SECBERT, a BERT Model for U.S. Securities and Exchange Commission (SEC) Document Analysis

Feasibility Study and Development Roadmap

Emanuele Ferrari[1]

[1]Department of Data Science, GISMA Univeristy of applied Science, Konrad-Zuse-Ring 11,
Potsdam, 14469, Brandenburg, Germany

30, March, 2024

Abstract

This paper will explore the feasibility of deploying a BERT model for U.S. Securities and Exchange Commission (SEC) documents classification. Companies publicly traded in the U.S. must submit financial statement to the SEC providing comprehensive insights into corporate financial health, governance, and strategic direction, we will explore an existing model performance on a multi label classification task, and lay the foundation to deploy a NLP tool for financial text analysis. The a state-of-the-art BERT model, will be able to label specific sections of SEC filings, reducing analyst research time. Due to time and budget limitations, this study focus only on 10-K (yearly) reports, aiming to label Environmental, Social, and Governance sections as benchmark for our proof of concept. The paper demonstrates the potential of a fine tuned model to significantly enhance the efficiency of financial text analysis, offering a promising direction for a larger scope deployment.

Source: All codes used for this paper can be found at the relative GitHub page.

1. Introduction

The main challenge for financial text analysis is the sacristy of domain-specific lexicon, corpora, and labeled data. Pre-trained language models, such as BERT, offer a viable solution to this issue; after gathering a comprehensive language understanding on a huge English-based dataset (Devlin et al. (2018)), it necessities fewer labeled examples for effective fine-tuning; these models can be refined on domain-specific corpora, significantly increasing their performance in specialized tasks (Sun et al. (2019)). The choice of a BERT-based model for our research was driven by the exceptional results in classification tasks, and the demonstrated understanding of context and subtle meanings in text (Devlin et al. (2018)).

The SEC's EDGAR database is an invaluable resource for financial analysis, offering access to all regulatory filings that all U.S. publicly traded companies must submit to the authorities. Among these, 10-K reports are the most inspected by analysts, as they provide the annual overview of a company's financial performance, risk factors, and management's discussion of the fiscal year. The information contained in 10-K reports makes them the most informative texts for financial analysis (LI (2010)). Therefore, we focus only on these reports for a small-scale application as a proof of concept for the SEC-BERT model.

In modern financial analysis, Environmental, Social, and Governance (ESG) factors play a crucial role, as they increasingly influence investment decisions and risk assessments (Friede et al. (2015)). However, the lack of standardization in ESG reporting results in mixed and fragmented disclosures that complicate the analysis (Amel-Zadeh & Serafeim (2017)), challenging automated text processing and analysts' ability to quickly scan the documents. Additionally, the absence of dedicated sections for ESG information within 10-K filings further complicates the extraction of relevant data. This lack of uniformity and specificity poses significant challenges for analysts asserting a company's ESG performance and integrate it into their work (Sullivan & Mackenzie (2017)). These obstacles depict an excellent testing ground for LLM; therefore, we chose ESG sections' labeling as a benchmark for our project.

Due to hardware, time, and budget constraints, this paper only compares and analyzes the performance of different pre-trained BERT models in ESG multi-label classification fed with the same fine-tuning dataset. The models under investigation are:

- BERT Base English (Devlin et al. (2018));
- Prosus FinBERT (Araci (2019)); and
- Yiyanghkust/FinBERT-ESG (Huang et al. (2023)).

2. Research Objectives

The final objective of this research is to examine the feasibility of and guide the development of an LLM tailored to SEC documents. The objective can be broken down into:

- **Overview of existing literature**
- **Design a schema for SECBERT development**
- **Evaluate existing model fine-tuned on a similar task on paragraph multi-labelling classification**
- **Evaluate results of different pre-trained models on our task**
- **Describe insights from experimenting and testing**
- **Guide future research and development**

3. Related Works

Given the widespread adoption of Google's BERT model and its extensive application across various domains and tasks (Rogers et al. (2020)), we have chosen not to delve into the model's technical details in this paper. For readers interested in BERT's underlying mechanisms, we refer to the original work by Devlin et al. (2018) and Vaswani et al. (2017).

3.1 BERT Base English

The BERT Base English model was introduced by Devlin et al. (2018), was pre-trained on a vast and diverse corpus. The original scope of BERT was to create a model that could be fine-tuned with just one additional output layer to excel across a broad range of NLP tasks without requiring task-specific architectural modifications.

3.2 ProsusAI/FinBERT

FinBERT, developed by ProsusAI, is a language model specifically pre-trained for the financial domain, based on BERT architecture. Introduced in Araci (2019) publication, the final FinBERT is pre-trained on the TRC2-financial dataset, a large corpus consisting with more than 29M words and almost 400K sentences with a maximum sequence length of 64 tokens.

The model is fine-tuned for sentiment analysis using the FiQA Sentiment database (Maia et al. (2018)), which presents sentences labeled by various experts as they will have a positive, negative, or neutral impact on the company stock price. Interestingly, the dataset report also shows the agreement levels between experts, further underlying the human difficulties in financial text classification and explaining part of the model errors.

In the paper, the authors compared the results of the BERT "Vanilla" version with the pre-trained (FinBERT-domain) and a smaller Financial PhraseBank corpus (FinBERT-task) consisting of 4845 English sentences selected randomly from LexisNexis database's Financial news (Malo et al. (2013)). The results showed a similar marginal improvement (+0.01 F1 score) compared to the vanilla model. The author assumed that it is due to the already high performance of the BERT base model, with an accuracy of 0.96 on sentences where all authors agree, therefore leaving little margin for improvement for the fine-tuned models.

3.3 Yiyanghkust/FinBERT-ESG

Yiyanghkust/FinBERT, created by Huang, Wang, and Yang (the two models have the same name but different authors), is pre-trained on a vast sample of financial documents, consisting of corporate filings (10-Ks and 10-Qs), financial analyst reports, and earnings conference call transcripts, to a total of 4.9 billion tokens, and a maximum token length of 512, perfectly tailored to our scope. FinBERT-ESG is fine-tuned for multi-class classification with four mutually exclusive classes: "None," "Environmental," "Social," and "Governance," using the FiQA Sentiment dataset ('FiQA 2018 Task: Aspect-Based Financial Sentiment Analysis' (2018)).

3.3.1 Milti Class and Multi Label Classification

Yiyanghkust/FinBERT-ESG originally approached ESG content categorization as a multi-class problem where sentences and text blocks were exclusively assigned to one of the four classes (None, E, S, G). However, during the manual labeling for the training set, we realized that many text sections could be categorized with more than one label. For example, the following section:

> *"- regulations relating to worker safety and environmental protection;"*

It can be categorized as both Environmental and Social. Therefore, we opted for a multi-label classification, where each class can be assigned individually.

The main differences reside in the output layer. Multi-class classification requires a "non-ESG" class. A softmax function is applied as it allocates the probability across various classes, forcing their sum to one, and the one with the highest probability is chosen as the predicted class, making it ideal for instances designated by a singular label (Devlin et al., 2019). The softmax funtion is:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

where z represents the input vector, and K classes.

In contrast, multi-label classification does not require a "neutral" label as it relies on the Sigmoid function, which assesses each class independently to generate distinct probability scores for each label. Each probability is capped between zero and one, and each class over a certain threshold (usually 0.5) is labeled as positive for the relative class. The Sigmoid funtion is:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

where z is the input vector.

3.4 ESG and European Sustainability Reporting Standards (ESRS)

ESG reporting is marked by an absence of a universally accepted framework and definition, leading to significant fragmentation of ESG disclosures. This complicates efforts to compare and label ESG data across different documents (Eccles & Klimenko (2019)). We decided to follow the definitions crafted by the European Financial Reporting Advisory Group (EFRAG) for the European Sustainability Reporting Standards (ESRS) (European Financial Reporting Advisory Group (2023)) for our labeling criteria. Although still in proposal form, the ESRS aims to standardize ESG reporting, providing a comprehensive set of definitions for consistent disclosure. Despite its draft stage, we opted to embrace the ESRS due to the European Union's record of setting global regulatory standards. Historically, EU regulations have often preceded broader international adoption, and we believe the ESRS could follow this trend (Timothy Busch & Orlitzky (2016)). The ESRS frame divides ESG matters into:

Environmental

- E 1 - Climate Change
- E 2 - Pollution
- E 3 - Water and marine resources
- E 4 - Biodiversity & ecosystems
- E 5 - Resource use & circular economy

Social

- S 1 - Own workforce
- S 2 - Workers in the value chain
- S 3 - Affected communities
- S 4 - Consumers & end-users

Governance

- G 1 - Business conduct

Source: European Financial Reporting Advisory Group (2023), sub-classes can be found in the related slides.

4. Project Schema

The project can be divided into "Development and Training" and "Deployment". The following image represents the whole project.

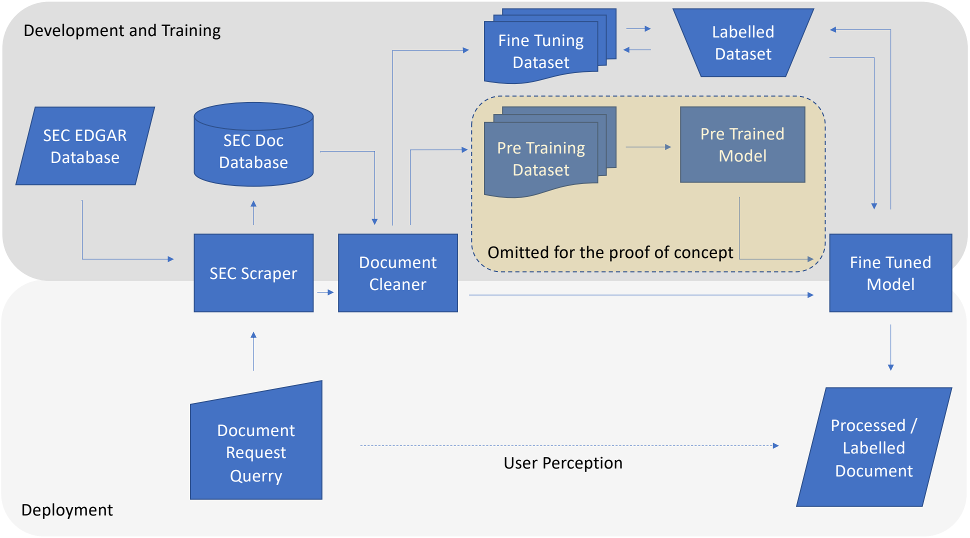

Figure 1: Visual representation of SECBERT.

4.1 Development and Training

In this preliminary evaluation phase, we focused our limited resources on developing and fine-tuning SECBERT, omitting a time-consuming pre-training, opting for an existing pre-trained model, and comparing them.

4.1.1 SEC Scraper and Document Cleaner

All documents can be retrieved via SEC EDGAR (Electronic Data Gathering, Analysis, and Retrieval system) through provided APIs. However, these APIs are limited, and we need to code an "SEC Scraper" that merges them with HTML scraping; the consistent layout of the centralized database allows us to retrieve multiple documents and document types. SEC provides a list of CIK (Central Index Key), a unique number assigned to a company for every possible Ticker Symbol, a unique character-based reference of traded stocks. A single company can have multiple tickers but always refer to a unique CIK as an identifier. For example, Alphabet CIK is 1652044 and has two traded stocks: GOOG and GOOGL.

Given a CIK, our scraper can retrieve the main Ticker and all fillings, identify the type, year, and unique code, and output documents as HTML. For this proof of concept, we focused only on S&P500 companies' 10-K filings and stored all documents as a JSON file. Future expansion will integrate other types of documents, such as 10-Q (quarterly reports) and 8-K (proxy statements) while using an SQL-based database to integrate companies and document objects better.

These documents are then processed by the "Document Cleaner", a script that pre-processes HTML files and outputs a clean Pandas dataframe with all separate text blocks, their type (narrative, list, title, table, image, or item heading), and other parameters, while removing recurrent heading/footers and split text blocks over 448 tokens (512 minus a safe margin).

This project element was the most challenging, requiring an extended, complex code integrating regex and HTML scraping. The best attempt was to process all pages individually, and it was able to clean about 65% of the selected documents.

4.1.2 Pre-Training and Fine-Tuning

Although we omitted the pre-training in favor of comparing existing models, in future development, we aim to compare a model pre-trained exclusively on a large corpus of cleaned SEC documents that will match the input data. When our cleaner can process the majority of documents, our goal is to pre-train the model with all filling from 2000 to 2021 while using documents from 2022 and 2023 as validation and testing. We will also explore custom vocabulary expansion to evaluate possible performance increases.

Regarding pre-training, we selected 3 companies for each of the eleven GICS sectors (Information Technology, Health Care, Financials, Consumer Discretionary, Communication Services, Industrials, Consumer Staples, Energy, Utilities, Real Estate, and Materials) and 67 random selected companies for a total of 100. Out of these companies, we selected one random yearly report. The composition by sector is shown in the following table.

GICS Sector	S&P 500 ratio	Fine-tuning dataset ratio
Industrials	0.155	0.16
Financials	0.143	0.10
Health Care	0.127	0.14
Information Technology	0.127	0.14
Consumer Discretionary	0.105	0.06
Consumer Staples	0.076	0.06
Real Estate	0.062	0.08
Utilities	0.060	0.08
Materials	0.056	0.09
Energy	0.046	0.03
Communication Services	0.044	0.06

Table 1: Ratio of GICS Sectors in S&P 500 and fine-tuning dataset

Considering the unbalanced nature of the target, we used FinBERT-ESG to select text blocks where the model was most confident about each label, using a 0.9 threshold for None, Environmental, and Social, and 0.65 for Governance. Then, we selected 250 for each detected label, followed by 500 text blocks where the model was uncertain (all scored less than 0.6) and 500 random samples for a total of 2000 text blocks. We manually labeled all of them into three separate categories (E, S, and G), with 0 if the text block is not part of that label and 1 if it is.

We divided training, validation, and testing with a 0.7/0.15/0.15 ratio. The resulting composition is shown in the following graphs.

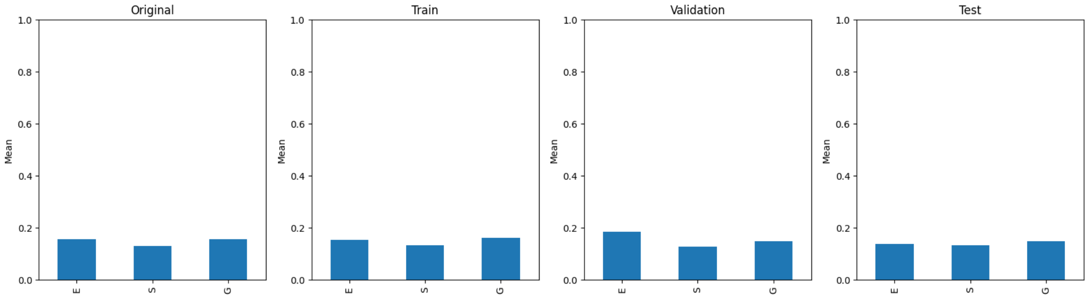

Figure 2: Original dataset, train, validation, and test breakdown.

As reported in the project schema (fig. 1), we plan to use an initial fine-tuned model to help generate a larger manually labeled dataset, aiming to reduce the classes' unbalanceness compared to the null category.

We acknowledge our basic knowledge in the financial analysis realm; therefore, without access to domain experts' advice, the quality of the training dataset is not ensured. We accepted a lower-quality training dataset for this proof of concept, although we will rely on expert advice for future development.

4.2 Deployment

The deployment phase will consist of creating a front-end and back-end system. Through this system, users will be able to input specific query parameters such as the Central Index Key (CIK), filing type, year, and desired categorization. Users will display a processed, cleaned document; this document will highlight paragraphs/sections relevant to the user's query. This design ensures an efficient and user-friendly interface for retrieving information needed for financial document analysis.

5. Evaluate FinBERT-ESG's Performance

To evaluate FinBERT-ESG's performance, we must consider that it has been trained for multiclass classification while we aim to create a multilabel classifier. We decided to compare the micro average and individual labels' F1 scores. We chose the F1 score due to its balance between precision and recall, which is useful when the target is unbalanced. FinBERT-ESG metrics are reported in the table below.

The micro average F1 score of 0.64 is 28% lower than the 0.89 reported in the original paper (Huang et al. (2023)). We assume the reason lies in the differences between its training task and our request. While the model was trained and tested on a single sentence

	F1	Acc	Prec	Rec
Micro Avg	**0.6364**	0.8756	0.5731	0.7153
E	0.8468	0.9433	0.8393	0.8545
S	0.4510	0.8133	0.3594	0.6053
G	0.5895	0.8700	0.5490	0.6364

Table 2: FinBERT-ESG's: micro average metrics and metrics by label

classification, we utilized the model to predict multiple labels of a paragraph or text section.

To corroborate this hypothesis, we analyzed the word and sentence counts of correct and wrong labels and found that the model performed better on shorter texts with three or fewer sentences. The following images report the results.

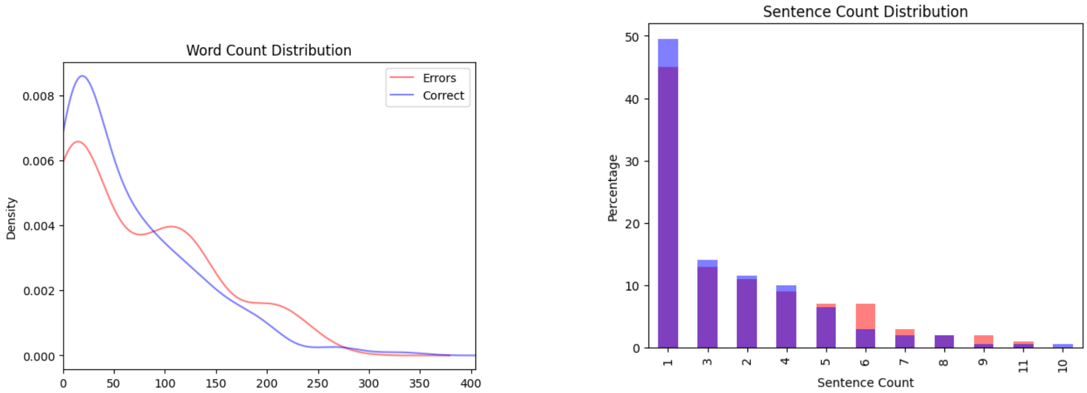

Figure 3: Word and sentence counts for FinBERT-ESG correct and wrong predictions

The best single category was Environmental, with an F1 score of 0.85, while Social and Governance performed considerably worse, 0.45 and 0.59, respectively. While manually labeling the data, we found labeling the environmental section simpler and more straightforward, with little doubt about the labeling. Aligned with Araci and Dogu's research, an LLM struggles in tasks that humans find more challenging.

6. Evaluate Fine-Tuned Models' Performances

We fine-tuned all three pre-trained models with our training dataset on Google Collab with one Nvidia A100, with training time ranging from 3 to 6 minutes. We found that the best hyperparameters were a performing learning rate of 3e-5, ArdamW optimizer, batch size 32, early stopping delta of 0.01, and patience 5.

Our best model was able to score a **micro average F1 score of 0.813**. The pre-trained model was *bert-base-uncased*. However, in agreement with Araci and Dogu's research, we did not find a huge gap between all the models. The best performer was inconsistent, and **all models score a micro average F1 score between 0.77 and 0.81** during multiple tests. The results highlight the ability of BERT to understand natural language and deeply adapt to new tasks quickly. The limited size of our dataset does not allow us to distinguish the performance graph from random fluctuation; therefore, we chose to show only our best model results as representative of all pre-trained models.

	F1	**Acc**	**Prec**	**Rec**
Micro Avg	**0.8136**	0.7433	0.8629	0.7697
E	0.9111	0.9733	0.9111	0.9111
S	0.7674	0.9333	0.8918	0.6734
G	0.7586	0.9300	0.7857	0.7333

Table 3: BERT base fine-tuned model: micro average metrics and metrics by label

Coherent with FinBERT-ESG, our model performs better in categorizing the Environmental label but with a smaller gap between metrics. Environmental labels score an F1 of 0.91, while Social and Governance score 0.77 and 0.76, respectively.

Despite our limited training data size and quality, BERT demonstrates its potential in financial multi-labeling tasks. We aim to increase our model performance up to state of the art. To achieve this goal, we plan to increase our fine-tuning dataset by up to 10,000 samples while increasing the quality and variety of the data. We believe that with a larger dataset, we will be able to distinguish the actual performance of the three models and use the result for a cost/benefits analysis of future pre-training.

7. Testing Insights

We tested two optimizers, AdamW and SGD (Stochastic Gradient Descent), and both converged to similar best model results on the test set. However, AdamW was able to converge in two or three epochs, while SDG converged over 30 epochs and required a more complicated learning rate schedule. The advantages of the first optimizer go beyond the saved time and deployment simplicity; with lower epochs needed and a lower learning rate. The following images show the learning rate curves utilizing AdamW.

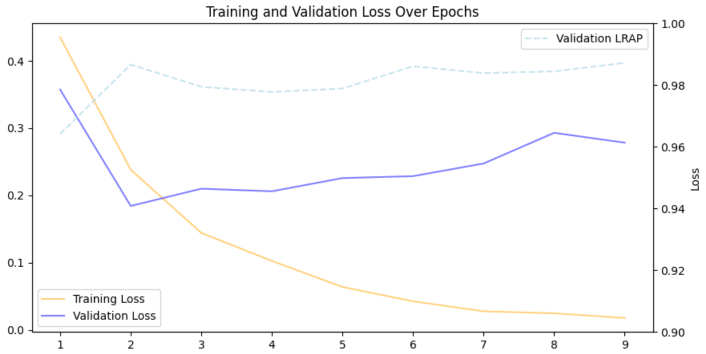

Figure 4: Model's learning rate curves with AdamW optimizer.

AdamW optimizer quickly found the minimum validation loss and started slowly rising after it; meanwhile, the training loss constantly decreased. This highlights a clear overfitting, which may be caused by Catastrophic Forgetting. To counter this trend, we adopt Early stopping, avoiding unproductive epochs, and saving considerable time. GU and TLR did not improve the test results, increasing the epoch and time needed to converge considerably.

8. Conclusions and Future Research

After inspecting existing literature, we were able to deploy an effective development schema that allows for recursive assistants to deploy better training datasets. The existing model for a similar task performs mediocre on multi-labeling tasks, outperforming our fine-tuned model. Despite room for improvement, we demonstrate the efficacy of BERT models on domain-specific documents, encouraging future work on a promising application. Lessons learned during experimentation and testing will guide cost-effective future research and development.

We will continue working on improving the SEC Cleaner to reduce unprocessed documents and expand its scope to 10-Q filings and other financial documents available at SEC. We will implement a prototype of the model to support manual labeling and expand the training dataset. We will repeat the experiments with a greater dataset to better understand the cost and benefit of developing a pre-trained model. We will also explore the split of the sentence-level dataset and expand the scope to other labels besides Environmental, Social, and Governance.

The promising result of this research highlights the potential and feasibility of applying BERT for financial text analysis.

We plan to explore different BERT variants and models that allow hierarchical and entire document understanding, such as HAN (Hierarchical Attention Networks) and GPT (Generative Pre-trained Transformer).

References

Amel-Zadeh, A., & Serafeim, G. (2017). Why and how investors use esg information: Evidence from a global survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2925310

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. https://doi.org/10.48550/ARXIV.1908.10063

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

Eccles, R. G., & Klimenko, S. (2019). The lack of standards in esg reporting. *MIT Sloan Management Review*, 60(4), 12–15.

European Financial Reporting Advisory Group. (2023). European sustainability reporting standards (esrs) [Accessed: date-of-access]. https://www.efrag.org/lab6

FiQA 2018 Task: Aspect-Based Financial Sentiment Analysis [Accessed: date-of-access]. (2018).

Friede, G., Busch, T., & Bassen, A. (2015). Esg and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance amp; Investment*, 5(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

Huang, A. H., Wang, H., & Yang, Y. (2023). <scp>finbert</scp>: A large language model for extracting information from financial text*. *Contemporary Accounting Research*, 40(2), 806–841. https://doi.org/10.1111/1911-3846.12832

LI, F. (2010). The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of Accounting Research*, 48(5), 1049–1102. https://doi.org/10.1111/j.1475-679x.2010.00382.x

Maia, M., Handschuh, S., Freitas, A., Davis, B., McDermott, R., Zarrouk, M., & Balahur, A. (2018). Www'18 open challenge: Financial opinion mining and question answering. https://doi.org/10.1145/3184558.3192301

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2013). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782–796. https://doi.org/10.1002/asi.23062

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8, 842–866. https://doi.org/10.1162/tacl_a_00349

Sullivan, R., & Mackenzie, C. (2017). The importance of esg factors in financial analysis and decision making. *Responsible Investment*, 1–15.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? https://doi.org/10.48550/ARXIV.1905.05583

Timothy Busch, R. B., & Orlitzky, M. (2016). Esg and financial performance: Aggregated evidence from more than 2000 empirical studies. *Journal of Sustainable Finance & Investment*, 5, 210–233.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/ARXIV.1706.03762