (2024). The original work involved labeling approximately 2,000 sentences per model, assigning a '1' to sentences addressing the respective ESG aspect (e.g., 'E' for Environmental) and '0' otherwise. It's noteworthy that while a sentence labeled '0' in the 'E' dataset might not discuss environmental issues, it could still pertain to social or governance themes. Each of the three models developed by Schimanski et al. (2024) focuses on a specific ESG aspect, yet their datasets largely comprise identical sentences.

Our methodology involved a more selective approach, utilizing roughly 100 sentences from each of Schimanski et al. (2024)'s datasets. For example, from the dataset designated for the 'E' aspect, approximately 100 sentences were selected that were marked '1,' indicating a focus on environmental concerns. In total, our environment training set comprises 105 unique sentences, our social training set 95 unique sentences, and our governance training set 102 unique sentences. The discrepancies arose because we determined that some of the sentences were not suitable for governance labeling and that some sentences for 'E' and 'S' were not included in the initial 100 sentences that were deemed important. The labeling process entailed the identification of the specific ESG aspect and sentiment within each sentence. This process was conducted by a team of three, comprising the authors of this paper, allowing for an initial intimate understanding and subsequent refinement of the labeling.

Our labeling process diverges from conventional methods by not starting with a predefined set of labels, as the relevant entities can vary significantly across sentences. Initially, one team member labeled the dataset to identify potential entities. In the second phase, a second reviewer examined the dataset for inconsistencies in labeling, which were then discussed and resolved collaboratively. Finally, a third reviewer, provided with a manual for labeling, compared their independent analysis with the previously labeled dataset to highlight and discuss discrepancies.

ESG-Subcategories, Entities and Sentiment

To achieve a more detailed analysis, which identifies potential underrepresentation of ESG-categories in reports, we employ predefined subcategories. These subcategories are detailed in Table 5, which outlines the ESG subcategories utilized in our labeling process. Although a variety of definitions exist, they generally converge on the same fundamental subcategories, exhibiting minimal variation. During the labeling process, these subcategories guided our selection of relevant entities from each category. The table also includes examples of entities from our training set.

In assessing the sentiment of sentences, context played a crucial role in determining the appropriate polarity. Sentences were deemed positive if they indicated that a company was taking steps to enhance its performance concerning ESG criteria. On their own, most sentences may appear neutral. For instance, a statement about a company offering employee training would typically be considered neutral. However, within the context of ESG implementation, such an initiative is regarded positively, leading us to label these sentences as positive. Conversely, sentences merely stating the existence of certain criteria without indicating the company's adherence were labeled neutral. Sentences that mentioned a company's failure to implement or improve upon ESG-related practices were labeled negative. Given the tendency of companies to enhance their image in such reports, the majority of sentences were labeled positive, resulting in imbalanced classes. This imbalance might adversely affect the model's ability to predict sentiment accurately, a challenge that is elaborated upon in section 4.

Training Process

Table 2: Training Results for Entity and Sentiment

Metric	Entitiy	Sentiment
Accuracy	0.9173	0.7917
Precision	0.9204	0.8096
Recall	0.9173	0.7917
F1 Score	0.9184	0.7999

A single base model was trained on our NVIDIA RTX A5000 GPU. While other models may exhibit slight improvements, a comprehensive five-fold cross-validation would be necessary to ascertain whether another model could be deemed significantly superior to the base model employed. The base model selected was 'sentence-transformers/paraphrase-mpnet-base-v2,' chosen due to its relatively fast training time of approximately five hours and ten minutes on the GPU. Additionally, the model exhibited satisfactory performance. However, larger models, which could potentially yield better results, cannot be trained on the GPU due to memory limitations. Further research may be warranted to investigate the potential for enhancing the model by selecting a more optimal base model.

A grid search was deemed unnecessary due to the impracticality of the extensive training process. Default parameters were used, consisting of one set for fine-tuning the sentence transformer and one for the classification head. For the sentence transformer, we used a batch size of 16, one epoch, and a body learning rate of 2e-05. For the classification head, we used a batch size of 2, 16 epochs, and a body learning rate of 1e-05. The default head learning rate for the entire model was set, and the CosineSimilarityLoss function was chosen for the entire model's loss function.

As a sampling method, oversampling was used to ensure

Table 3: Performance Comparison of E, S and G Models.

Model	Accuracy	F1 Score	Precision	Recall
SetFit-Model E	0.9950 ± 0.0050	0.9952 ± 0.0048	0.9957 ± 0.0043	0.9945 ± 0.0050
EnvRoBERTa	0.9565 ± 0.0098	0.9319 ± 0.0140	0.9330 ± 0.0399	0.9331 ± 0.0314
SetFit-Model S	0.9600 ± 0.0292	0.9543 ± 0.0341	0.9566 ± 0.0329	0.9600 ± 0.0292
SocRoBERTa	0.9341 ± 0.0140	0.9190 ± 0.0179	0.9035 ± 0.0345	0.9366 ± 0.0292
SetFit-Model G	0.9750 ± 0.0194	0.9738 ± 0.0207	0.9747 ± 0.0205	0.9750 ± 0.0194
GovRoBERTa	0.8961 ± 0.0113	0.7848 ± 0.0262	0.8562 ± 0.0184	0.7252 ± 0.0378

an even number of positive and negative sentence pairs until every sentence pair had been drawn. This methodology ensures that all sentence pairs are included at least once, thereby preventing an imbalance of positive and negative pairs. Given that our polarity data is imbalanced, this approach is beneficial to our model, as oversampling serves to balance the training data and improve the model's performance. To assess the performance of the models, an 80% training set and 20% test set split was employed. The structure of the training and test sets can be found in Table 4 in the Appendix.

Training Results

In evaluating our performance, it is essential to distinguish between two key aspects: entity accuracy and sentiment accuracy. Entity accuracy pertains to the ability of the spaCy model to correctly classify aspect candidate spans as either true entities or non-entities. Sentiment accuracy, on the other hand, concerns the model's capacity to correctly categorize only the filtered aspect candidate spans into their respective classes. With default parameters, the entity prediction accuracy is 91.73%, while the sentiment prediction accuracy is 78.13%. The F1 scores for the entity and sentiment predictions can be found in Table 2. Although our model is more complex than the three distinct E, S, and G models presented in Schimanski et al. (2024), our performance is comparable.

To compare the SetFit model with the models trained by Schimanski et al. (2024), we trained three distinct SetFit models on a considerably smaller training set. We utilized the initial 200 sentences from Schimanski et al. (2024), implementing a 5-fold cross-validation procedure. Our training set comprised 160 sentences, with 40 sentences in our test set. Our model consistently demonstrated superior performance relative to the three models presented in Schimanski et al. (2024). The results are presented in Table 3. For a meaningful comparison with our ABSA model results, the same base model was employed. Despite Schimanski et al. (2024) arguing that the extensive and necessary nature of pretraining is a limitation in the ESG framework, our model, which was not pretrained on our subdomains, still demonstrated superior performance.

Panel VAR estimation

To estimate the pVAR, we employed our sentiment scores and the ESG scores for the main pillars E, S, and G. Additionally, we attempted to estimate more granular pVARs at a subcategory level; however, we only had data from Refinitiv for the years 2018 until 2022. Unfortunately, the available data for this period was insufficient for a reliable pVAR estimation. As a model, we estimated panel VARs with fixed effects and a System Generalized Method of Moments (GMM) approach. The System GMM approach enhances efficiency and addresses potential endogeneity issues.

For each ESG dimension (E, S, G), the dependent variables included the respective ESG score and the net sentiment score. A one-lag approach was employed for the dependent variables, based on the assumption that a company report in a specific year would not have a long-lasting impact on the ESG scores, nor vice versa. To mitigate potential issues associated with serial correlation in the transformed error terms, which could result in less efficient and reliable estimates, and to address potential non-stationarity, we employed forward orthogonal deviations. A two-step estimator was used to obtain robust standard errors. To avoid the proliferation of instruments and the resulting overfitting, we opted to collapse the instruments. The models were estimated separately for each ESG dimension (E, S, G) and each rating agency (Bloomberg, Refinitiv, S&P) using the panelvar package in R (Sigmund and Ferstl 2021).

For each ESG dimension (E, S, G), the pVAR model without exogenous variables and predetermined variables can be represented as follows:

$$\boldsymbol{y}_{i,t} = \boldsymbol{\mu}_i + \sum_{l=1}^{p} \boldsymbol{A}_l \boldsymbol{y}_{i,t-l} + \boldsymbol{\epsilon}_{i,t}$$

Where:

- $\boldsymbol{y}_{i,t}$ is the vector of endogenous variables for company i at time t. In our context, $\boldsymbol{y}_{i,t}$ includes the ESG score and the sentiment score:

$$\boldsymbol{y}_{i,t} = \begin{bmatrix} \text{ESG}_{i,t} \\ \text{Sentiment}_{i,t} \end{bmatrix}$$

- $\boldsymbol{\mu}_i$ represents the individual fixed effects for company i.
- p is the number of lags of the endogenous variables.
- $\boldsymbol{A}_l$ is the matrix of coefficients for the l-th lag of the endogenous variables.
- $\boldsymbol{\epsilon}_{i,t}$ is the vector of idiosyncratic error terms.

Since we used forward orthogonal deviations, the model becomes:

$$\tilde{\boldsymbol{y}}_{i,t} = \boldsymbol{\mu}_i + \sum_{l=1}^{p} \boldsymbol{A}_l \tilde{\boldsymbol{y}}_{i,t-l} + \boldsymbol{\epsilon}_{i,t}$$

The transformed variables, denoted by $\tilde{\boldsymbol{y}}_{i,t}$, represent a deviation from the average of future observations for the same individual. For our specific case with one lag ($p = 1$) and the System GMM approach, the model can be simplified to:

$$\tilde{\boldsymbol{y}}_{i,t} = \boldsymbol{\mu}_i + \boldsymbol{A}_1 \tilde{\boldsymbol{y}}_{i,t-1} + \boldsymbol{\epsilon}_{i,t}$$

In matrix form, considering $\boldsymbol{y}_{i,t} = [\text{ESG}_{i,t}, \text{Sentiment}_{i,t}]^\top$, the model for each ESG dimension (E, S, G) for each rating agency (Bloomberg, Refinitiv, S&P) is:

$$\begin{bmatrix} \tilde{\text{ESG}}_{i,t} \\ \tilde{\text{Sentiment}}_{i,t} \end{bmatrix} = \boldsymbol{\mu}_i + \boldsymbol{A}_1 \begin{bmatrix} \tilde{\text{ESG}}_{i,t-1} \\ \tilde{\text{Sentiment}}_{i,t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{\text{ESG},i,t} \\ \epsilon_{\text{Sentiment},i,t} \end{bmatrix}$$

In contrast to the fixed effects model proposed in Schimanski et al. (2024), this equation is designed to capture the dynamic interactions between ESG scores and sentiment scores over time for each company in the dataset. It should be noted, however, that no exogenous variables were included, such as company fundamentals like the current ratio or revenues. Moreover, the data utilized in this study was limited to that of the EUROSTOXX50, whereas Schimanski et al. (2024) employed data from the EUROSTOXX600.

5 Results

Descriptive Results

The output of our model comprises ESG-related entities and their corresponding sentiment for each report under consideration. This detailed data can be invaluable for an analyst focusing on specific companies. However, while a detailed examination of each company is possible, it would provide too much information for the scope of this work. Therefore, we have summarized the results for all reports and years under consideration simultaneously. Figure 2 illustrates the relative importance and sentiment of all ESG-related entities across all reports, years, and ESG categories. The entities displayed in the word cloud represent a summary of issues from reports spanning from 1999 to 2023. Each word cloud displays the 100 most frequently occurring words. The size of the font indicates the frequency of occurrence of the corresponding word in a given report. The average sentiment score of the overall word cloud is 0.5642. Green shades indicate a positive tone, red shades indicate a negative tone, and shades of yellow represent words discussed with neutral sentiment, lying between the most positively and most negatively discussed terms.

It appears that ”board” and ”employee” are the most pivotal words, with ”employee” being discussed in a more favorable manner on average and ”board” in a more unfavorable manner. While not as crucial, terms such as ”stakeholder”, ”culture”, and ”community” are frequently discussed in a positive light. These terms indicate the ESG topics that companies claim to be improving or are already on a positive trajectory. Some of the less frequently but negatively discussed topics include ”depreciation”, ”volatility”, and ”consolidated financial statement”.

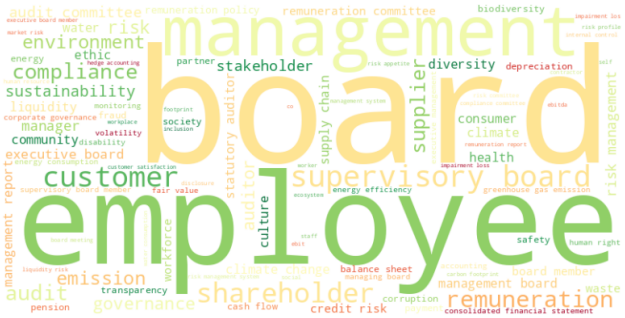


Figure 2: The mean sentiment for the top 100 words in all reports across all ESG pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

Moreover, a word cloud was generated for the top 100 words in each of the three pillars. A comparison of the three pillars reveals that the social pillar is the most positive, with an average sentiment score of 0.7025. The governance pillar

appears to be more negative, with an average sentiment score of 0.47081. In contrast, the environment pillar has a more positive tone, with an average sentiment score of 0.55072, though this is less positive than that of the social pillar. In general, companies tend to report the most negative experiences with governance-related issues. However, the subjective nature of the reporting does not permit the assumption that they experience greater difficulties with governance issues than with social issues. Nevertheless, an analysis of the sentiment scores on their own reveals the key areas of concern for companies in the ESG context.

Upon examining the environment pillar in Figure 3, it becomes evident that the term ”sustainability” is a significant topic with a positive tone. The term ”environment” is also frequently mentioned, although with a less positive tone. The term ”emission” appears to be a source of greater concern, potentially leading to difficulties for the companies in question. While not frequently discussed, other environmental issues, such as weather conditions, weather events, oil prices, gas prices, and droughts, also appear to be significant concerns for companies. In contrast, topics such as recycling, energy efficiency, and ecosystems are associated with a more positive tone.

In conclusion, weather-related events and fluctuations in prices present significant challenges for companies within the EUROSTOXX50. In contrast, sustainability issues such as recycling, the carbon footprint, and energy efficiency appear to be less pressing concerns.

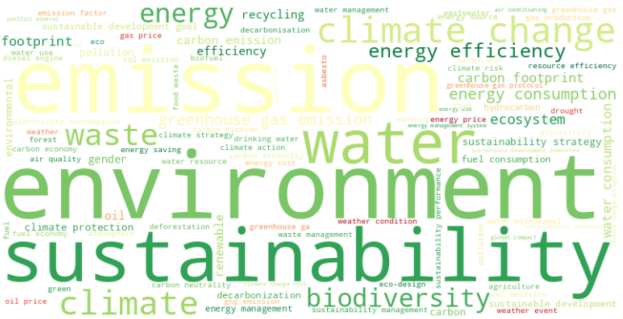


Figure 3: The mean sentiment for the top 100 words in all reports for the environment pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

As previously stated, the social setting is characterized by a positive tone. The color green is frequently associated with words in Figure 4. This is particularly evident in the instances of the most discussed words ”employee” and ”customer”. Conversely, the words ”community”, ”culture”, and ”diversity” appear to be the most positively mentioned in the social setting. The social pillar presents a challenge for companies in the areas of pension, employee benefits, and fair value. In general, companies tend to report more on the positive aspects and achievements related to social issues in the ESG context.

Figure 4: The mean sentiment for the top 100 words in all reports for the social pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

The wordcloud representing the governance pillar, as shown in Figure 5, is predominantly composed of words in shades of yellow and orange, which collectively convey a more negative tone. The most frequently discussed topics in the context of governance are ”management”, ”shareholder”, ”board”, and ”supervisory board”. These topics are associated with a more negative tone. It appears that the board and remuneration settings are a more negative issue for companies in terms of governance. The annotated words that are more positive in tone are ”stakeholder”, ”transparency”, and ”compliance”. These words are more closely related to the concept of business transparency.

While an analysis of individual ESG-related entities can provide detailed insights into companies’ reporting on ESG topics, displaying changes in tone over time for each entity is not feasible. Consequently, to analyze ESG issues in general, including a time component, the entities were grouped into ESG subcategories as previously mentioned in Section 4. Figure 6 depicts the average sentiment scores per year

Figure 5: The mean sentiment for the top 100 words in all reports for the governance pillar.

The words are shaded from red, representing a more negative sentiment, to green, representing a more positive sentiment. The size of the word indicates the extent of its discussion.

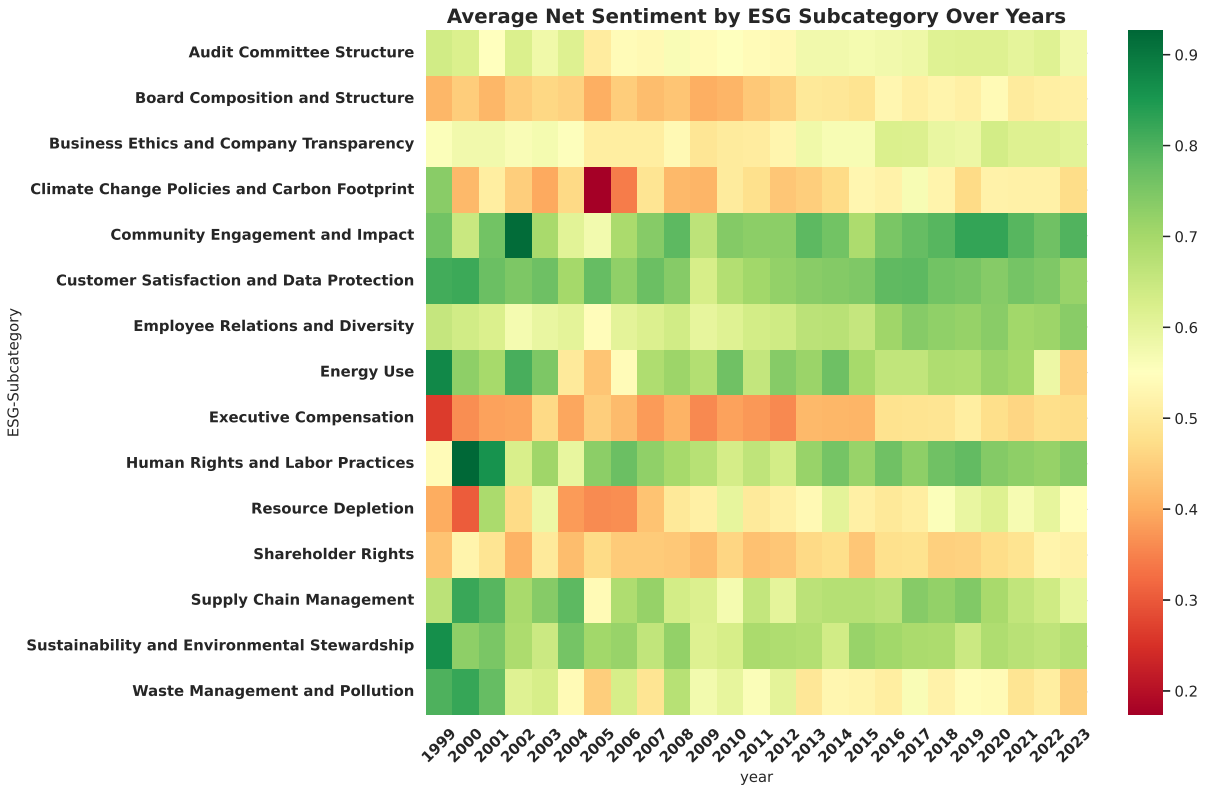


Figure 6: Average Sentiment for each ESG-Subcategory over time.
The legend ranges from a more negative sentiment in blur to a more positive senitment in red.

per ESG subcategory for all EUROSTOXX50 companies. A darker shade of red indicates a more negative tone, while a darker shade of green indicates a more positive tone. It is important to note that all sentiment scores over the years are positive. Consequently, our findings indicate a tendency towards more positive reporting on ESG-related issues in general, with individual reports exhibiting a negative sentiment score. However, when considering the overall tone across all companies and years, a positive sentiment is evident. As our model is trained using an oversampling method to prevent class imbalance, we still observe a greater tendency towards positivity in our training sample. This is likely because reports are, on average, more positive in nature. This is a logical conclusion, as companies appear to prioritize reporting on more positive achievements and topics, which could be perceived as greenwashing (Bingler et al. 2022). Nevertheless, to substantiate these assertions, further analysis is required, which will be briefly discussed in the Conclusion section.

Prior to 2005, there appeared to be a greater range of tones in reporting, with particularly positive coverage of ”Energy Use” in 1999, ”Human Rights” and ”Labor Practices” in 2000, and ”Community Engagement and Impact” in 2002. In contrast, more negative reporting was observed for ”Executive Compensation” in 1999 and ”Climate Change Policies and Carbon Footprint” in 2005. However, given the paucity of reports in the early years and the fact that special ESG reports were not introduced until 2004, the results from this period may be subject to bias.

Upon examination of the data from 2005 onwards, it becomes evident that certain subsections are more positively discussed than others. In general, the subcategories ”Community Engagement and Impact”, ”Customer Satisfaction and Data Protection”, ”Employee Relations and Diversity”, and ”Sustainability and Environmental Stewardship” appear to have been discussed in the most positive manner over time. The sentiment expressed towards ”Energy Use” is consistently positive, with the exception of the years 2004 to 2006 and 2023. On the less favorable end of the spectrum are the categories ”Board Composition and Structure”, ”Climate Change Policies and Carbon Footprint”, ”Executive Compensation”, and ”Shareholder Rights”. The remaining categories exhibit a more diverse range of results.

The scores generated for all companies in a given year can also be created for a specific company or a specific report. Consequently, this tool allows for the generation of alternative ESG scores. While such scores are inherently subjective and reflect the perspectives of the companies themselves, they can be used as an adjunct to traditional scores, which are subject to similar limitations. This aligns with the findings of Berg, Koelbel, and Rigobon (2022), who posited that

aggregate scores may be unreliable. Additionally, the correlation coefficients between the Bloomberg, Refinitiv, and S&P's scores are relatively low, illustrating the variability between the scores of the rating agencies and the questionable transparency of their methodologies.

Despite the subjectivity of our scores, they can provide valuable insights into a company's strengths and weaknesses on ESG topics. Furthermore, changes in score values can be interpreted as improvements or declines in certain ESG areas within a company.

Panel VAR model results

To interpret the results of our pVAR models, we employ generalized orthogonal response functions. The results can be found in Figure 8 in the appendix. In all models, except for the model for governance with S & P data, where we find a positive effect in the first period for both sentiment on score and score on sentiment, there are no significant reactions from the ESG scores to a shock in sentiment, nor vice versa. This may be attributed to model misspecification due to a lack of exogenous variables. Alternatively, this could be viewed as another point of criticism regarding ESG scores from rating agencies. In future research, it would be beneficial to include company fundamentals to improve the predictive value of the pVAR models, though this would also increase model complexity.

Due to the criticism outlined in the literature (Berg, Koelbel, and Rigobon 2022), we also calculated the correlations between the ESG scores from all agencies. Figure 7 illustrates the results, demonstrating a lack of correlation between the E scores, S scores, and G scores across the agencies. This aligns with the general criticism of these scores and shows significant variation between ESG scores, raising questions about their reliability. Consequently, the lack of significance in the relationship between the scores and our sentiment scores does not necessarily imply a deficiency in the quality of our sentiment scores. It is possible that the rating scores do not adequately capture the subjective views of the companies regarding ESG issues.

Consequently, our scores may be employed as an alternative measure, albeit subjective, in conjunction with the ESG scores provided by rating agencies. The aggregation of our scores into ESG subcategories provides a more detailed view of ESG issues. Alternatively, companies can be analyzed on an entity-by-entity basis, without the necessity of manually reading their reports. Furthermore, stakeholders may utilize reports from a single company and employ the provided tool to identify the issues that are most negatively discussed in the company report. This information can then be subjected to further investigation in greater detail. To provide a more objective measure, the model could be trained and applied to news data (Fischbach et al. 2023). This approach would diversify the sources of ESG information, potentially reducing bias and offering a broader perspective on ESG performance.

6 Conclusion

In this study, we developed an ABSA model that serves as a systematic tool for extracting entities related to ESG issues and their associated sentiment from company reports. Unlike previous studies that have employed NLP on company reports (Schimanski et al. 2024), our model provides more detailed insights into a company's perspective on ESG issues. While Schimanski et al. (2024) adopted a quantitative approach by examining the frequency of mentions of E, S, and G in company reports, our approach is more qualitative, incorporating the tone of reporting. This method provides deeper insights into a company's performance in relation to ESG matters beyond mere quantity. Furthermore, we employed a distinct time series analysis methodology. By estimating a pVAR, we examined the bidirectional effects of the variables, considering that ESG scores could influence the tone of reporting and vice versa. Despite this comprehensive approach, no significant relationship, except for one model, was found between the ESG scores provided by rating agencies and the sentiment scores. This result may be attributed to the controversy surrounding the quality of ESG scores, as discussed in Berg, Koelbel, and Rigobon (2022). Our correlation analysis between the ESG scores revealed a lack of consistency among the rating agencies, indicating inconsistencies in their methodologies. Consequently, there appears to be no predictive capacity of ESG sentiment in company reports on ESG scores from rating agencies.

The ABSA model may prove to be a valuable resource for stakeholders seeking insights into a company's stance on ESG issues. The tool can provide a comprehensive understanding of ESG topics and their associated tone at the entity level. This can be achieved by examining detailed topics or aggregating data at a higher level, focusing on specific ESG subcategories or even on the E, S, and G pillar levels. The tool can be applied to specific reports or all reports over time from a company to identify which ESG matters might be problematic or unproblematic at specific points in time. As company reports are inherently subjective, future research could apply our tool to news data, as suggested by (Fischbach et al. 2023), to obtain a more objective measure. Given the ongoing debate regarding the quality of ESG scores from rating agencies (Schimanski et al. 2024; Berg, Koelbel, and Rigobon 2022), our tool offers an alternative measure that aligns with the companies' own views on ESG matters. Furthermore, our findings indicate that the SetFit few-shot model with standard parameters yields superior outcomes on the same dataset, despite using only one-tenth of the training data.

Nevertheless, our analysis is subject to several limitations. In our time series analysis, we did not incorporate any exogenous variables, such as company fundamentals, which

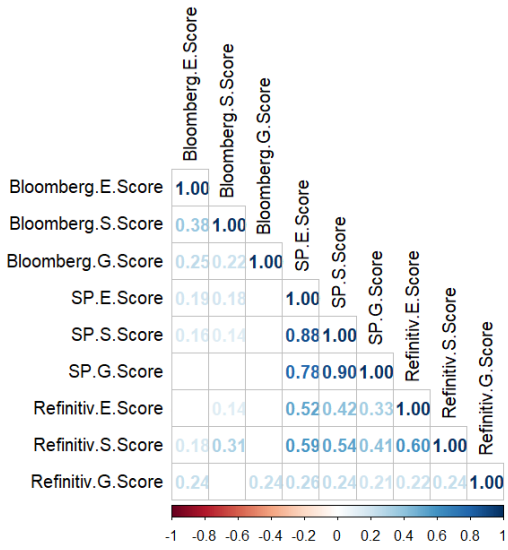


Figure 7: Correlation Analysis between Ratings from Refinitiv, S&P, and Bloomberg.
The figure displays the Pearson correlation coefficients between the ratings obtained from Refinitiv, S&P, and Bloomberg. Significant correlations are shown with their respective values, while non-significant correlations are blanked out.

could have enhanced the analysis. The absence of a correlation between the scores may be attributed to this limitation, suggesting an opportunity for future research to include such variables. Additionally, the available data did not permit a time series analysis on an ESG subcategory basis. With more than five years of data from Refinitiv matched to our ESG subcategories, future research could explore relationships at a more granular level. Regarding the training process of our model, there is considerable scope for improvement. The ABSA model currently employs standard parameters. While a grid search could potentially improve the model, it requires significant computational resources, and the sustainability of such extensive training must be considered. Furthermore, comparing different base models, especially larger ones, may yield more favorable outcomes. Future research could also incorporate additional variables beyond those provided by ESG scores from rating agencies to assess the quality of our sentiment scores. A larger sample size, such as that from the EUROSTOXX600 or S&P 500, could facilitate the development of more robust time series models and provide insights into industry-specific ESG issues.

Acknowledgements

Christoph Funk acknowledges the funding by the German Academic Exchange Service (DAAD) from funds of Federal Ministry for Economic Cooperation (BMZ), SDGnexus Network (Grant No. 57526248), program “exceed - Hochschulexzellenz in der Entwicklungszusammenarbeit.”

Christian Haas acknowledges funding from the project ”safe Financial Big Data Cluster” (safeFBDC) by the Federal Ministry for Economic Affairs and Climate Action.

References

Auzepy, A.; Tönjes, E.; Lenz, D.; and Funk, C. 2023. Evaluating TCFD reporting—A new application of zero-shot analysis to climate-related financial disclosures. *PLOS ONE*, 18(11): e0288052.

Berg, F.; Koelbel, J. F.; and Rigobon, R. 2022. Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6): 1315–1344.

Bingler, J. A.; Kraus, M.; Leippold, M.; and Webersinke, N. 2022. Cheap talk and cherry-picking: What ClimateBert has to say on corporate climate risk disclosures. *Finance Research Letters*, 47: 102776.

Fischbach, J.; Adam, M.; Dzhagatspanyan, V.; Mendez, D.; Frattini, J.; Kosenkov, O.; and Elahidoost, P. 2023. Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool. In *2023 IEEE International Conference on Big Data (BigData)*, 2823–2830.

Friederich, D.; Kaack, L. H.; Luccioni, A.; and Steffen, B. 2021. Automated Identification of Climate Risk Disclosures in Annual Corporate Reports. arXiv:2108.01415.

Föhr, T. L.; Schreyer, M.; Juppe, T. A.; and Marten, K.-U. 2023. Assuring Sustainable Futures: Auditing Sustainability Reports using AI Foundation Models. Available at SSRN: https://ssrn.com/abstract=4502549 or http://dx.doi.org/10.2139/ssrn.4502549.

Goloshchapova, I.; Poon, S.-H.; Pritchard, M.; and Reed, P. 2019. Corporate social responsibility reports: topic analysis

and big data approach. *The European Journal of Finance*, 25(17): 1637–1654.

Jain, Y.; Gupta, S.; Yalciner, S.; Joglekar, Y. N.; Khetan, P.; and Zhang, Q. 2023. Overcoming Complexity in ESG Investing: The Role of Generative AI Integration in Identifying Contextual ESG Factors. *SSRN Electronic Journal*. Available at SSRN: https://ssrn.com/abstract=4495647 or http://dx.doi.org/10.2139/ssrn.4495647.

Jo, Y.; and Oh, A. H. 2011. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 815–824. New York, NY, USA: Association for Computing Machinery.

Koch, G.; Zemel, R.; and Salakhutdinov, R. 2015. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2. Lille.

Laperdon, R.; Aarsen, T.; Tunstall, L.; Korat, D.; Pereg, O.; and Wasserblat, M. 2023. SetFitABSA: Few-Shot Aspect Based Sentiment Analysis using SetFit. https://huggingface.co/blog/setfit-absa. Accessed: 2024-06-07.

Lee, H.; Lee, S. H.; Lee, K. R.; and Kim, J. H. 2023. Esg discourse analysis through bertopic: comparing news articles and academic papers. *Computers, Materials & Continua*, 75(3): 6023–6037.

Liu, B. 2020. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; and Raffel, C. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. arXiv:2205.05638.

Luccioni, A.; Baylor, E.; and Duchene, N. 2020. Analyzing Sustainability Reports Using Natural Language Processing. arXiv:2011.08073.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; et al. 2016. SemEval-2016 Task 5: Aspect Based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 19–30. San Diego, California: Association for Computational Linguistics.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Manandhar, S.; and Androutsopoulos, I. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 486–495. Denver, Colorado: Association for Computational Linguistics.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Saeidi, M.; Bouchard, G.; Liakata, M.; and Riedel, S. 2016. SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. *CoRR*, abs/1610.03771.

Schimanski, T.; Reding, A.; Reding, N.; Bingler, J.; Kraus, M.; and Leippold, M. 2024. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61: 104979.

Shen, Z.; Lo, K.; Wang, L. L.; Kuehl, B.; Weld, D. S.; and Downey, D. 2021. VILA: Improving Structured Content Extraction from Scientific PDFs Using Visual Layout Groups. Last accessed Feb. 28, 2022.

Sigmund, M.; and Ferstl, R. 2021. Panel vector autoregression in R with the package panelvar. *The Quarterly Review of Economics and Finance*, 80: 693–720.

TCFD. 2017. Final Report: Recommendations of the Task Force on Climate-related Financial Disclosures. Technical Report 11 (1), TCFD. (TCFD, 2017).

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Ferrer, C. C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.-A.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288.

Tunstall, L.; Reimers, N.; Jo, U. E. S.; Bates, L.; Korat, D.; Wasserblat, M.; and Pereg, O. 2022. Efficient Few-Shot Learning Without Prompts. arXiv:2209.11055.

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

Xu, Y.; Li, M.; Cui, L.; Huang, S.; Wei, F.; and Zhou, M. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1192–1200.

Zhang, L.; Wang, S.; and Liu, B. 2018. Deep Learning for Sentiment Analysis: A Survey. *CoRR*, abs/1801.07883.

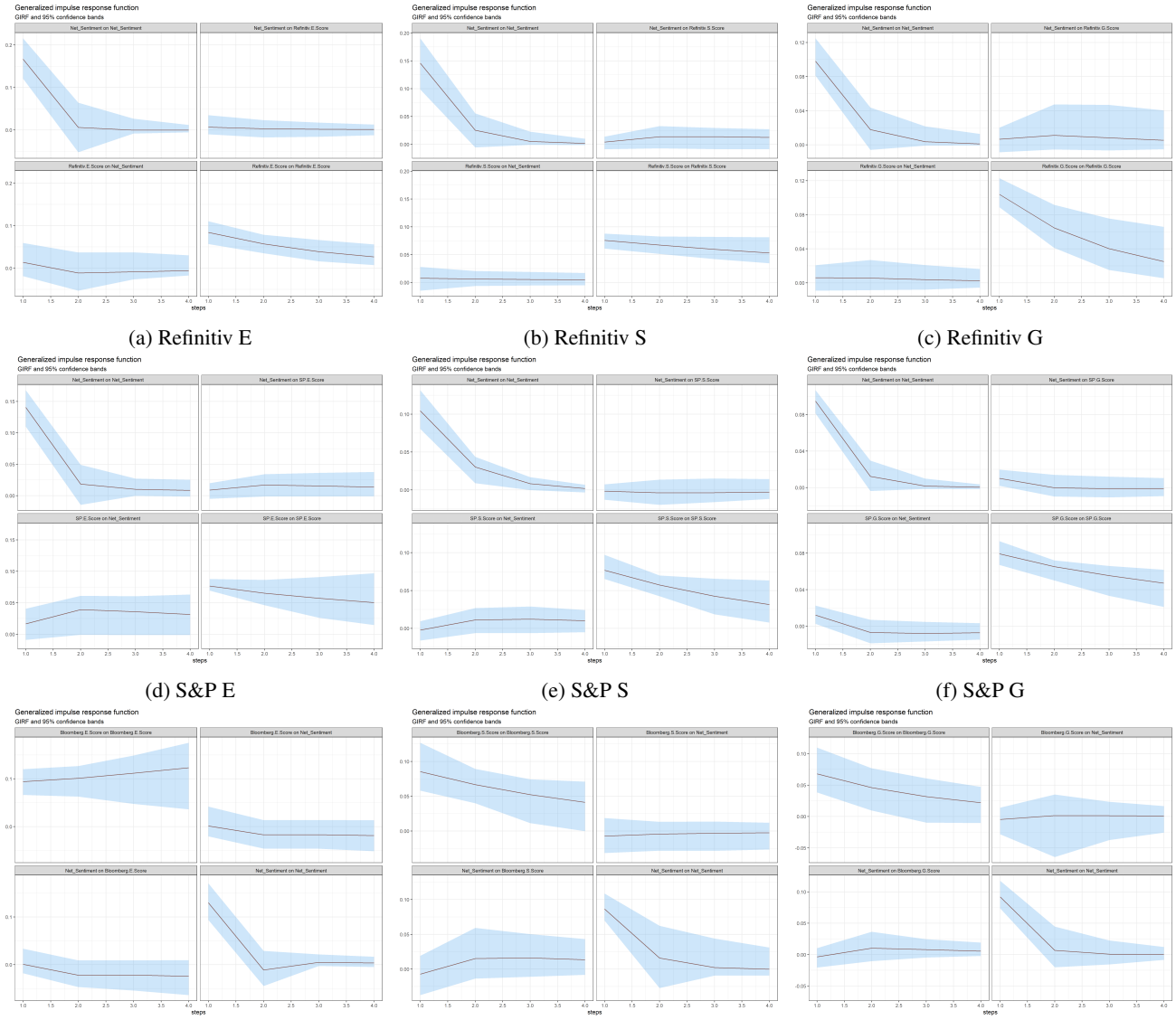
Appendix

Table 4: Distribution of Sentiment Labels in Datasets

Dataset	**Set**	**Positive**	**Neutral**	**Negative**	**Total**
Environment	Training	86	30	21	137
	Test	14	9	7	30
Sum		100	39	28	167
Social	Training	122	6	3	131
	Test	27	1	0	28
Sum		149	7	3	159
Governance	Training	119	23	4	146
	Test	33	3	2	38
Sum		152	26	6	184

Table 5: ESG Subcategories and Definitions

ESG Category	Subcategory Definition	Example Entities
Environmental	Climate Change Policies and Carbon Footprint: Measures the company's contribution to climate change through greenhouse gas emissions and carbon footprint management.	Greenhouse Gas Emission, Carbon Emission, Decarbonization, Greenhouse Gas, Climate Risk
	Energy Use: Assesses the company's energy efficiency and renewable energy usage.	Energy Efficiency, Energy, Renewable, Energy Sector, Fuel Efficiency
	Waste Management and Pollution: Evaluates waste management practices, pollution prevention, and handling of toxic emissions.	Recycling, Carbon Dioxide, Waste Management, Food Waste, Air Pollution
	Resource Depletion: Considers the company's use of resources, such as water and raw materials, and its impact on biodiversity.	Fuel Economy, Natural Resource, Forest, Resource Management, Coal
	Sustainability and Environmental Stewardship: Looks at the company's overall commitment to environmental sustainability practices.	Environment, Sustainability, Climate, Environmental, Sustainable Development
Social	Employee Relations and Diversity: Involves employee treatment, diversity, labor standards, and fair wages.	Employee, Diversity, Health, Woman, Culture
	Customer Satisfaction and Data Protection: Focuses on product quality, customer service, data security, and privacy.	Customer, Fair Value, Customer Satisfaction, Cybersecurity
	Community Engagement and Impact: Looks at how the company contributes to the communities in which it operates, including charitable efforts and community service.	Community, Society, Social, Corporate Social Responsibility, Citizen
	Human Rights and Labor Practices: Assesses the company's adherence to fair labor practices, human rights, and avoiding exploitation.	Human Right, Discrimination, Harassment, Refugee, Child Labor
	Supply Chain Management: Evaluates the social aspects of the supply chain, including labor practices and human rights of suppliers.	Global Supply Chain, Supply Chain, Supplier, Contractor
Governance	Board Composition and Structure: Analyzes the diversity, independence, and expertise of board members.	Board, Management, Executive Board, Supervisory, Top Management
	Executive Compensation: Looks at how executives are compensated and whether it aligns with the company's long-term goals and shareholders' interests.	Renumeration, Management Renumeration, Supervisory Board Renumeration, Board Renumeration, Cash Renumeration
	Audit Committee Structure: Evaluates the quality and independence of internal audits and controls.	Audit, Compliance, Auditor, Tax, Accounting
	Business Ethics and Company Transparency: Considers ethical business practices, transparency in reporting, and avoiding conflicts of interest.	Risk Management, Fraud, Crisis Management, Business Ethics, Credit Risk Management
	Shareholder Rights: Examines the rights of shareholders and how well the company listens to and integrates their feedback.	Shareholder, Minority Shareholder, Ordinary Shareholder, Minority Interest, Shareholder Renumeration



(a) Refinitiv E (b) Refinitiv S (c) Refinitiv G

(d) S&P E (e) S&P S (f) S&P G

(g) Bloomberg E (h) Bloomberg S (i) Bloomberg G

Figure 8: Generalized Impulse Response Function (GIRF) of pVAR Model with 95% confidence bands.
The figure displays the Generalized Impulse Response Function (GIRF) of the pVAR model with 95% confidence bands for ratings obtained from Refinitiv, S&P, and Bloomberg. The confidence bands are based on 1000 bootstrap samples.

Table 6: Summary of ESG scores from S&P.

Company Name	From Year	To Year	Count of Years
adidas	2014	2023	10
adyen	2019	2023	5
airbus	2014	2023	10
allianz	2014	2023	9
anheuserbusch	2014	2023	10
asml	2014	2023	9
axa	2014	2023	9
banco bilbao	2014	2023	10
banco santander	2014	2023	9
basf	2014	2023	10
bayer	2014	2023	10
bmw	2014	2023	10
bnp paribas	2014	2023	9
crh	2014	2023	9
danone	2014	2023	10
deutsche boerse	2014	2022	9
deutsche post	2014	2023	8
deutsche telekom	2014	2023	8
enel	2014	2023	9
eni	2014	2023	9
essilorluxottica	2014	2023	10
flutter	2016	2023	6
hermes	2014	2023	10
iberdrola	2014	2023	9
industria de	2014	2023	8
infineon	2014	2022	9
ing	2014	2023	8
intesa sanpaolo	2014	2023	9
kering	2014	2023	9
ahold delhaize	2014	2023	10
air liquide	2014	2023	10
loreal	2014	2023	10
lvmh	2014	2023	9
mercedes	2014	2023	9
munich re	2014	2023	9
nokia	2014	2023	10
nordea	2014	2023	9
pernod ricard	2014	2022	8
prosus	2020	2022	2
safran	2014	2023	10
sanofi	2014	2023	9
sap	2014	2023	8
schneider electric	2014	2023	8
siemens	2014	2023	9
stellantis	2015	2023	9
totalenergies	2014	2023	8
unicredit	2014	2023	9
vinci	2014	2023	9
volkswagen	2014	2023	9
vonovia	2015	2023	9

Table 7: Summary of ESG ratings from Bloomberg.

Company Name	From Year	To Year	Count of Years
adidas	2015	2022	8
adyen	2018	2022	5
airbus	2015	2022	8
allianz	2015	2022	8
anheuserbusch	2015	2022	8
asml	2015	2022	8
axa	2015	2022	8
banco bilbao	2015	2022	8
banco santander	2015	2022	8
basf	2015	2022	8
bayer	2015	2022	8
bmw	2015	2022	8
bnp paribas	2015	2022	8
crh	2015	2022	8
danone	2015	2022	8
deutsche boerse	2015	2022	8
deutsche post	2015	2022	8
deutsche telekom	2015	2022	8
enel	2015	2022	8
eni	2015	2022	8
essilorluxottica	2015	2022	8
flutter	2021	2022	2
hermes	2015	2022	8
iberdrola	2015	2022	8
industria de	2015	2022	8
infineon	2015	2022	8
ing	2015	2022	8
intesa sanpaolo	2015	2022	8
kering	2015	2022	8
ahold delhaize	2016	2022	7
air liquide	2015	2021	7
loreal	2015	2022	8
lvmh	2015	2022	8
mercedes	2015	2022	8
munich re	2015	2022	8
nokia	2015	2022	8
nordea	2015	2022	8
pernod ricard	2015	2022	8
prosus	2020	2022	3
safran	2015	2022	8
sanofi	2015	2022	8
sap	2015	2022	8
schneider electric	2015	2022	8
siemens	2015	2022	8
stellantis	2015	2022	8
totalenergies	2015	2022	8
unicredit	2015	2022	8
vinci	2015	2022	8
volkswagen	2015	2022	8
vonovia	2015	2022	8

Table 8: Summary of ESG ratings from Refinitiv.

Company Name	From Year	To Year	Count of Years
adidas	2002	2022	21
adyen	2018	2022	5
airbus	2002	2022	21
allianz	2002	2023	22
anheuserbusch	2002	2022	21
asml	2002	2022	21
axa	2002	2022	21
banco bilbao	2002	2022	21
banco santander	2002	2022	21
basf	2002	2022	21
bayer	2002	2022	21
bmw	2005	2022	18
bnp paribas	2002	2022	21
crh	2005	2022	18
danone	2005	2022	18
deutsche boerse	2002	2022	21
deutsche post	2005	2022	18
deutsche telekom	2002	2022	21
enel	2002	2022	21
eni	2002	2022	21
essilorluxottica	2002	2022	21
flutter	2005	2022	18
hermes	2005	2022	18
iberdrola	2002	2022	21
industria de	2002	2023	22
infineon	2002	2023	22
ing	2002	2022	21
intesa sanpaolo	2002	2022	21
kering	2002	2022	21
ahold delhaize	2002	2023	21
air liquide	2005	2022	18
loreal	2002	2022	21
lvmh	2002	2022	21
mercedes	2002	2022	21
munich re	2002	2022	21
nokia	2002	2022	21
nordea	2005	2023	19
pernod ricard	2002	2022	21
prosus	2020	2023	4
safran	2002	2022	21
sanofi	2002	2022	21
sap	2002	2022	21
schneider electric	2002	2022	21
siemens	2002	2023	22
stellantis	2002	2022	21
totalenergies	2002	2022	21
unicredit	2007	2022	16
vinci	2002	2022	21
volkswagen	2002	2022	21
vonovia	2015	2022	8