Creating a Systematic ESG (Environmental Social Governance) Scoring System using Social Network Analysis and Machine Learning to Influence Company Practices to be More Sustainable

Student: Aarav Patel

Mentor: Dr. Peter A Gloor (MIT Center of Collective Intelligence)

March 4th 2022

Table of Contents

Abstract

ESG, or Environmental Social Governance, is a widely used metric that measures the sustainability and societal impact of a company's practices. ESG is determined using self-reported corporate filings, meaning companies often portray themselves in an artificially positive light. As a result, ESG evaluation is extremely subjective and inconsistent, and this is an issue since it gives companies mixed signals on what to improve. The purpose of this project is to create a data-driven ESG evaluation system to systematically and holistically analyze the good a company does for society by incorporating social sentiment. To do this, Python web-scrapers were made to collect posts from Wikipedia, Twitter, LinkedIn, Glassdoor, and GoogleNews across all S&P 500 companies. Irrelevant data was filtered using Named Entity Recognition, and the remaining data was cleaned via regex. Next, the Flair NLP algorithm calculated sentiment for each post, which was then averaged and mean-normalized to obtain scores for ESG subcategories. Using these features, machine learning algorithms such as SVR, KNN, XGBoost, Random Forest Regression were trained and calibrated to S&P Global ESG Ratings. The SVR algorithm displayed the strongest results with a mean absolute error of 13% and p-value of 0.021, showing it is well-calibrated to be implemented in a public setting. The student has carried out all parts of the project while the mentor provided guidance. Overall, systemizing ESG can encourage companies to adapt their supply chains and corporate practices to be more sustainable. This can rewire over $6.1 trillion to more sustainable/ethical initiatives.

Introduction

Companies often prioritize profits over social responsibility. 100 companies have been responsible for 71% of the global greenhouse gas emissions that cause global warming since 1998 (Carbon Majors Database). Despite this, many business leaders have said that they are fully on board with incorporating sustainability measures. In 2016, a UN survey found that 78% of CEO respondents believe corporate efforts should contribute to the UN Standard Development Goals, which are goals adopted by the United Nations as a universal call to action to end poverty and protect the planet (UN). While many executives have pledged to focus more on these areas of concern, people feel there has been little tangible progress made to more sustainability. In a more recent 2019 UN survey, only a fifth of responding CEOs felt that businesses are making a difference in the worldwide sustainability agenda (UN). These highlight a clear disconnect between sustainability goals versus sustainability actions.

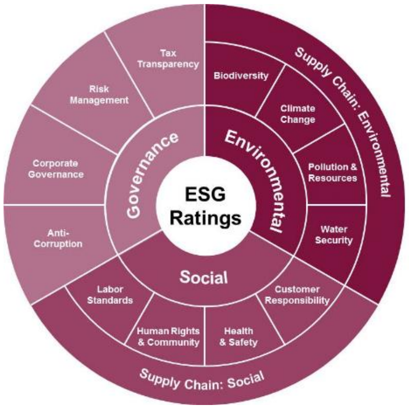


Figure 1: FTSE ESG evaluation framework

ESG, or Environmental Social Governance, is a commonly used metric that determines the sustainability and societal impact of a company's practices. ESG raters such as MSCI (Morgan Stanley Capital International), S&P Global, and FTSE (Financial Times Stock Exchange) do this by measuring sub-categories such as pollution, diversity, human rights, community impact, etc (figure 1). Measuring these areas are necessary since they encourage companies to rectify bad practices. This is because ESG ratings can influence public perception, credit ratings, government regulation, investor capital, etc.

At the moment, ESG is subjectively assessed using self-reporting company filings. As a result, companies can often portray themselves in an artificially positive light. This has caused what some feel is a disconnect between ESG ratings and a company's actual sustainability. According to Kenneth Pucker, a researcher at Tufts, self-disclosure has created "a problem where analysis is not complete, results are mostly unaudited, and they are not comparable." Therefore, this has led to subjective and inconsistent analysis between different ESG rating organizations, despite the fact they essentially seek to measure the same thing (Kotsanonis). For instance, the correlation among 6 prominent ESG rating agencies is 0.54. In comparison, mainstream credit ratings have a stronger correlation of 0.99 (Berg et al). This highlights how subjective assessment and limited data transparency can cause inconsistent ratings.

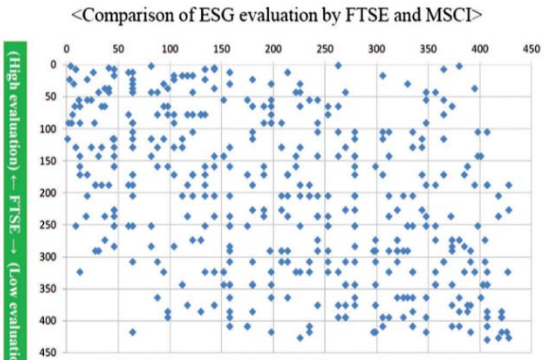


Figure 2: Low correlation between ESG firms such as MSCI and FTSE

Having more consistent and accurate ESG evaluation is important. Divergence and imprecision in ESG ratings hamper motivation for companies to improve since they are given mixed signals on what to change. Furthermore, failure to have accurate ESG standards limits the use of ESG for regulatory purposes. Finally, inconsistent ESG ratings allow larger companies with more resources to "game the system" by only publishing the highest score. This is why there is a significant positive correlation between a company's size, available resources, and ESG score (Drempetic). These issues ultimately defeat the purpose of ESG by failing to significantly motivate companies to improve their practices.

Purpose

The purpose of this project was to create a systematic ESG rating system that gives a more precise, balanced, and accurate view of a company's practices. To do this, a machine-learning algorithm was created that employs social network data to quantitatively evaluate ESG. social network data was used as opposed to self-reported filings to holistically determine ESG. This, in turn, can increase the prevalence of ESG and influence company practices to be more socially responsible.

To ensure that the proposed index does not deviate much from current ESG solutions, the project should have a statistically significant correlation with current ESG ratings as well as a mean absolute average error (MAAE) <20%. However, some potential constraints include limited access to high volumes of data, limited computational resources, and overall the complexity of such a large-scale project.

Methods

The creation of this project was divided into 3 steps. The first step was data collection, where posts/news surrounding a company were gathered and cleaned. Afterward, text data was converted into numeric scores using Natural Language Processing. Finally, machine-learning algorithms were trained using this data to systematically compute a more balanced ESG score.

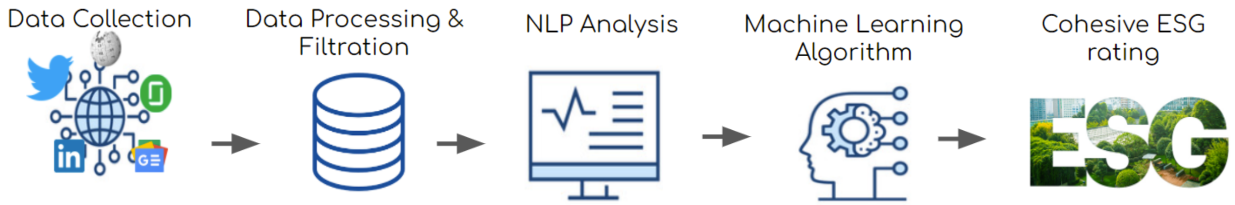
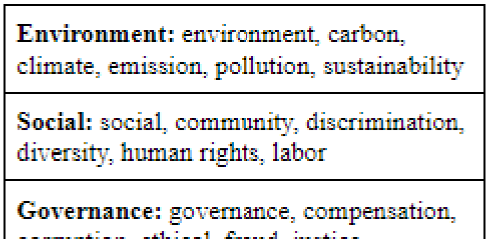


Figure 3: An overview of how the data-driven ESG index uses social network data to compute a cohesive ESG rating

Data Collection

Rather than use biased self-reported corporate filings, social network data was used to holistically quantify ESG. Social network analysis and web scraping can be used to identify trends (Gloor). Popular social networks such as Twitter, LinkedIn, and Google News have a plethora of data pertaining to nearly any topic. This data can provide a balanced view of company ESG practices, and it can help cover both short-term and long-term company ESG trends.



Environment: environment, carbon, climate, emission, pollution, sustainability
Social: social, community, discrimination, diversity, human rights, labor
Governance: governance, compensation, corruption, ethical, fraud, justice

Figure 4: Keywords/Topics used for data collection

To do this, a comprehensive list of ESG relevant keywords was created (figure 4). This list was used to help collect publicly available company data from Wikipedia, LinkedIn, Twitter, Google News, and Glassdoor. To collect data, web scrapers were developed in Python. Afterward, data for each sub-category was stored in a CSV file.

- **Wikipedia:** Wikipedia data was collected using the Wikipedia Application Programming Interface (API). Wikipedia serves to give a general overview of a company's practices
- **Google News:** Google News data was collected by identifying top news articles based on a google search. The links to these articles were stored. The news serves to give overall updates on notable ESG developments
- **Twitter:** Twitter data was collected with the help of the snscrape library. Snscrape is a lightweight API that allows users to collect near-unlimited Tweets (with certain restrictions on how many can be collected per hour) from almost any timeframe. Twitter was chosen to primarily give consumer sided feedback of a company's practices
- **LinkedIn:** Since the LinkedIn API does not support the collection of LinkedIn posts, an algorithm was created from scratch to do so instead. The algorithm utilized the Selenium Chromedriver to simulate a human scrolling through LinkedIn pertaining to a query. Based on this, each post's text was collected and stored using HTML requests via

BeautifulSoup. LinkedIn serves to provide more professional sided information on a company's practices

- **Glassdoor:** The Glassdoor API, unfortunately, had limitations that prevented me from using it to collect data. As a result, Glassdoor company meta-reviews were manually collected and recorded to a spreadsheet. This helps consider employee-sided feedback within the ESG evaluation model

These five social networks allow the majority of ESG data pertaining to a company to be collected. Data was collected for most S&P 500 companies (excluding real estate). This ensures the collected companies were well balanced across sectors and industries. The web-scrapers attempted to collect at least 100 posts/articles for each keyword on a social network. In order to speed up collection, multiple scripts were run simultaneously. At first, the programs would often get rate-limited for collecting so much data in such a short timeframe. To resolve this, safeguards were added to pause the program in case it encountered this. It took around a week to collect all the data, and on average, it took between 20-30 minutes for each company. All data collection was done in accordance with each site's terms and conditions

Once all data was collected, it was exported onto a spreadsheet for further analysis. Data was preprocessed using RegEx (Regular Expressions). Preprocessing helped remove/reformat URLs, mentions, reserved words, emoji, or other uncommon characters from the analysis. This helps filter out words/characters that might interfere with NLP analysis.

NLP Analysis

After data was cleaned and organized, an algorithm was built for analysis. Firstly, an ESG relevancy algorithm was created in order to filter out ESG irrelevant data that might obstruct results. In order to do this, keyword detection was used to see if the post/article discussed the current company as well as one or more of the ESG sub-categories. Next, Python's Natural Language Toolkit (NLTK) Named Entity Recognition was used to ensure that a post discussed the company rather than something else. For example, if the query "apple climate" was searched, then a post might come up saying "Spring is the optimal time to grow apple trees." However, named entity recognition would be able to identify that this sentence is not ESG relevant since "Apple" is being used as an adjective. Therefore, the algorithm will disregard it from the analysis. On the other hand, if the post said "Apple is pouring 500 Million dollars into initiatives for climate change," then the algorithm would determine that the post is talking about Apple the organization. This filtration step helps remove most irrelevant information which helps ensure only high-quality data is used.

After filtration, NLP sentiment analysis algorithms were used to numerically score whether a post was ESG positive or negative. Two NLP algorithms were created to do this: the short-post NLP algorithm analyzed shorter bodies of text (Twitter, LinkedIn Posts), while the long-article NLP algorithm analyzed longer ones (News, Wikipedia).

At first, the short-post NLP algorithm was created using NLTK sentiment analysis. However, it was found that this algorithm was ineffective at properly evaluating sentiment. So, a meta-analysis of different Python sentiment analysis libraries was carried out. After comparing various sentiment analysis libraries such as TextBlob, VADER, FastText, Flair+BERT(Bidirectional Encoder Representations from Transformers), Flair+ELMO(Embeddings from Language Model), Flair Casual Transformer, it was found that algorithms, such as Flair, that incorporated contextual word embeddings outperformed the other classifiers. This is because the simple bag-of-words classifiers, such as NLTK, failed to identify the relations that different words had with each other. On the other hand, Flair uses contextual word vectors to analyze the unique relationship between different words (figure 5). This allows it to better understand the text, and therefore generate superior results. As a result, when these algorithms were tested on the Stanford Sentiment Treebank (SST) to rate sentiment on a scale of 1-5, it was found that the Flair algorithm performed the best with an F1 score of 49.90% (Akbik et al, Rao et al)(figure 6). So, the short-post algorithm was built using the Flair sentiment analysis library. The long-article algorithm is essentially the short-post algorithm but averaged across all relevant body paragraphs in an article.

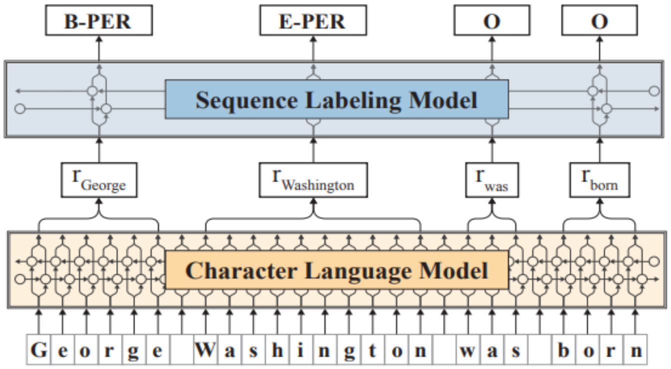


Figure 5: Diagram explaining how Flair contextual string embeddings allow it to learn relations on a character level and word level.

These umbrella algorithms were further optimized for each specific social network. For example, the LinkedIn algorithm analyzed the author profile of a LinkedIn post to eliminate self-reporting. This is because, oftentimes, executives at companies will only discuss their positive initiatives and goals. However, this can dilute other unbiased observations, thus construing results. Additionally, for the Twitter and LinkedIn algorithms, if a link address was found within the text, then the algorithm would analyze that article as well. These added measures help to further boost accuracy.

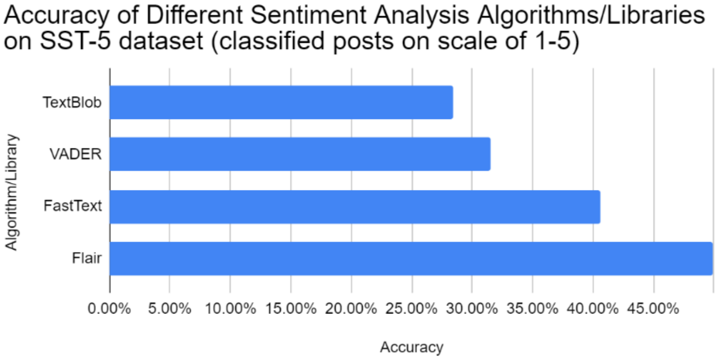


Figure 6: Comparison of accuracy of different sentiment analysis algorithms on SST-5 database

Initially, the analysis algorithm was very slow since it would take Flair 3-4 seconds to analyze one post. So, a variation called “Flair sentiment-fast” was installed. This allowed Flair to analyze multiple posts simultaneously, which significantly cut down on analysis time. It took the program 2 days to analyze all the data.

Once all raw data was scored, the scores were then averaged into one cohesive spreadsheet. Mean imputing was used to fill in any missing sub-score data. Additionally, a publicly licensed ESG rating scraper on GitHub retrieved S&P Global ESG scores for all companies that were going to be analyzed (Shweta-29). This spreadsheet served as the features I trained my machine learning algorithm with.

Machine Learning Algorithms

After compiling the data, different machine learning models were tested. The goal of these models was to predict an ESG score from 0-100. Most of these supervised learning models were lightweight regression algorithms that can learn complex patterns with limited data. Some of these algorithms include:

- Random Forest Regression: It operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.
- Support Vector Regression: identifies the best fit line within a threshold of values
- K-Nearest Neighbors Regression: predicts a value based on the average value of its neighboring data points
- XGBoost (Extreme Gradient Boosting) Regression: it uses gradient boosting by combining the estimates/predictions of simpler regression trees

These regression algorithms were trained using 20 features (19 keyword averages + 1 Glassdoor Meta score). They were calibrated to public S&P Global ESG ratings to ensure they did not diverge much from pre-existing solutions. Optimization techniques such as SMOTE data augmentation were used to further bolster accuracy. SMOTE data augmentation works by selecting two random neighboring points and drawing a new sample randomly across the line (figure 7). SMOTE is commonly used to help deal with limited data. Additionally, regularization was used to prevent overfitting. Overfitting is essentially when an algorithm aligns so closely to the training data that it fails to make generalizations that also apply to the testing data

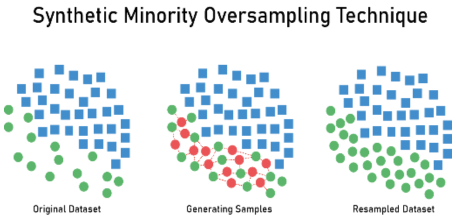


Figure 7: Diagram explaining Synthetic Minority Over-sampling (SMOTE) for data augmentation

In order to create the algorithm, ~375 companies were used as training data, while ~95 companies were used for testing data. This 80%-20% train-test split ensures accurate results.

Results

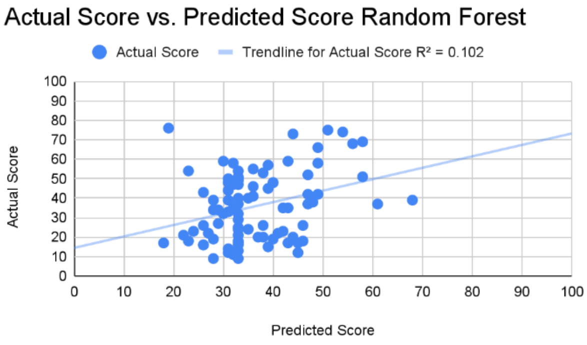


Figure 8: Random Forest model predictions v actual scores

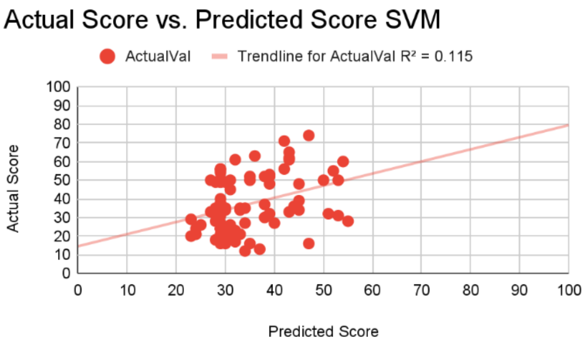


Figure 9: Support Vector Regression predictions v actual scores

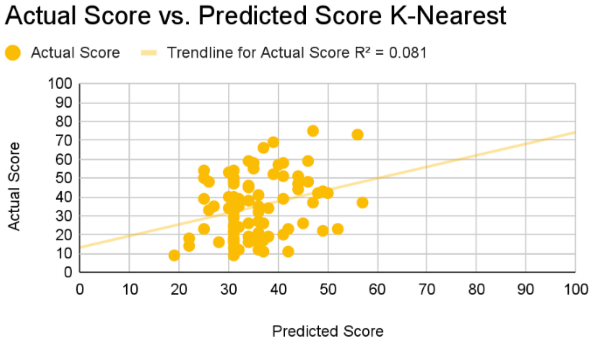


Figure 11: K-Nearest Neighbor model predictions v actual scores

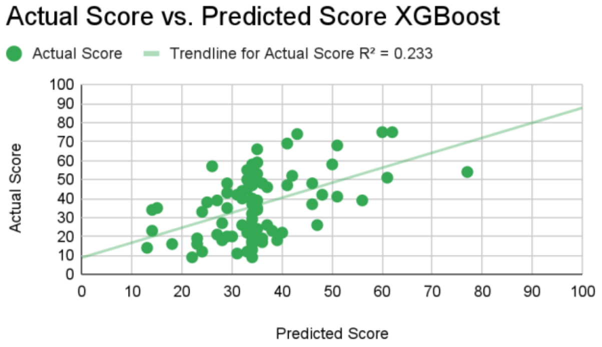


Figure 10: XGBoost model predictions v actual scores

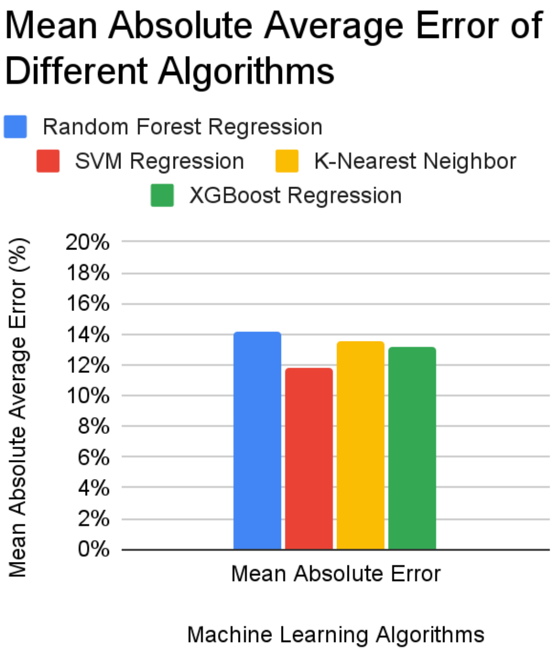


Figure 13: Mean Absolute Average Error of different machine learning algorithms against S&P Global ESG scores

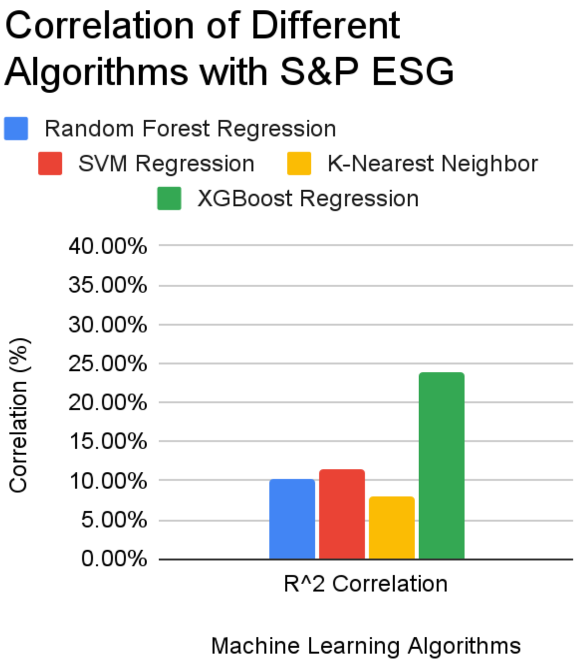


Figure 12: R^2 correlation of different machine learning algorithms

The XGBoost Regressor displayed the strongest overall results. The XGBoost had the strongest correlation with current S&P Global ESG scores with a statistically significant correlation coefficient of 0.23 (Figure 12) and a mean absolute average error (MAAE) of 13.1%. This means that the algorithm has a P-value of 0.022 (<0.05), showing that it is well-calibrated to existing ESG solutions. On the other hand, while the other models have similar MAAE, they also have lower correlation coefficients that do not prove to be statistically significant (Figure 12). For example, Random Forest Regression algorithm had a correlation of 0.102 and MAAE of 14.2%, which results in a P-value of 0.32 (Figure 9). The Support Vector Regressor had a correlation of 0.115 and MAAE of 11.9%, which results in a P-value of 0.26 (Figure 10). Finally, the K-Nearest Neighbors algorithm had a correlation of 0.081 and a MAAE of 13.5%, which is a P-value of 0.43 (Figure 11). However, all of the algorithms had a similar MAAE that fell between 12%-14%, with the SVR algorithm having the lowest at 11.9% (Figure 13). All the algorithms surpassed the MAAE criteria of 20%.

These differences in correlation were seen because most algorithms had their predictions cluster between the 20-60 range. The XGBoost algorithm was the only model that could accurately generalize data over longer periods. Overall, the XGBoost algorithm was the only one that could successfully meet both the correlation and MAAE criteria.

Discussion

The proposed algorithm was able to exceed the criteria, thus highlighting that it is well-calibrated to pre-existing solutions. However, unlike the current ESG raters who determine ESG using voluntarily disclosed sustainability reports and subjective judgments, the proposed algorithm's data-driven approach provides an unparalleled degree of precision and systemization. Additionally, the use of social networks as opposed to self-reported filings allows for a more holistic and balanced evaluation.

While the algorithm has met the set after criteria, there is more room for improvement that can be tackled in future research. Some of this might include:

- **Gathering more data:** This can be by analyzing more companies beyond the S&P 500, collecting more keywords and ESG sub-topics, or by gathering data across more social networks
- **Improving ESG relevancy classification:** While the current filtration method does get rid of most irrelevant data, some bad data still gets through. So, to solve this, a new algorithm can be trained to identify related bodies of text using TF-IDF vectorization. The algorithm can be trained on the data that has already been collected
- **Optimizing long-article/short-post NLP algorithms:** While Flair does provide good results, a significant number of articles seem to be misclassified, which might be a source of error for the algorithm. Further optimizing the long-article and short-post algorithms by creating a sentiment analysis algorithm specifically tailored to ESG classification can help fix this. This can be done by either creating a custom ESG lexicon or training a novel NLP algorithm against classified ESG data.
- **Testing more Machine-learning/Deep-Learning algorithms:** In this project, I only tested out four machine learning algorithms. However, by implementing deep learning algorithms, the accuracy of the algorithm can be further improved
- **Evaluating Post Credibility:** While small amounts of non-credible information would not significantly alter results, evaluating this can serve as an added safeguard

Conclusion

The use of the proposed ESG analysis algorithm can standardize ESG evaluation for all companies. This is because it limits bias by systematically incorporating social network analysis for more balanced results. Additionally, the model can help evaluate the sustainability of smaller companies that do not have analyst coverage. This will help nearly every public company receive ESG ratings in an automated way, which can help socially responsible firms broaden their impact. By solving the current divergence in ESG, ESG can be more readily used, and this will help influence companies to be more responsible/sustainable. This will help rewire over $6.1 trillion worth of capital to more sustainable and ethical initiatives (Reuters).

Acknowledgments

Thank you Dr. Peter Gloor your assistance and guidance throughout the duration of this project. Additionally, thank you Mrs. Catherine Piscitelli for you unwavering support, advice, and countless opportunities

Bibliography

Akhik, Blythe, and Vollgraf. “Contextual String Embeddings for Sequence Labeling.” Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649 Santa Fe, New Mexico, USA, August 20-26, 2018

Berg, Florian, et al. “Aggregate Confusion: The Divergence of ESG Ratings.” *SSRN Electronic Journal*, 2019, doi:10.2139/ssrn.3438533.

Drempetic, Samuel, et al. “The Influence of Firm Size on the ESG Score: Corporate Sustainability Ratings Under Review.” Journal of Business Ethics, vol. 167, no. 2, 2019, pp. 333–360., doi:10.1007/s10551-019-04164-1

“ESG Evaluation.” *S&P Global Ratings*, www.spglobal.com/ratings/en/products-benefits/products/esg-evaluation#:~:text=S&P Global Ratings ESG Evaluation&text=The methodology is founded on,, past, present and future.

“ESG Investing: ESG Ratings.” *MSCI*, www.msci.com/our-solutions/esg-investing/esg-ratings.

Gloor, Peter A., et al. “Web Science 2.0: Identifying Trends through Semantic Social Network Analysis.” *2009 International Conference on Computational Science and Engineering*, 2009, doi:10.1109/cse.2009.186.

Kotsantonis, Sakis, and George Serafeim. “Four Things No One Will Tell You About ESG Data.” Journal of Applied Corporate Finance, vol. 31, no. 2, 2019, pp. 50–58., doi:10.1111/jacf.12346

“New Report Shows Just 100 Companies Are Source of over 70% of Emissions.” CDP, www.cdp.net/en/articles/media/new-report-shows-just-100-companies-are-source-of-over-70-of-emissions.

Person, and Simon Jessop Ross Kerber. “Analysis: How 2021 Became the Year of ESG Investing.” Reuters, Thomson Reuters, 23 Dec. 2021, www.reuters.com/markets/us/how-2021-became-year-esg-investing-2021-12-23/.

Rao, Prashanth. “Fine-Grained Sentiment Analysis in Python (Part 1).” *Medium*, Towards Data Science, 9 Sept. 2019, towardsdatascience.com/fine-grained-sentiment-analysis-in-python-part-1-2697bb111ed4.

Stackpole, Beth. “Why Sustainable Business Needs Better ESG Ratings.” MIT Sloan, 6 Dec. 2021, mitsloan.mit.edu/ideas-made-to-matter/why-sustainable-business-needs-better-esg-ratings.

shweta-29. “Shweta-29/Companies_ESG_Scraper: This Repository Includes a Tool to Extract Companies' ESG Ratings &amp; Financial Metrics and Load Them on SQL.” GitHub, github.com/shweta-29/Companies_ESG_Scraper.

Global Compact-Accenture CEO Study on Sustainability.” Accenture, www.accenture.com/us-en/insights/sustainability/ungc.