Automatically assessing social disclosure in sustainability reports

Rike Benjamin, Nisi Asbjørn
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

***Abstract*—During the last decade, the interest in Environmental Social and Governance (ESG) scores and rankings have skyrocketed. Investors, policymakers, and the public are concerned with how companies deal with the climate, their employees, and society as a whole. Concurrently, Deep Learning and Natural Language Processing (NLP) have undergone a revolution thanks to increased processing power, data availability, and recently more advanced algorithms. In this paper, we combine the two by automatically assessing the level of social disclosure in sustainability reports. We achieve this by fine-tuning a SmallBERT transformer on classifying a sentence as either social disclosure or not according to the GRI standards. In the end, we achieve an accuracy of 95% and an F1-score of 95%. Furthermore, we use the sentence scores to build company social scores for 89 companies on the S&P 500. The company scores are then used in a panel data regression to determine social scores' impact on financial performance. We find no statistically significant relationship between social scores and companies' excess return over the S&P 500 index.**

I. Introduction

Environmental, social, and corporate governance (ESG) has steadily gained prominence over the last two decades. Executives no longer solely focus on shareholder value. Instead, factors like the environment, the community, and the employees must be considered together with the shareholder value [1]. Institutional investors have started to complete ESG due diligence as part of the deal process. Bain & Company found that 70% of private equity companies consider ESG in their investment policies [2]. The Norwegian Bank Investment Management, the owner of, on average, 1.3 percent of the listed stock in the world, avoids investing in coal, tobacco and other ethical controversial operations [3]. As a part of this trend, the focus has increased on assigning ESG scores or ratings to companies to enable comparing companies against each other.

Today, most ESG ratings come from third-party vendors like Bloomberg, Dow Jones, and MSCI. However, Dimson et al. [4] show that ESG ratings tend to diverge and exhibit little correlation across vendors. Additionally, the data from such vendors are primarily limited to paying users and are therefore unavailable to retail investors or academics without access to these portals. In this paper, we propose a method for mitigating those problems by using NLP methods for automatically creating a social score from Corporate Social Responsibility (CSR) reports. We limit ourselves to evaluating the amount of disclosure a company makes on its social pillar by matching sentences to GRI standards [5]. The evaluation is done by fine-tuning a pre-trained SmallBERT transformer to distinguish between sentences that disclose on a social GRI standard or not. However, we also propose how this could be extended to develop a social score by changing the problem formulation of our NLP model.

Impact investing has gained popularity over the last decade due to the previously mentioned factors and the hypothesis that strong ESG fundamentals correlate with a strong executive team [6]. However, some argue that stock performance and ESG are a trade-off. There are divergent opinions on whether ESG performance impacts financial performance in the literature as well [6]. Therefore, we utilize the social score we build in this report to investigate whether we find any significant correlation between social scores and financial performance. We use the sentence scores from our classification model to create report scores. These scores are then used in a panel data regression model to investigate how social disclosure impacts financial performance.

Main contributions:

- We propose a fine-tuned BERT transformer capable of determining whether a sentence discloses on a social GRI standard with an accuracy of 95% and an F1-score of 95% on the test set.
- We show how one can use the sentence-wise BERT transformer to automatically assess the level of social disclosure in CSR reports.
- We find no statistically significant relationship between the aforementioned social scores and financial performance.

The code, dataset and models used in the report can be found in our GitHub repository[1].

II. Related works

A. Building social scores

Historically, most works on building ESG and social scores have focused on subjective evaluations towards a reporting standard. However, with the advent of the new era

[1]https://github.com/benjaminrike1/social_score

of deep learning in 2012, researchers have tried exploiting the power of big data to draw conclusions.

Luccioni et al. [7] develop a tool, ClimateQA, which enables users to automatically identify climate-relevant sections of sustainability reports. They use a RoBERTa transformer on a dataset of 2249 sustainability reports to pre-train a word embedding that is context-specific to sustainability reports. Furthermore, they fine-tune the model on labeled data to determine whether a sentence answers a TCFD question. Every TCFD question is a question developed by the Task Force on Climate-related Financial Disclosures to assess the climate disclosures of companies. The tool then uses the transformer to automatically search a document for climate-relevant sections.

Chen et al. [8] use various NLP methods to measure companies' alignment with the UN's Sustainable Development Goals (SDG). They extract sustainability reports from companies on the Russel 1000 index and extract features by identifying important unigrams and bigrams in the reports. Furthermore, they add features by finding synonyms for the uni- and bi-grams through a word2vec model. They associate every report to SDGs by using Refinitiv's ESG scores as labels. The authors then apply logistic regression, SVM, and random forest to create a supervised model automatizing the process. The SVM model is able to measure alignment with an accuracy of 80%.

B. Financial analysis

There is an abundance of papers investigating the relationship between ESG and financial performance. The papers typically consider an ESG ranking from external sources as input to an econometric model which assesses the financial impact.

Borovkova and Wu [9] investigate 2000 companies over nine years to assess whether their ESG score affects financial performance. They use ESG scores from Refinitiv, a global provider of financial market data, in a panel data regression to identify ESG's impact. The authors find that larger companies tend to have better ESG scores. They control for this effect by introducing the company's size in the model. The panel data model also uses cash flow, book price, and stock volatility as independent variables. Their results show that ESG performance seems to have a different impact in different regions. The correlation is negative in the US and Asia but nearly neutral in Europe and Australia. In addition, the authors find highly ESG-performing companies to have lower volatility.

La Torre et al. [10] study how the ESG index impacts companies' financial performance. Looking at 46 companies from the Eurostoxx50, the authors use a two-step methodology consisting of a panel data model with fixed and random effect models and a multivariate linear regression. The multiple linear regression separates the effect into environment, social sustainability, and governance, to assess which factors are most important. In addition, the authors add macro variables such as the Euribor rate and the unemployment rate to the model. They find the correlation between ESG and stock performance very weak or absent.

Whelan et al. [6] perform a meta-study of papers examining the relationship between ESG and financial performance. Assessing more than 1000 research papers, they find a positive relationship in 58% of the studies focusing on stock return, while 13% of the studies showed neutral, 21% mixed, and 8% negative relationships. Further, they find that ESG performance becomes more critical in long-time horizons and periods of economic and social crisis. The authors mention the lack of standardization of the ESG data as a complicating factor in many of the studies. Furthermore, they compare studies looking at the effect of ESG disclosure alone with studies using performance based ESG measures. They find that ESG performance and ESG disclosure do not affect financial performance equally and that ESG disclosure on its own does not drive financial performance.

III. Methodology

A. The data

To build social scores and assess financial performance, we studied sustainability reports, CSR reports, and ESG reports from 89 companies listed on the S&P 500 index (Appendix A). We retrieved reports annually from 2016 to 2020, giving a total of 500 reports.

The reports are read into Python using the PDF reader from the PyMuPDF-library [11]. PyMuPDF is an efficient way to read plain text from reports. However, it does not perform as well on tables and graphics. As reports come in all forms and shapes without a standard structure, finding a method that perfectly reads the reports is difficult, but the approach is good enough for our purposes. After reading the reports, we use the spaCy NLP library [12] to detect sentences in the extracted text. The method uses a dependency parser to determine the boundaries of the sentences. After the sentences are detected for each report, they are passed through the model pipeline. This methodology works out of the box and is helpful since punctuation often is deficient in the reports.

For the financial modeling, we extract yearly stock prices, total revenue, and market capitalization for each company

over five years. Furthermore, we retrieve the data for stock prices and total revenue from Yahoo finance [13], whereas market cap is retrieved from the data provider TIKR [14]. We use the last closing price for a given year for the stock prices and market capitalization.

B. Modelling the problem as a supervised learning task

The work in this report is based on the research question: *"Do companies who disclose more information on their social pillar provide superior returns?"* The question is two-folded. First, it requires us to define what a company's social pillar means and find a method to measure this in the form of a *social score*. Thereafter, we must examine if the social score built in the first part indicates a correlation between social sustainability reporting and the company's financial performance.

To build the social score, we must find a common definition of social disclosure and how to measure this for each company. For the definition, we utilize the standards of the Global Reporting Initiative (GRI) [5]. GRI is an independent and international organization working to standardize and communicate sustainability reporting. The GRI is the world's most widely used sustainability reporting standard. 16 of the GRI standards are directly related to social sustainability, covering areas like diversity, child labor, and the rights of indigenous people. The standards define the metrics' scope and standardize how to measure them. To link these standards with the sustainability reports' content, we define it as a classification problem, where every sentence in a report is classified as social disclosure or not. The idea is to train a model on classifying whether a sentence is a social disclosure as defined by the GRI standards. Below are three examples of sentences from reports, and excerpts from their related GRI standards.

Table I: Three example sentences and corresponding GRI standards.

Sentence	**GRI standard**
1) 127 550 children protected against the risk of child labor since 2012	GRI 408: Child Labor "Operations and suppliers at significant risk for incidents of child labor"
2) Alcoa-managed bauxite mines in Juruti and Western Australia are located on lands of significance to Indigenous or Traditional Communities	GRI 411: Rights of indigenous people "Incidents of violations involving rights of indigenous peoples"
3) We are further developing technologies on the basis of our open innovation concept	No related GRI standard

Table I showcases how a sentence from a report links to a concrete metric. We label sentences that link to a standard as 1s and the rest as 0s. We manually labeled the dataset by collecting 1311 sentences from around 50 sustainability reports (Appendix A). Out of these, 822 are social disclosure, and 489 are not. In addition, we added 3255 sentences from non-sustainability sources (Appendix A), giving a total of 4566 sentences for training and testing. Finally, we split the data into a 60/20/20 training, validation, and test set.

C. BERT transformer

To classify the sentences as explained in the previous part, we utilize a SmallBERT transformer fine-tuned with an extra dense layer. The SmallBERT transformer is a compact version of the more famous BERT transformer. To properly understand BERT and, more generally, transformers, one must first seek some historical context.

1) Reccurent Neural Networks: Before transformers were introduced, Recurrent Neural Networks (RNN), often in the form of Gated RNNs or Long Short-term Memory (LSTM), were state of the art for all things sequence modeling and hence language modeling. Shortly explained, the vanilla RNN uses hidden states to enable the network to remember information from previous states. The hidden states receive as input both the new input as well as its output from the previous step. However, the vanilla RNN poses several problems, 1) It suffers from exploding and vanishing gradients, 2) The network gradually forgets previous states, and 3) It operates sequentially, which makes training a time-consuming process [15].

The LSTM was introduced in 1997 by Hochreiter and Schmidhuber [16]. The LSTM solves the Vanilla RNN's long-term dependencies problem by splitting the memory into long and short-term. The model can store important inputs in the long-term memory to remember them for a long time, while it can use the short-term memory for short-term dependencies. The LSTM also allows an uninterrupted gradient flow through the network, which avoids the problems of exploding and vanishing gradients [16].

The state-of-the-art architectures pre-transformer used an encoder-decoder structure and the attention mechanism [17]. The encoder-decoder structure address the neural network's lack of ability to model sequences when the input and output size varies. The encoder encodes the input into a representation space, and the decoder extracts the output sequence from the representation vector [18]. Initially, the vector in the representation space was fixed in length. However, Bahdanau et al. [19] suggested using the attention mechanism to let the model itself decide on the vector's length. Attention achieves this by searching for the parts of the input sequence that is relevant for a particular

output.

2) Transformers: Although the modified RNNs presented in the previous section achieved significant progress, their intrinsic sequential nature was preserved. The sequentiality precluded parallelizing within training examples [17]. To address this issue, the transformer was introduced in 2017 by Vaswani et al. in the infamous paper "Attention is all you need" [17]. The authors suggested removing the models' recurrent part and relying only on an attention mechanism: *multi-headed self-attention*. Self-attention is an attention mechanism that only uses dot products between input vectors and softmax to create the output vectors [20]. The multi-headed part runs several self-attention layers in parallel to account for the possibility that different parts of the input will attend to different parts of the input sequence [20]. Transformers enable us to combine the long-term dependencies power of RNNs with the contextual power of Convolutional Neural Networks (CNN). The transformer introduced by Vaswani et al. [17] kept the encoder-decoder architecture, but, in later transformers, this structure has been dispensed with [20]. A more thorough walkthrough of transformers can be found in [17], [20], [21].

3) BERT: Devlin et al. [22] introduced BERT, short for Bidirectional Encoder Representations from Transformers, in 2018. At the time of the paper, BERT achieved state-of-the-art scores on many language tasks. The main strength of the architecture, in contrast to the previous state-of-the-art models like GPT, is that it is bidirectional, i.e., it can attend to context on both sides of the sequence [22]. This way, the model can be fine-tuned with only one extra layer to achieve state-of-the-art performance on NLP tasks [22].

The BERT architecture consists of several transformer blocks, stacked together to create an encoder. The base model has 12 transformer blocks, outputs a representation vector in a 768-dimensional representation space, and uses 12 attention heads [22]. Furthermore, BERT utilized WordPiece tokens as described in [23] to model the input sequences, e.g., by splitting the word playing into play and ##ing. This enables the model to match words both on semantic and grammatical functions.

To achieve the previously mentioned bidirectional self-attention, BERT is trained using masking. Masking allows attending to tokens on both sides of the current token by masking parts of a sentence and asking BERT to fill in the blanks. BERT's training pipeline also includes next-sequence prediction, i.e., given two sequences, BERT is trained to predict whether the sequences are consecutive or not. The original model is pre-trained on a large general corpus from English books and Wikipedia totaling around 3.3B words [20].

4) SmallBERT: The model we decided to use is SmallBERT, a compact version of BERT with a smaller number of transformer blocks, lower embedding dimension, and a corresponding lower number of attention heads. SmallBERT was introduced by Turc et al. [24] in 2019. The authors show that pre-training a compressed architecture of a large model can be more effective than using advanced techniques to compress a pre-trained large language model.

D. Our model

In the modeling, we use the SmallBERT model with 8 transformer blocks, an embedding dimension of 256, and 4 attention heads. We add a dense layer, with input size 256 and output size 2 with a dropout of 0.1, to the end of the encoder. The output is softmaxed to output probabilities for whether the sentence is a social disclosure.

Our model pipeline is in Figure 1. First, the input sentences are sent in batches of 32 to the preprocessing layer. The preprocessing layer converts the input sentences into WordPiece tokens and sends them into the SmallBERT model. Next, the SmallBERT model outputs a 256-dimensional embedding vector, the input for the dense classifier. To optimize the model, we use the binary cross entropy objective function and the Adam optimizer, which uses an adaptive learning rate and momentum to improve upon the vanilla Stochastic Gradient Descent (SGD) [25]. Lastly, we use early stopping during training. Early stopping is an effective technique for mitigating overfitting when training a complex model on a small dataset [26], a vital consideration as our dataset is only around 4500 sentences.

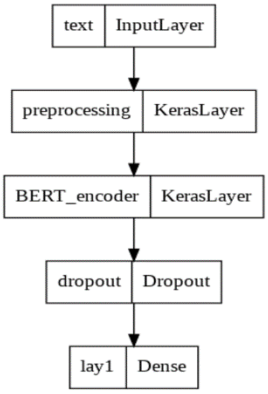

Figure 1: The architecture used for fine-tuning the model.

The advantage of using a compact model such as SmallBERT is that training and inference are fast while preserving performance. Also, transformer-based architecture has been shown to outperform other methods and become the go-to

model for natural language processing tasks [27]. A BERT-like architecture enables a vector embedding of sentences that provides solid representations for fine-tuning.

E. Social score

Running the model on our reports, we get the number of social disclosure sentences according to the GRI standard. We then build social scores for a report by standardizing this sum. The rationale is based on our research question. We aimed to assess how much a company discloses on its social pillar, not the relative frequencies of social disclosure in a report. An alternative way to build the scores is to normalize it by the total number of sentences in each report. However, the latter would mean that a report with 100 sentences, all related to social disclosure, would outperform a report of 10 000 sentences where 10 % of them are social disclosures. This does not assess how much a company discloses.

F. Financial modelling

To accurately determine the relationship between social scores and financial performance, we create a panel data regression. We utilized panel data as it combines data from a tabular and time-series dimension and thus models variations in both dimensions. The data is fitted as a pooled OLS, a fixed-effect model, and a random effect model. The pooled OLS disregards the panel data structure of the data and fits like an ordinary OLS. The fixed-effect model controls for time-invariant factors for individuals that bias the results. On the other hand, the random effect model assumes that the variation across entities is random and uncorrelated with the predictor or independent variables. We utilize the Hausman test to determine which model of random and fixed effects we use. The Hausman test tests whether the unique errors (υ_i) are correlated with the regressors. The Null hypothesis is that they are not and that one should use the random effect model.

As exogenous variables, we use the social score, the market capitalization, and the revenue growth. We choose market capitalization to control for company size as suggested by Engelhardt et al. [28]. Revenue growth is added as a growth measure as it is likely to impact a company's financial performance. Market capitalization is standardized to have values in the same range as the other exogenous variables. The endogenous variable is the yearly excess return of a company over the S&P 500 index. The equation for the fixed effect model can be found in Equation 1. Random effects and Pooled OLS is fitted with the same exogenous variable but vary slightly in structure.

$$excess_{i,t} = \alpha + \beta_1 score_{i,t} + \beta_2 cap_{i,t} + \beta_3 growth_{i,t} + \upsilon_{it} \quad (1)$$

where $excess_{i,t}$ is the excess stock return over the S&P 500 in year t for company i, $scores_{i,t}$ is the social score, $cap_{i,t}$ is the market capitalization, and $growth_{i,t}$ is the revenue growth.

IV. Results

A. Classification

Below, we report the results from the SmallBERT classification of sentences. In Table II, you find the accuracy and F1-scores for the train and test data. The model achieves an accuracy of 95% and an F1-score of 95% on the test set. Next, in Figure 2, we show the loss and accuracy development during training. Interestingly, the model performs strongly on the validation set after only three epochs. Finally, in Table III, one can find example sentences and their assigned scores. Note example sentence 4, which the model misclassifies.

Table II: Accuracy and F1-score for training and test data. The baseline model is always predicting the majority class which in this case is not social disclosure.

	Train		Test	
Model	**Accuracy**	**F1-score**	**Accuracy**	**F1-score**
SmallBERT	0.99	0.99	0.95	0.95
Baseline	0.82	N/A	0.78	N/A

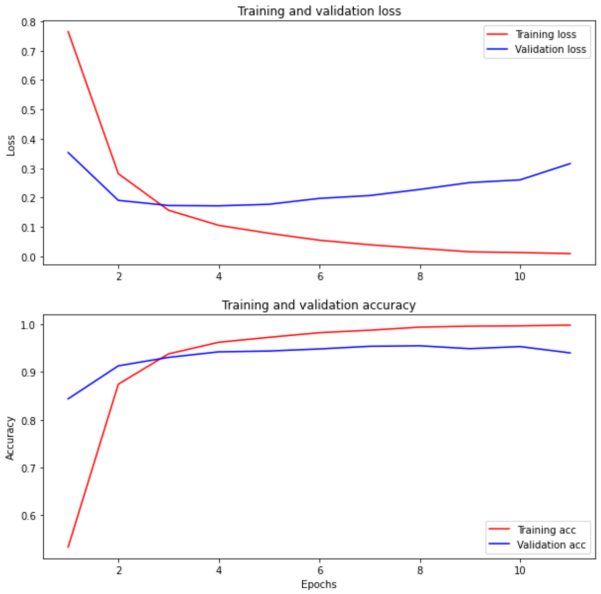

Figure 2: The loss and accuracy development during training.

Table III: Example sentences with corresponding output probabilities from the model. Especially note sentence 4 which is misclassified.

Sentence	**P(Social disclosure)**
1) As signatories of the Valuable 500 pledge, we are committed to putting disability on the business leadership agenda	1
2) Football is a sport	0
3) Our diversity was reduced this year	1
4) Across these three pillars we are driving our sustainability and technology ambitions to bring Credit Suisse into a sustainable future	0.77
5) As a leading global financial institution, Credit Suisse is deeply aware of its responsibility to clients	0.4
6) We lowered CO2 emissions this year	0
7) Zero fatalities occurred during the year	1

B. Social score results

Table IV shows the best and worst-performing companies and their respective social scores in 2020. Eleven companies were removed from the data set due to issues in the preprocessing. The final model therefore consists of the 89 remaining companies. The full results for all companies can be found in our GitHub repository [2].

Table IV: The five companies with the best and worst social scores in 2020.

	Company	**Social Score**
1	3M	3.53
2	Cigna	3.44
3	Procter and Gamble	3.324
4	Johnson and Johnson	2.936
5	Abbot	2.389
...	...	...
85	Accenture	-0.823
86	Starbucks	-0.837
87	J.MSmucker	-0.874
88	Stanley	-1.098
89	Boeing	-1.143

Figure 3 shows the distribution of social scores. The distribution is skewed towards the higher scores.

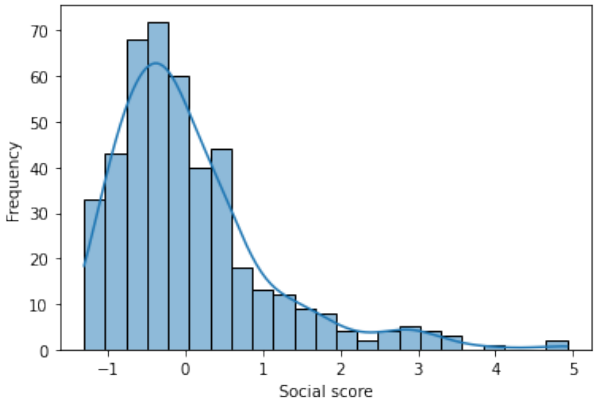

Figure 3: Distribution of social scores.

Interestingly we find that the social score is highly correlated with the total number of sentences in the reports. By running a linear regression we find that $R^2 = 0.76$ and $correlation = 0.87$ between the two. The scatterplot together with the regressed line can be found in Figure 4.

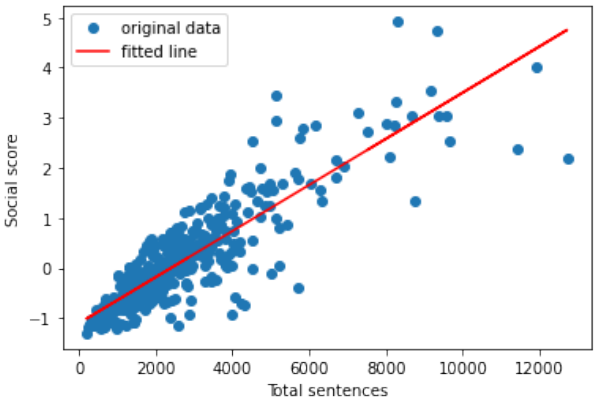

Figure 4: The social score and total number of sentences together with the regressed line. The R^2 is 0.76.

C. Finance results

We fit the panel data as a pooled OLS, a fixed-effect model, and a random effect model. For all three models, we find no statistically significant relationship between social scores and financial performance (p-values ranging from 0.20 to 0.29). We perform the Hausman test with the null hypothesis that we should use random effects. We reject the null hypothesis at a significance level of 1% (p-value = 0.0002) and decide to use the fixed-effects model. In Table V and Figure 5 you can find the correlations for social score with other variables.

[2]https://github.com/benjaminrike1/social_score/blob/main/df_financial.csv

Table V: The correlation between social score and the exogenous variables in the model.

	# sentences	Market cap	Rev. growth	Excess return
Social score	0.87	0.05	0.07	-0.03

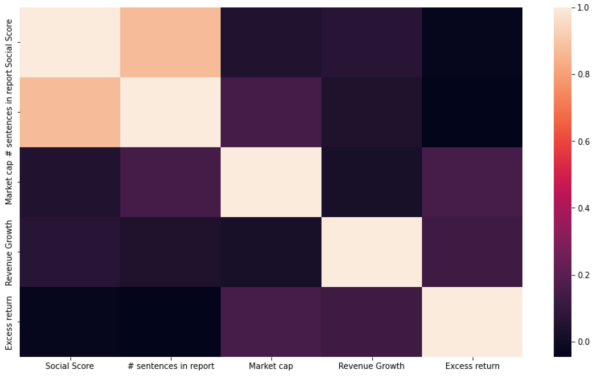

Figure 5: Correlations of the variables used in the panel data regressions. Lighter values indicate positive correlation, darker indicate little correlation.

Moreover, in Figure 6 the results from the fixed effects model can be found. The p-value for social score is 0.235 and we do not reject the null hypothesis that social scores do not impact financial performance at a significance level of 5%. Moreover, the 95% confidence interval for social scores includes values on both sides of zero, further indicating that there is no evidence for a relationship between social score and financial performance. The F-statistic for the regression is 0.72 which tells us that there is no evidence for joint significance either.

Fixed effects Estimation Summary

Dep. Variable:	excess_return	R-squared:	0.0061
Estimator:	PanelOLS	R-squared (Between):	-0.0461
No. Observations:	445	R-squared (Within):	0.0061
Date:	Wed, Jun 08 2022	R-squared (Overall):	-0.0093
Cov. Estimator:	Unadjusted	Log-likelihood	41.211
		F-statistic:	0.7239
		P-value	0.5383
		Distribution:	F(3,353)

Parameter Estimates

	Parameter	Std. Err.	T-stat	P-value	Lower CI	Upper CI
constant	0.0139	0.0120	1.1557	0.2486	-0.0098	0.0375
score	-0.0238	0.0200	-1.1905	0.2347	-0.0632	0.0155
market_cap	-0.0125	0.0266	-0.4683	0.6399	-0.0649	0.0399
growth	0.0207	0.0243	0.8513	0.3952	-0.0271	0.0686

Figure 6: The fixed effects model and its parameters.

V. Discussion

A. ML modelling

As shown in Table II, the model achieves an accuracy of 95% and an F1-score of 95% on the test set. The model strongly outperforms the baseline and is largely capable of learning the desired task. In Table III, we observe that the model's performance also is strong on a sample of sentences. It classifies all except for one example correctly. The misclassified example is sentence 4, where Credit Suisse states its ambitions about sustainability and a sustainable future. This is not a social sentence as it weighs more on the environmental side of ESG, even though sustainability also could be argued to be a social measure. This proves that the model still is not perfect.

The training data used for fine-tuning is limited in size. A high-value activity would be to manually gather more labeled data. We discovered an example illustrating the importance of data during training. At first, we only utilized 800 random neutral sentences in the model. We then achieved an F1-score of 87%. However, by changing to 3000 random neutral sentences, we achieved an F1-score of 95% on the test set. Therefore, it is natural to assume that more positive and neutral labeled data would increase the performance further.

Furthermore, we manually gathered the data for training, validation, and testing. This introduces a risk of a biased dataset due to subjective considerations. One problem this may have caused is that the dataset contains too many sentences in the following format: "We prioritize diversity." These sentences are no proof of actions or disclosures of actual status in the company. They are solely a subjective statement from the company. It would be valuable to be more thorough during the data gathering and only classify sentences that provide specific evidence, like numbers or actions, as positive. Another way to mitigate this problem could be to weight statements containing quantitative data. For example, when a company reports gender diversity as a percentage. This is a very concrete disclosure and should thereby be weighted more. However, we discarded weighting as deciding on the weighting scheme is a complicated exercise that quickly turns subjective, a property we wish to avoid in these ratings.

There is also room for improvement on the modeling side. We used the original BERT transformer in a compact form to achieve fast training and interference. The task we are learning is also lower in complexity than more advanced NLP problems, and we, therefore, argued that the loss in model complexity is a worthy trade-off. Nevertheless, it could be advantageous to utilize the whole model, as the vector representations will be better. Bert is also outperformed by more robust algorithms in NLP tasks today. For example, RoBERTa, a model that addresses BERT's undertraining, outperforms BERT on vital benchmarks [29].

B. Social score

The social scores developed in this project are essentially a quantification of companies' social disclosure in their sustainability report. As motivated earlier in this paper, this is a simple way to measure how much a company focuses on social sustainability. However, this methodology has some weaknesses. One major challenge is the lack of structure and standards in sustainability reporting. The GRI standards have contributed a lot to better this, but companies still provide their information in various ways. For example, some companies disclose a large part of their sustainability reporting in their annual reports, while others almost exclusively do this in their sustainability reports. We are thus likely to get biased results by only looking at the sustainability reports. A solution could be to add the annual reports to the data set. However, this could lead to double-counting of disclosures related to the same GRI standards. This effect is hard to control for and we decided to not do this.

Another weakness of our methodology is the lack of qualitative sentence assessment. We count a sentence related to social disclosure as one, regardless of whether the sentence is related to a violation of social rights, a neutral remark, or a measure to improve it. A way to avoid this problem would have been to frame the problem as a categorical learning task. Then, we could train the model on three categories of sentences: positive, negative, and neutral. With such an approach, a qualitative reporting assessment would have been possible. Another approach would be distinguishing sentences containing quantitative and concrete information from vague sentences about future visions. Such an approach would require much more data and a different approach to the data gathering process.

Lastly, it is essential to emphasize that our methodology cannot separate between lies and truth in the reporting. For example, if a company says it focuses heavily on reducing child labor in the value chain, our methodology will believe it. However, research has shown that greenwashing in sustainability reports is not necessarily a big problem. For example, Aikaterini and ManMohan [30] show that what companies claim in their report strongly correlates with a third-party assessment of their sustainability performance. Furthermore, a company reporting much on its social pillar will likely have this as a focus area. Therefore, we consider the social score an efficient tool to measure this.

C. Financial modelling

As Figure 6 shows, there is no evidence for a relationship between social scores and financial performance. We fit a pooled OLS, a random effect model, and a fixed-effect model but found no statistically significant coefficients for social scores in any of them. Through the Hausman test, we found evidence that the fixed-effect model is the correct model, and we decided to use that as our final model. The p-value for the social score in the fixed effect model is 0.235, and we do not reject the null hypothesis that social scores do not impact financial performance at a significance level of 5%. Moreover, the 95% confidence interval for social scores includes values on both sides of zero, further indicating no evidence for a relationship between social score and financial performance. The F-statistic for the regression is 0.72, which means there is no evidence for joint significance either. However, there are some factors we must consider before rejecting any relationship altogether.

We only ran the model on 89 companies over five years, totaling 445 reports. This is a small dataset, and we could probably strengthen the model by increasing the number of companies and the time period. The confidence interval for the coefficient for the social score would probably become narrower and the p-value lower. However, even in our current model, the coefficient for social scores is minimal. It is unlikely that a more extensive dataset will lead to much more robust results.

Another component to remember is what was discussed in the previous paragraph: social scores measure the amount of social disclosure, not social performance. Therefore, our results are only valid for social disclosure, not for a more complex social score that also measures social performance.

VI. Summary

During the last decade, ESG performance and scores have gained incredible popularity. The traditional way to assign ESG scores to companies is by subjective evaluations from experts. However, thanks to the rise of deep learning and the availability of data, automatic scoring has become a possibility.

In this paper, we propose fine-tuning a SmallBERT transformer to determine whether a sentence is a social disclosure or not. We achieve this by manually labeling a dataset from sustainability reports matching sentences to GRI standards. The fine-tuned transformer achieves an accuracy of 95% and an F1 score of 95% on the test set.

Furthermore, we use the individual sentence scores to build social scores for 89 companies listed on the S&P 500 over five years. The company scores are further used in a panel data regression to determine if there is any relationship between social scores and financial performance. We find no statistically significant relationship between social scores and the excess return over the S&P 500 for the 89 companies in the sample.

References

[1] D. Gelles and D. Yaffe-Bellany, "Shareholder value is no longer everything, top c.e.o.s say," *The New York Times*. [Online]. Available: https://www.nytimes.com/2019/08/19/business/business-roundtable-ceos-corporations.html

[2] "Esg in private equity." [Online]. Available: https://www.bain.com/industry-expertise/private-equity/esg-in-private-equity/

[3] "Responsible investment," Dec 2018. [Online]. Available: https://www.nbim.no/en/the-fund/responsible-investment/

[4] E. Dimson, P. Marsh, and M. Staunton, "Divergent esg ratings," *The Journal of Portfolio Management*, vol. 47, no. 1, pp. 75–87, 2020.

[5] "Continuous improvement." [Online]. Available: https://www.globalreporting.org/standards/

[6] T. Whelan, U. Atz, T. Van Holt, and C. Clark, "Esg and financial performance," *Uncovering the Relationship by Aggregating Evidence from*, vol. 1, pp. 2015–2020, 2021.

[7] A. Luccioni, E. Baylor, and N. Duchene, "Analyzing sustainability reports using natural language processing," *CoRR*, vol. abs/2011.08073, 2020. [Online]. Available: https://arxiv.org/abs/2011.08073

[8] M. Chen, G. Mussalli, A. Amel-Zadeh, and M. O. Weinberg, "Nlp for sdgs: Measuring corporate alignment with the sustainable development goals," *The Journal of Impact and ESG Investing*, 2021. [Online]. Available: https://jesg.pm-research.com/content/early/2021/12/12/jesg.2021.1.035

[9] S. Borokova and Y. Wu, "Esg versus financial performance of large cap firms: The case of eu, us, australia and south-east asia," *Probabilty and partners*, 2020. [Online]. Available: https://probability.nl/wp-content/uploads/2020/08/Refinitive_ESG_Analysis_in_4_Regions.pdf

[10] M. La Torre, F. Mango, A. Cafaro, and S. Leo, "Does the esg index affect stock return? evidence from the eurostoxx50," *Sustainability*, vol. 12, p. 6387, 08 2020.

[11] Python binding mupdf - pymupdf. [Online]. Available: https://pymupdf.readthedocs.io/en/latest/index.html

[12] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017, to appear.

[13] "Yahoo finance," https://finance.yahoo.com/, accessed: 2022-06-08.

[14] "Tikr," https://app.tikr.com/markets?fid=1&ref=3eixs1, accessed: 2022-06-08.

[15] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/abs/1706.03762

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014. [Online]. Available: https://arxiv.org/abs/1409.3215

[19] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014. [Online]. Available: https://arxiv.org/abs/1409.0473

[20] P. Bloem, "Transformers from scratch," Aug 2019. [Online]. Available: http://peterbloem.nl/blog/transformers

[21] A. Rush, "The annotated transformer," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 52–60. [Online]. Available: https://aclanthology.org/W18-2509

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: https://arxiv.org/abs/1810.04805

[23] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016. [Online]. Available: https://arxiv.org/abs/1609.08144

[24] I. Turc, M.-W. Chang, K. Lee, and K. Toutanova, "Well-read students learn better: On the importance of pre-training compact models," 2019. [Online]. Available: https://arxiv.org/abs/1908.08962

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[26] R. Caruana, S. Lawrence, and C. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13. MIT Press, 2000. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/059fdcd96baeb75112f09fa1dcc740cc-Paper.pdf

[27] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021. [Online]. Available: https://arxiv.org/abs/2106.04554

[28] N. Engelhardt, J. Ekkenga, and P. Posch, "Esg ratings and stock performance during the covid-19 crisis," *Sustainability*, vol. 13, no. 13, 2021. [Online]. Available: https://www.mdpi.com/2071-1050/13/13/7133

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[30] A. Papoutsi and M. S. Sodhi, "Does disclosure in sustainability reports indicate actual sustainability performance?" *Journal of Cleaner Production*, vol. 260, p. 121049, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0959652620310969

APPENDIX

Companies from the S&P 500 analyzed in the report:

3M Company, Abbott Laboratories, Accenture, Adobe Incorporated, Alaska Air Group Inc., AES Corporation, Amazon.com Inc., American Airlines Group Inc., Applied Materials Inc., Becton Dickinson and Company, Best Buy Co. Inc., Boeing Company, Boston Properties Inc., Caesars Entertainment Inc, Campbell Soup Company, Cardinal Health Inc., Carnival Corporation, Caterpillar Inc., Cigna Corporation , Cisco Systems Inc., Colgate-Palmolive Company, Comerica Incorporated, Conagra Brands Inc., ConocoPhillips, CSX Corporation, CVS Health Corporation, DENTSPLY SIRONA Inc., Duke Energy Corporation, Eastman Chemical Company, Equinix Inc., Estee Lauder Companies Inc., FMC Corporation, Goldman Sachs Group Inc., Hasbro Inc., Healthpeak Properties Inc., Henry Schein Inc., Hewlett Packard Enterprise Co., Home Depot Inc., Hormel Foods Corporation, Howmet Aerospace Inc., International Business Machines Corporation, International Paper Company, Interpublic Group of Companies Inc., Invesco Ltd., Iron Mountain Inc., J.M. Smucker Company, Johnson & Johnson, JPMorgan Chase & Co., KeyCorp, Keysight Technologies Inc, Kimco Realty Corporation, Leidos Holdings Inc., Lincoln National Corporation, Lockheed Martin Corporation, Lowe's Companies Inc., Micron Technology Inc., Microsoft Corporation, Mohawk Industries Inc., Mondelez International Inc., NiSource Inc, Northrop Grumman Corporation, NortonLifeLock Inc., Norwegian Cruise Line Holdings Ltd., NRG Energy Inc., NVIDIA Corporation, Omnicom Group Inc, Pentair plc, PepsiCo Inc., PNC Financial Services Group Inc., Procter & Gamble Company, Prologis Inc., PVH Corp., Qualcomm Incorporated, Quest Diagnostics Incorporated, Royal Caribbean Group, S&P Global Inc. , Seagate Technology Holdings PLC , Sealed Air Corporation, Southwest Airlines Co., Stanley Black & Decker Inc., Starbucks Corporation, Tapestry Inc., Texas Instruments Incorporated, TJX Companies Inc, Union Pacific Corporation, United Parcel Service Inc., United Rentals Inc., Verizon Communications Inc., Visa Inc., Vornado Realty Trust, Walt Disney Company, Waste Management Inc., Wells Fargo & Company, Whirlpool Corporation, Wynn Resorts Limited, Xylem Inc.

Companies for which we have used their sustainabilty or annual reports to label sentences for training data:

Apple, H&M, AGL, BHP, Orkla, Statnett, Statkraft, Veidekke, Nokia, John Deere, Airbus, ExxonMobil, Toyota, Nestle, Norsk Hydro, Meta, Volkswagen, NovoNordisk, Vattenfall, DNB, Lundin, Nordea, OOIL, Nutrien, Samsung, Huaweii, LVMH, MSC Sjippong, Berneck, Evergreen,Unilever, BCG, Alcoa, Sidel, Maersk, Zara, Rio Tinto,CHRB, BHP, Glencore, Ebay, McDonalds, General Motors, Walmart, SaudiAramco, AkerSoutions, Yara, Elkem, BHP, AGLenergy, Siemens, Schlumberger, British American tobacco, Nike, Lidl

Datasets used for unrelated neutral sentences:

- ASSET Corpus
- WikiQA