

Applying NLP to predict companies' ESG Readiness Score using Quarterly Earnings Transcripts, Press release and News Articles

Husein Adenwala

www.linkedin.com/in/Husein-Adenwala

huseinadenwala@gmail.com

Abstract

Sustainability investment based on Environmental, Social and Governance (ESG) factors has become mainstream and is now a key component of many investment strategies. However, the wide range of ESG factors, the complexity of comparing those factors between different companies and the lack of standardization in reporting and disclosure poses a challenge in predicting a company's ESG readiness score. The purpose of this research is to apply Natural Language Processing to capture the ESG factors in earning call transcripts and news articles to predict the ESG readiness score. Further, as research suggests that high ESG preparedness reduces systematic risk, a secondary goal was to test the robustness of the respective ESG readiness scores based on their correlation with the companies' stock price volatility (Beta).

This research used a BERTopic and Top2Vec for topic modeling on ESG/sustainability reports to create a training and test data set. After training and evaluating models, two pretrained Bidirectional Encoder Representations from Transformers models (BERT) were determined to be the most suitable – a finBERT-ESG model to classify E, S and G text followed by a pre-trained finBERT sentiment classifier to arrive at a sentiment count. The resulting sentiment count became the input to calculate a percentile ranking of mean ESG sentiments within the peer group for the companies that were part of this evaluation.

The findings of this research were inconclusive as the calculated ESG score and volatility had little to no correlation, as would have been expected. However, the topic modeling provided relevant topics that can be used to create label data for a classification model. The pretrained finBERT sentiment model did not perform as expected on ESG data as it incorrectly classified some of the text as negative words as it contained negative words, despite it actually having a positive meaning, like “decreasing carbon.” The finBERT-ESG model did perform well in classifying E, S and G text; however, this can be improved by using a larger training data set that can be created by topic modeling.

Introduction and problem statement

Integration of Environmental, Social and Governance (ESG) factors into investing strategies has gained mainstream popularity in the last few years. ESG factors cover a wide range of topics that may include corporate response to climate change, diversity, income equality and business ethics. These factors were traditionally not relevant to financial analysis, but climate change and human rights issues have driven a dramatic growth in sustainable investing, and this has made ESG financially relevant.

However, finding the data to support ESG scores is problematic. The primary source for ESG data is a company's annual ESG report based on guidelines provided by NGOs such as the Sustainability Accounting Standards Board (SASB) and Global Reporting Initiative (GRI) and news articles that provide ESG updates. As such, companies provide whatever reports they choose in whatever format they choose—there is no standard reporting and disclosure policy and there is a lack of timeliness in reporting

by companies. Additionally, ESG factors are so broad and cover so many facets, many of which are difficult to quantify and differ for each market sector and company type, that it is challenging to create a standard methodology to calculate the ESG score. As a result, there is a lot of variance in ESG scores between different ESG score providers.

These limitations of ESG reporting and variations in ESG ratings-based reports provide an opportunity to use NLP to identify and evaluate the ESG factors that are present in quarterly earnings calls.

Earnings calls are one of the key resources for investors and equity analysts. They are held on a quarterly basis by publicly traded companies and are the time at which earnings and other financial results are disclosed alongside the company's outlook. As companies include ESG into their business strategies, they include more information on ESG factors in these discussions.

Shareholders are increasingly demanding that organizations make progress on environmental, social and governance (ESG) issues. This has resulted in companies sending out press releases outlining their ESG initiatives.

As ESG news gets more attention online and drives traffic to financial news websites, these websites in turn provide more coverage on companies' ESG news.

NLP can be used to identify ESG factors that are discussed in earnings call transcripts. In addition to the earnings calls, press releases and financial news articles provide insights into companies' ESG factors. As such, NLP was applied to the earnings call transcripts, press releases and news articles in order to calculate a composite ESG readiness score.

Literature Review

Users find that there is lack of understanding on methodology that is used to calculate ESG ratings. Investors also find ESG ratings unreliable as the ratings provided by different rating agencies have low correlation with each other. This is surprising as they measure the same construct and should have high correlation (Larker et al 2022). For example, credit risk agency ratings have high correlation with each other.

There are no conclusive studies proving that ESG scores result in better investment performance (Bhagat 2022) but there is evidence that good corporate citizenship leads to higher ROE and enterprise value (Romero et al 2018, 36). There is also evidence ESG performance helps reduce stock price volatility (Zhou et al 2021, 202). This suggests that a high ESG readiness corresponds to lower relative volatility.

Because of the amount of manual work required to extract ESG information from transcripts and articles (and the high propensity for error), NLP is the best solution to extract ESG factors (Fishback et al, 2022).

The combination of unsupervised and supervised machine learning approaches can be a great solution to classify unlabeled data

First, using topic-modeling models such as BERTopic and Top2vec models for topic extraction, unlabeled data from sustainability reports were labeled E, S or G. One of the key advantages of these models over

models based on bag of words is that they use embeddings from a pretrained sentence transformer or universal sentence encoder to extract the semantic meaning of the sentences and therefore it can cluster documents/topics which are not just similar in words, but also similar in meaning.

Models such as Naive Bayes, support vector machine (SVM) and random forest are frequently used models for text classification; however, deep learning-based models have outperformed various classical machine-learning-based approaches in various text classification tasks, including sentiment analysis, news categorization, question answering, and natural language inference (Minaee et al, 2021, 2).

Of the deep learning models, attention-based models have outperformed simpler RNN-based models (Glassi et al, 2021). Attention-based models such as BERT and XLNet are currently among the best performing models for NLP tasks. However, attention-based models significantly increase the parameters of the model which make them computationally expensive to train.

With the explosion of large language models, the current approach to NLP modeling is to use a pre-trained language model or fine-tune a pre-trained model for specific tasks such as classification or sentiment analysis.

BERT - Bidirectional Encoder Representations from Transformers, is a deep learning model that was developed in 2018 by researchers at Google AI Language and it was a big milestone in the field of NLP. It can perform most common language tasks, such as sentiment analysis and text classification. BERT is open source and easily accessible, which is why finetuning a BERT model to use as a benchmark model for classification was the best approach.

For sentiment analysis, using a pre-trained model was the best approach as training data was not available without extensive manual annotation. This pretrained BERT model (FinBERT) was trained on a large corpus financial text from company annual and quarterly financial filings for sentiment classification. Specifically, the FinBERT-ESG model, a FinBERT model fine-tuned on manually annotated sentences from companies' sustainability reports, provided more accurate ESG-relevant classifications.

Data

Earnings calls transcripts, press release summary and news articles summary and Beta (stock price volatility) came from the Financialmodelingprep.com API. The data to predict ESG scores consisted of:

- 1st Quarter 2023 earnings calls transcripts for 493 companies listed in the S&P500 index
- Press releases and news articles for the same companies dated in the period ranging from the beginning of 1st Quarter 2023 to the date when the ESG scores were calculated i.e., 05/23/2023

	Period	Number of Records
Transcripts	2023 Q1	481
News	1/1/2023 to 5/23/2023	3233
Press Release	1/1/2023 to 5/23/2023	2517

Data used for topic modeling used to train and evaluate E, S, G classifiers included:

- PDF annual sustainability reports for 233 companies gathered by web scraping

	Period	Number of Records
Annual Sustainability reports	2023 - 2022	233

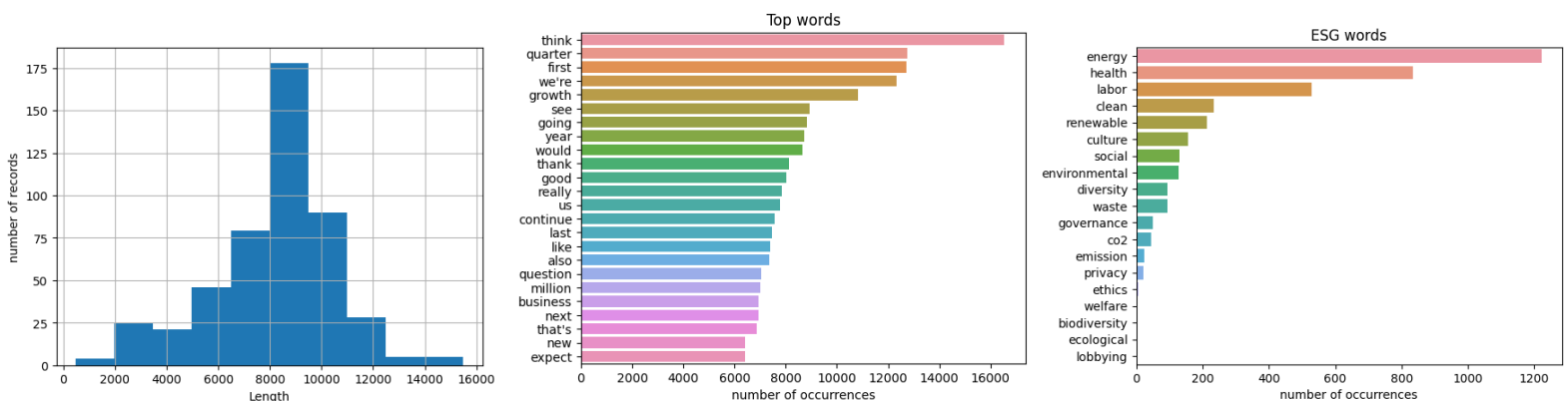
Data preparation, exploration, visualization

Text preprocessing included sentence tokenization, lowercasing, stop words removal, lemmatization, removing numbers, punctuation, white space and special characters. The text data in the earnings calls, news articles and press releases was converted to TF-IDF (Term Frequency - Inverse Document Frequency) and document term matrix (dtm) for training SVM, random forest and Naïve Bayes models.

Earnings calls had the most number of words on average, followed by sustainability reports then news articles and lastly press releases. Both news articles and press releases were summaries done by financialmodelingprep.com – even though they were summaries, they were still useful for identifying topics discussed.

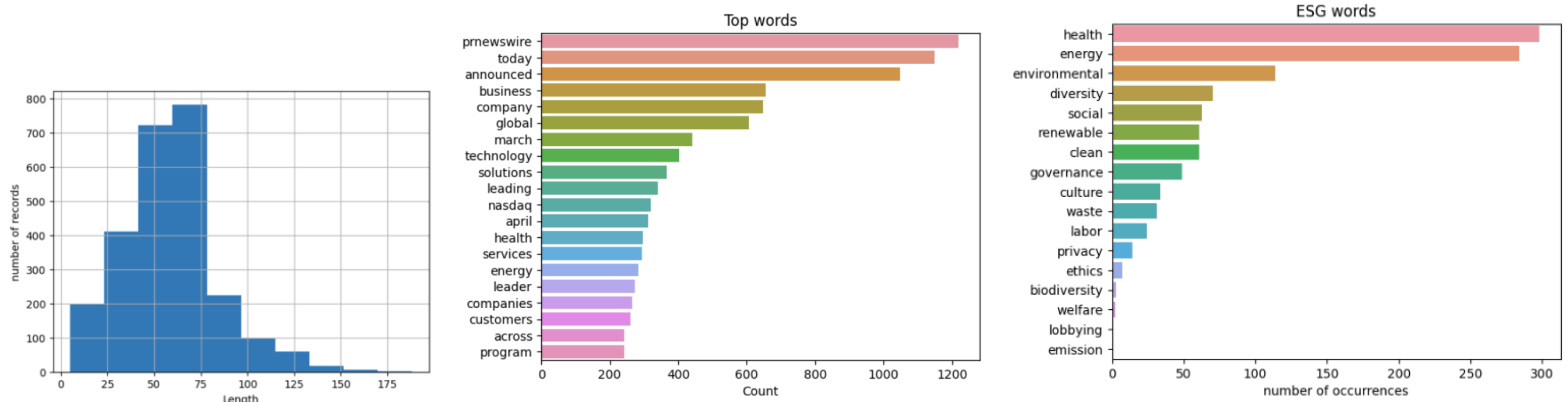
Plots: Earnings call transcripts

The earnings call transcripts used for this analysis had an average of 8400 words. The comparison of top words and ESG count show that ESG discussion was a relatively small part of the calls.



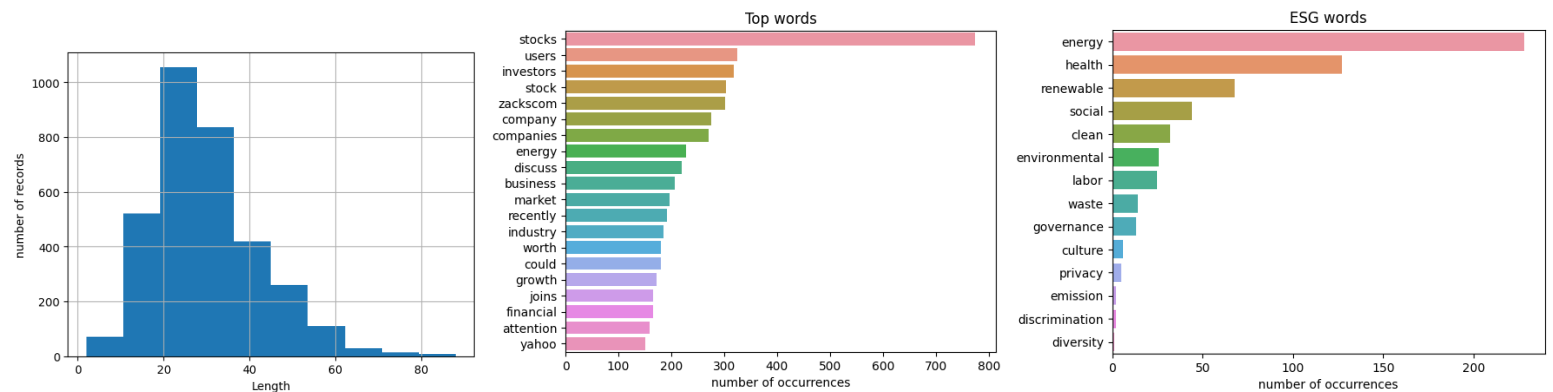
Plots for text in Press release summary

After stop word removal and lemmatization, press releases used in this analysis contained approximately 60 words on average per document. The comparison of top words and ESG count makes it evident that ESG discussion is relatively small but more frequent than in earnings call transcripts.



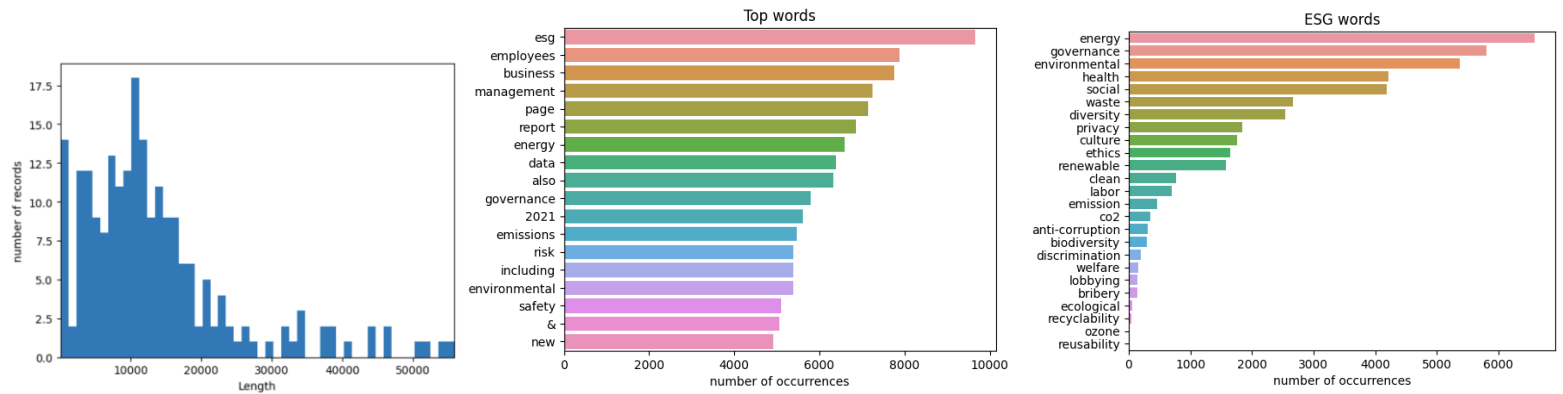
Plots: News article summary

News article summaries used in this analysis had approximately 30 words on average per document. The comparison of top words and ESG counts show that ESG discussion is relatively small but also more frequent than in earnings call transcripts.



Plots: Sustainability reports

On average, 1400 words extracted from PDFs of Annual sustainability reports. Top words contain many ESG terms, which suggest that the text is rich in ESG topics

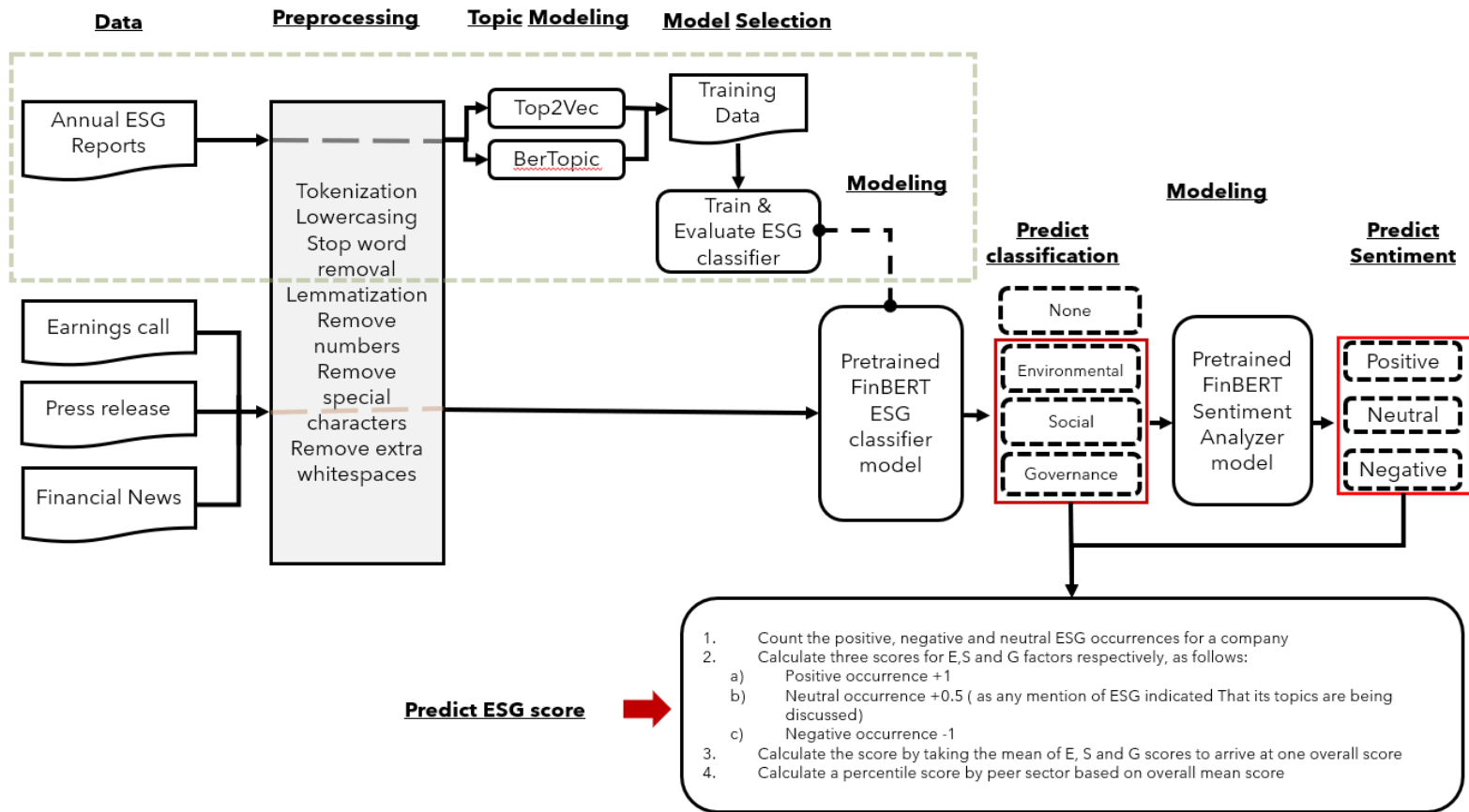


Methods

This research followed a 7-step process, from initial data collection through the ESG score calculation/prediction:

- 1- Find data sources
- 2- Preprocess data
- 3- Topic modeling
- 4- Model selection
- 5- Classification prediction
- 6- Sentiment prediction
- 7- ESG score prediction

The image below shows the relationship between these steps:



ESG score calculation:

The following formula was used to calculate the final ESG score:

$$Overall\ ESG\ Score = \sum_{k=symbol}^n E_{k,P} + 0.5 \times E_{k,N} - E_{k,Ne} + S_{k,P} + 0.5 \times S_{k,N} - S_{k,Ne} + G_{k,P} + 0.5 \times G_{k,N} - G_{k,Ne}$$

Where:

E is Environmental category

S is Social category

G is Governance category

k is list of ticker symbols

P is count of positive sentiment

N is Count of Neutral Sentiment

Ne is Count of Negative sentiment

$$\text{ESG score/percentile rank by peer group} = \left(\frac{f_b + \frac{1}{2}f_w}{N} \right)$$

f_b is the frequency below, the number of overall ESG score in peer group which are less than the score value of the percentile rank

f_w is the frequency within, the number of overall ESG core in peer group which have the same value as the score value of the percentile rank

N is the number of overall ESG score in the peer group

Method Explanation.

The first step was to use training data to train and evaluate different classifiers. However, because there are no publicly available ESG labeled training datasets, I leveraged topic modeling and text from 233 Annual ESG reports to create training and test data. These ESG reports specifically talk about ESG they are ideal candidates for extracting ESG topics.

Based on my model evaluation on the training and test data, a pre trained ESG Bert Classifier was chosen to extract ESG relevant text from earnings call transcripts, press releases and financial news feeds.

Next, a pretrained finbert sentiment classifier was used for sentiment analysis.

Lastly, using the positive, negative and neutral sentiment count, the composite ESG percentile score was calculated within each peer group.

Topic modeling on annual sustainability reports for 233 companies allowed for the extraction of 3000 sentences of E,S and G to create training data to train an E, S and G classifier.

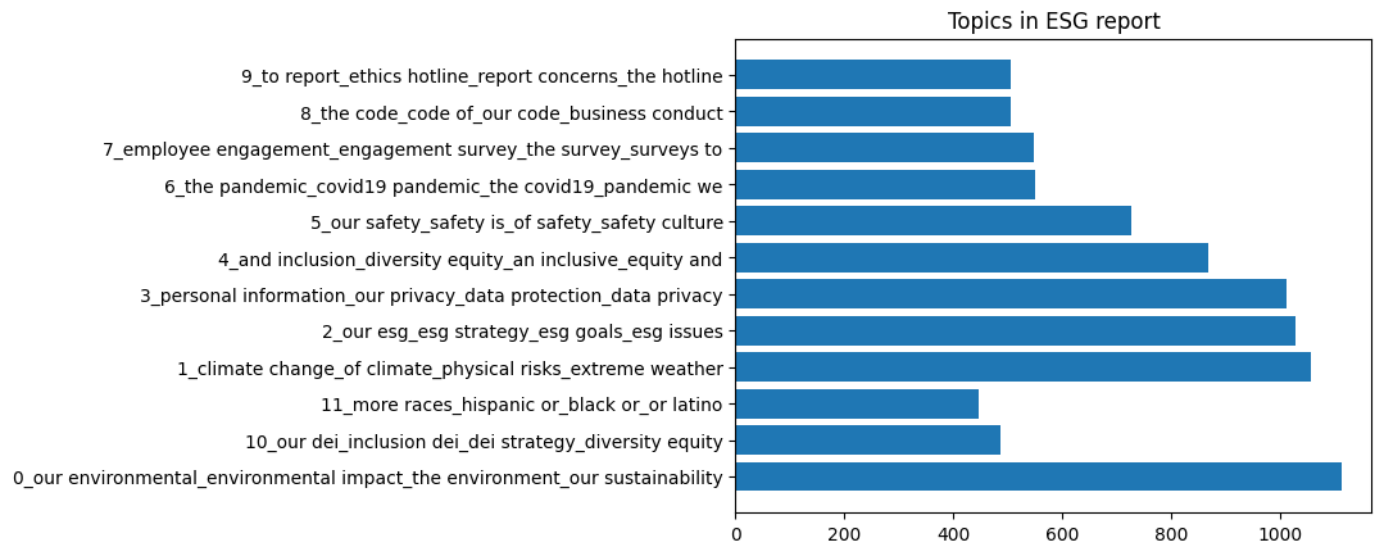
SVM, Naïve Bayes and Random Forest classifiers were trained and evaluated for classifying E, S and G; a BERT model was then fine-tuned using the training data.

The FinBERT ESG model (a pretrained Bidirectional Encoder Representations from Transformers (BERT) model), was specifically developed to classify E,S and G categories; the FinBERT model for sentiment analysis provided positive, negative and neutral sentiment scores for each category.

Results

Topic modeling

Topic modeling identified 10 Topics related to E, S and G factors by using top2vec and BERTopic. Below is the result of BERTopic. The 10 factors were mapped to E, S and G. Example of this and of Top2vec topics are in Appendix A.



ESG classification model

	Train/Validation		Test	
	Accuracy	F1 score	Accuracy	F1 score
SVM	99%	96%	52%	50%
Naïve Bayes	99%	98%	66%	64%
Random Forest	97%	95%	52%	50%
Fine-tuned BERT	94%	91%	68%	67%
pretrained FinBert-ESG	-	-	78%	77%

Using the data extracted by topic modeling, different models were trained and evaluated. The machine learning models such as Naïve bayes, SVM (Support Vector Machines) and random forest, which are efficient in speed, had a test accuracy of 55-65%. The fine-tuned BERT model had a test accuracy of 68% and a pre trained domain specific ESG model had the highest accuracy of 78% and was my model of choice.

ESG Score:

The ESG score was calculated in a linear manner by taking the sentiment type and ESG category type with a weight and then computing the percentile rank. The reason behind this was to keep the ESG score methodology interpretable—the percentile rank for a given sector gives it a rank which makes the score relative. Typically, financial metrics are compared on a relative basis depending on sector and business type. The image below shows ESG scores for the communications sector.

symbol	name	sector	Total_mean_score	ESG_score/Percentile_rank
ATVI	Activision Blizzard	Communication Services	1.833	0.396
GOOGL	Alphabet Inc. (Class A)	Communication Services	5.167	0.917
GOOG	Alphabet Inc. (Class C)	Communication Services	5.500	0.958
T	AT&T	Communication Services	1.333	0.250
CHTR	Charter Communications	Communication Services	2.333	0.500
CMCSA	Comcast	Communication Services	3.667	0.729
DISH	Dish Network	Communication Services	-1.167	0.042
DIS	Disney	Communication Services	4.500	0.854
EA	Electronic Arts	Communication Services	3.500	0.667
FOXA	Fox Corporation (Class A)	Communication Services	1.833	0.396
FOX	Fox Corporation (Class B)	Communication Services	3.667	0.729
IPG	Interpublic Group of Companies (The)	Communication Services	2.500	0.583
LYV	Live Nation Entertainment	Communication Services	2.500	0.583
MTCH	Match Group	Communication Services	2.167	0.458
META	Meta Platforms	Communication Services	-0.333	0.083
NFLX	Netflix	Communication Services	1.667	0.333
NWSA	News Corp (Class A)	Communication Services	1.500	0.292
NWS	News Corp (Class B)	Communication Services	1.333	0.208
OMC	Omnicom Group	Communication Services	4.333	0.792
PARA	Paramount Global	Communication Services	2.500	0.583
TMUS	T-Mobile US	Communication Services	4.500	0.854
TTWO	Take-Two Interactive	Communication Services	1.000	0.167
VZ	Verizon	Communication Services	8.000	1.000
WBD	Warner Bros. Discovery	Communication Services	0.833	0.125

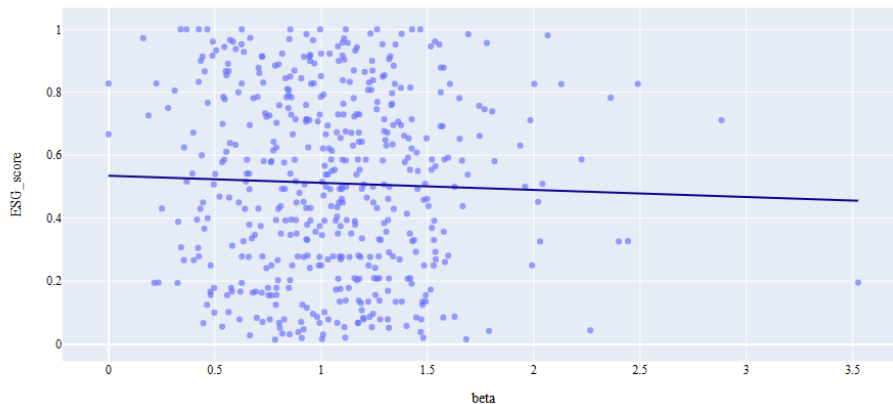
ESG score vs Beta

As there is no true label for ESG scores, comparing the scores with beta (stock price volatility) should show negative correlation based on the hypothesis that high ESG preparedness reduces stock price volatility. As beta gets closer to 0, there is less volatility; above 1 indicates higher volatility compared to the benchmark.

The regression plot for calculated ESG scores gathered in this analysis does not show any strong or moderate correlation with beta.

Instead, there is only a slight negative slope, which is statistically not significant, as shown by the t-test below. However, in some sectors, there is a moderate negative correlation (see Appendix B).

ESG Score vs Beta for S&P500 Companies



OLS Regression Results						
Dep. Variable:	y	R-squared:	0.001			
Model:	OLS	Adj. R-squared:	-0.001			
Method:	Least Squares	F-statistic:	0.5549			
Date:	Fri, 02 Jun 2023	Prob (F-statistic):	0.457			
Time:	04:25:17	Log-Likelihood:	-85.973			
No. Observations:	493	AIC:	175.9			
Df Residuals:	491	BIC:	184.3			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.5349	0.034	15.513	0.000	0.467	0.603
x1	-0.0225	0.030	-0.745	0.457	-0.082	0.037
Omnibus:	330.479	Durbin-Watson:	1.966			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30.003			
Skew:	-0.001	Prob(JB):	3.05e-07			
Kurtosis:	1.791	Cond. No.	5.16			

Analysis and Interpretation

Lack of True labels for ESG scores makes it difficult to interpret the outcome; however, the correlation with BETA can be used as a benchmark. Prior studies have shown that companies with high ESG preparedness have low price volatility. If this assumption holds true, there would be a negative correlation between ESG scores and beta. However, analysis shows that overall, there is no or little correlation between ESG and Beta.

One of the reasons for this may be that comparison of ESG score and beta requires longitudinal studies as beta is affected by other market factors.

Also, the E, S and G classification in the data is unbalanced. The number of data classified as Social is significantly higher than the Environmental and Governance data. Based on a review of the classifications, this is likely due to many false positive classifications, as some of the key words that are part of the Social topic area, such as safety and community, are commonly used in discussion, but not related to companies' social responsibility. The data was tokenized by sentence and where the sentences were short and had key words that are used in ESG, but not related to ESG, were still classified as ESG.

The pre-trained BERT model (FinBERT sentiment analyzer) that was trained on financial data did not perform as expected on ESG data as some sentences containing negative words such as reducing carbon output actually have a positive sentiment, but was classified as a neutral or negative sentiment.

Results from Topic modeling on annual sustainability show that the E, S, G and topics were approximately evenly distributed, which suggests that the imbalance between E, S and G classifications is not representative of the actual focus of the companies.

The richness of the E, S, and G topics in the annual sustainability reports suggests that those reports must be included in calculating ESG scores, but the analysis of text in periodic reports such as earnings calls, news, and press releases are still valuable as supplemental texts.

In summary, this work has borne out that:

- Topic modeling can be used for creating a classification model.
- FinBERT-ESG classification model does well in extracting relevant ESG sentences, but needs to be further refined, especially so that Social factors are not misclassified.
- FinBERT sentiment model does poorly on Environmental texts- misclassifying phrases such as “reducing carbon output” as negative or neutral when they should be positive.
- The correlation between calculated ESG score and beta is not clear. There is some indication that there is a small negative correlation, but not what would be expected. This means the model likely needs to be refined or the calculation of the ESG score itself might need to be revisited.

Conclusions

In this paper, we apply transformed based language model to classify E, S and G and Sentiment in earnings calls, news articles and press release text to calculate a percentile ESG score. However, as there is variation between ESG scores and rankings between the diverse ESG rating agencies, the accuracy of ESG scores cannot be assessed. Therefore, we use correlation with BETA to assess the validity of the ESG scores calculated in this paper as prior studies have shown that companies with higher ESG performance have low Beta. However, the comparison of calculated ESG score and Beta in this paper is inconclusive, despite a modest negative correlation in some sectors. More work is required to refine the method and model used here and experiment with a larger dataset.

Directions for Future Work

- Use Large dataset for Topic modeling to finetune Bidirectional Encoder Representations from Transformers models (BERT) classification model to create granular factor classification which would provide the ability to assign different weight to different factors depending on the sector type.
- A domain-specific sentiment model should be trained, and this will require manual annotation.
- Granular ESG factors should have different weights depending on the Sector type.
- Calculate ESG score for a large set of companies and for multiple periods as part of a longitudinal study.
- Compare scores resulting from this analysis across those determined by multiple ESG ratings companies, with the understanding that a comparison will always be relative given differences in scoring. Aggregate ratings can be used to create a regression model to predict weights of the sentiment and ESG classification parameters.

References:

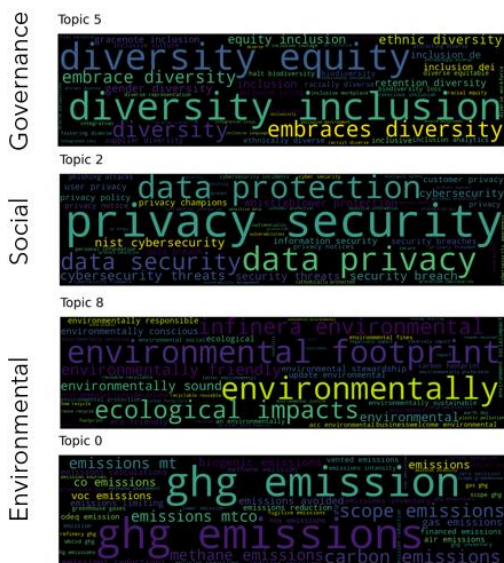
1. Bhagat, Sanjai. "An Inconvenient Truth about ESG Investing." *Harvard Business Review*, March 31, 2022. <https://hbr.org/2022/03/an-inconvenient-truth-about-esg-investing>.
2. Cerqueti, Roy, Rocco Ciciretti, Ambrogio Dalò, and Marco Nicolosi. "ESG investing: A chance to reduce systemic risk." *Journal of Financial Stability* 54 (2021): 100887.
3. Fischbach, Jannik, Max Adam, Victor Dzhagatspanyan, Daniel Mendez, Julian Frattini, Oleksandr Kosenkov, and Parisa Elahidoost. "Automatic ESG Assessment of Companies by Mining and Evaluating Media Coverage Data: NLP Approach and Tool." *arXiv preprint arXiv:2212.06540* (2022).
4. Géron, Aurélien. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.", 2022.
5. Grootendorst, Maarten. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure." *arXiv preprint arXiv:2203.05794* (2022).
6. Huang, Allen H., Hui Wang, and Yi Yang. "FinBERT: A large language model for extracting information from financial text." *Contemporary Accounting Research* (2022).
7. Jonsdottir, Bjorg, Throstur Olaf Sigurjonsson, Lara Johannsdottir, and Stefan Wendt. "Barriers to using ESG data for investment decisions." *Sustainability* 14, no. 9 (2022): 5157.
8. Larcker, David F., Lukasz Pomorski, Brian Tayan, and Edward M. Watts. "ESG ratings: A compass without direction." *Rock Center for Corporate Governance at Stanford University Working Paper Forthcoming* (2022).
9. Lee, Ook, Hanseon Joo, Hayoung Choi, and Minjong Cheon. "Proposing an integrated approach to analyzing ESG data via machine learning and deep learning algorithms." *Sustainability* 14, no. 14 (2022): 8745.
10. Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep learning--based text classification: a comprehensive review." *ACM computing surveys (CSUR)* 54, no. 3 (2021): 1-40.
11. Mukherjee, Mukut. "ESG-BERT: NLP meets Sustainable Investing". *Towardsdatascience* (blog). 10 September 2022. <https://towardsdatascience.com/nlp-meets-sustainable-investing-d0542b3c264b>
12. Niu, Liqiang, Xinyu Dai, Jianbing Zhang, and Jiajun Chen. "Topic2Vec: Learning distributed representations of topics." In *2015 International conference on asian language processing (IALP)*, pp. 193-196. IEEE, 2015.
13. Oeltz, Daniel, Jan Hamaekers, and Kay F. Pilz. "Parameterized Neural Networks for Finance." *arXiv preprint arXiv:2304.08883* (2023).
14. Romero, Silvia, Agatha E. Jeffers, Frank Aquilino, and Laurence DeGaetano. "Using ESG ratings to build a sustainability investing strategy." *The CPA journal* 88, no. 7 (2018): 36-43.
15. Whelan, Tensie. "ESG Reports Aren't a Replacement for Real Sustainability." *Harvard Business Review*, July 27, 2022. <https://hbr.org/2022/07/esg-reports-arent-a-replacement-for-real-sustainability>.
16. Zhou, Dongyi, and Rui Zhou. "ESG performance and stock price volatility in public health crisis: evidence from COVID-19 pandemic." *International Journal of Environmental Research and Public Health* 19, no. 1 (2021): 202.

Appendix A:

Custom name of ESG factors

Topic	Count		Name	CustomName
0	1665	0_of water_our water_water consumption_water use	Environmental	
1	1278	1_personal information_data protection_our pri...		Social
2	1160	2_our environmental_environmental impact_the e...	Environmental	
3	835	3_our safety_safety is_of safety_and safety		Governance
5	661	5_and inclusion_diversity equity_an inclusive_...		Governance
6	648	6_covid19 pandemic_the pandemic_the covid19_pa...		Social
7	639	7_climaterelated risks_physical risks_climate ...	Environmental	
8	584	8_employee engagement_engagement survey_the su...		Governance
11	459	11_human rights_rights and_rights policy_of human		Social
48	188	48_scope emissions_our scope_scope category_sc...	Environmental	

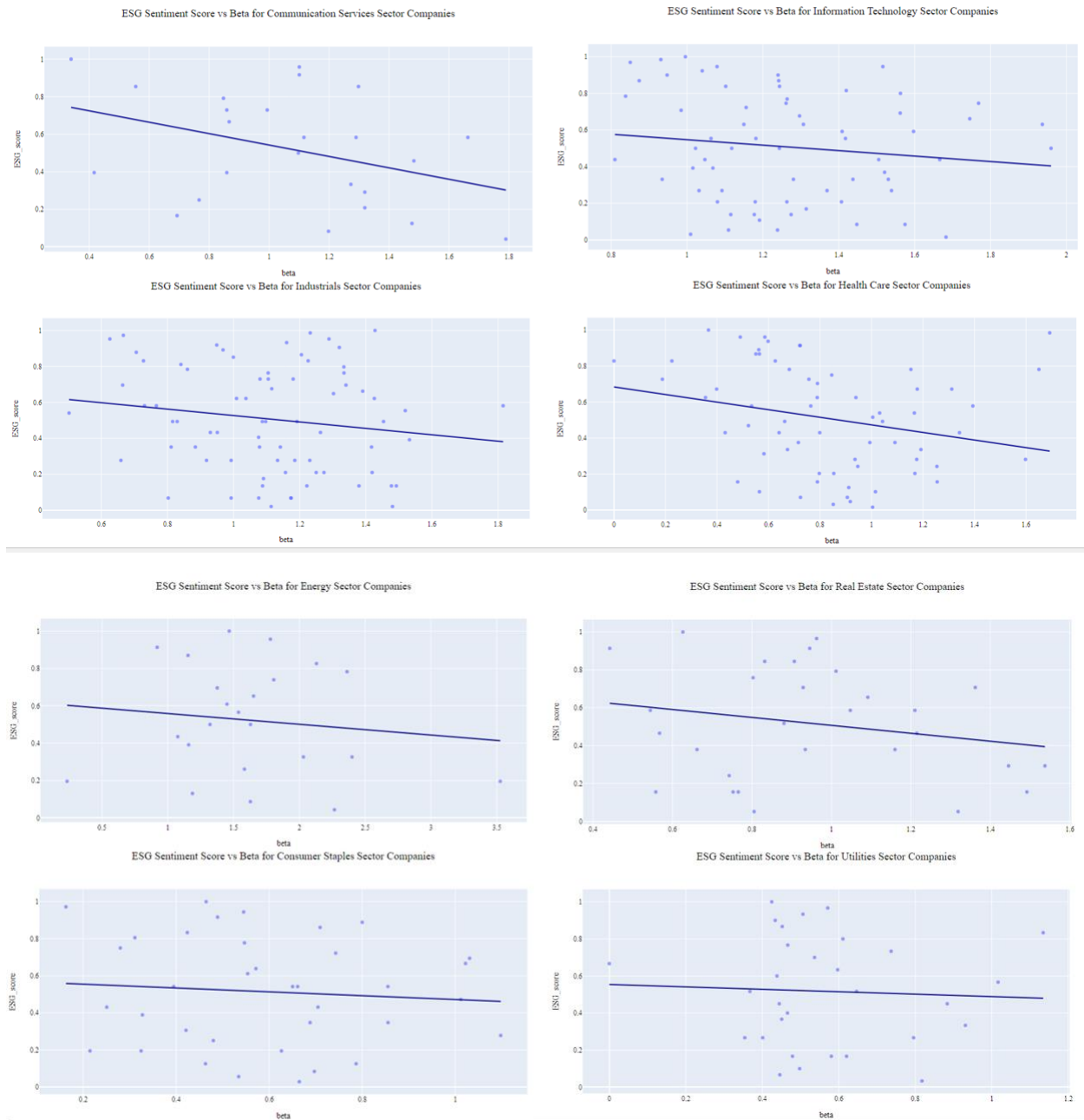
4 of the 10 topics as example of Top2Vec model output:



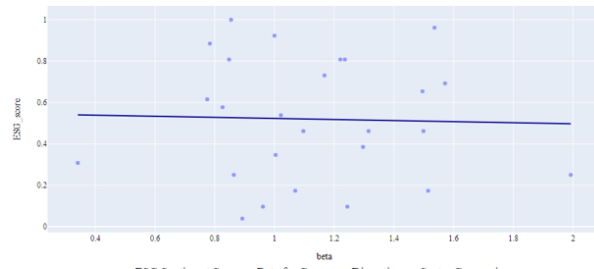
Appendix B:

ESG score vs Beta by sector

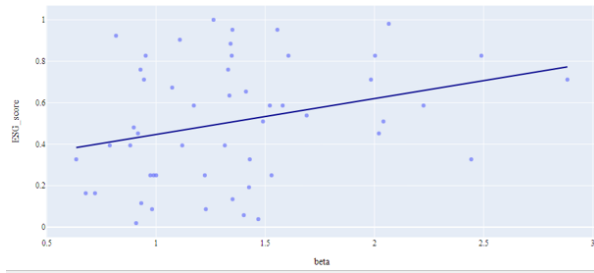
Analysis – ESG scores by sector



ESG Sentiment Score vs Beta for Materials Sector Companies



ESG Sentiment Score vs Beta for Consumer Discretionary Sector Companies



ESG Sentiment Score vs Beta for Financials Sector Companies

