Leveraging Multimodal Large Language Models and Natural Language Processing Techniques for Comprehensive ESG Risk Score Prediction

Abhiram Nandiraju[1] and Siddha Kanthi[2]
[1]*Frisco High School, Frisco, U.S.A.*
[2]*Reedy High School, Frisco, U.S.A.*

Keywords: Natural Language Processing, ESG Risk Assessment, S&P 500, Corporate Sustainability, Financial Decision Making.

Abstract: Companies are subject to stringent expectations in terms of social responsibility, particularly in managing risks associated with their environmental, social, and governance (ESG) practices. These practices are evaluated using ESG risk scores. Traditionally, ESG risk scores are generated by firms like Sustainalytics and MSCI, which primarily focus on larger corporations. Consequently, entities investing in smaller companies, such as venture capital firms, private equity firms, and individual investors, face a challenging and resource-intensive process for initial risk assessment. However, our research has uncovered a novel approach through the application of machine learning techniques and the use of multimodal large language models based on publicly released company reports. This approach enables the prediction of ESG risk scores with an accuracy of 68.09%, offering a viable tool for preliminary analysis. Significantly, this research introduces a pioneering framework that utilizes a new architecture for analyzing ESG practices, transforming the traditional assessment process for both large and small companies alike. Our research shows high accuracy in predicting risk assessments and simplifies the evaluation process. Nonetheless, there is potential for enhancing this accuracy through further refinement of the models, improvements in data extraction, and continued exploration of additional modeling techniques.

1 INTRODUCTION

Environmental, Social, and Governance (ESG) criteria have become increasingly pivotal in assessing a company's impact on environmental, societal, and corporate sustainability. Alongside the rise of ESG is the transformative field of artificial intelligence (AI), particularly the development of chatbots, which have revolutionized numerous sectors by automating interactions and processing large volumes of data efficiently. Their role in business and finance is no exception, offering new avenues for data analysis and customer engagement. One of the most significant developments in the domain of ESG assessment has been the introduction of ESG Risk Scores, which provide a quantitative measure of a company's exposure to and management of ESG-related risks. However, the current process of calculating these scores faces several challenges: a limitation to larger companies due to resource constraints, a lack of standardization across different scoring systems, and a time-intensive and expensive evaluation process. Additionally, the calculation of these scores has not fully embraced automation, making the process less efficient than it could be. This paper seeks to address these limitations by proposing an automated, machine learning, and Chat-GPT 4-based approach to predict ESG risk scores using publicly available company reports, such as annual reports, from S&P 500 companies. Through this, we aim to extend the applicability of ESG risk assessment to a broader range of companies, including smaller firms often overlooked in current methodologies. By leveraging AI and machine learning, there is potential to not only democratize ESG risk assessment but also enhance its accuracy and efficiency, paving the way for more inclusive and sustainable corporate practices. It is important to note the subjectivity and inconsistencies in ESG scoring, which could hinder the accuracy of traditional AI models trained on ESG data, as seen with different firms providing vast analyses of the same data and information on the same company. The remainder of the paper is structured as follows: Section 2 presents a review of the literature, Section 3 outlines the methodology

Nandiraju, A. and Kanthi, S.
Leveraging Multimodal Large Language Models and Natural Language Processing Techniques for Comprehensive ESG Risk Score Prediction.
DOI: 10.5220/0012725700003717

In *Proceedings of the 6th International Conference on Finance, Economics, Management and IT Business (FEMIB 2024)*, pages 69-78
ISBN: 978-989-758-695-8; ISSN: 2184-5891

we used, Section 4 presents our results, Section 5 discusses the implications and limitations of our findings, and Section 6 concludes our research paper.

2 LITERATURE REVIEW

2.1 Machine Learning Techniques in Finance

Machine Learning's ability to "learn" and then predict based on data especially suits the financial industry which is all about strategic use of data to make predictions that will beat the competition (Rundo et al., 2019). Thus, it is not outlandish to visualize a future in finance where firms are competing to create models with the highest accuracy and lowest loss to outplay their competitors. Previously, some machine learning applications for finance included: algorithmic trading (Hansen, 2020), credit scoring (Dastile et al., 2020), fraud detection (Dornadula & Geetha, 2019), regulatory compliance (Bauguess, 2017), and portfolio management (Aithal et al., 2023).

2.2 Machine Learning Techniques for ESG Risk Assessment

Similar to the rest of the financial industry, machine learning has been utilized for ESG Risk Assessments as well. What's unique about machine learning is that it gives a degree of creativity to the user as to which model they specifically want to use that will be the most effective given their goal and the data available. In our previous review of ESG Risk score predicting research, the most common types of machine learning algorithms were random forest classifiers (Chen & Liu, 2021; Chowdhury et al., 2023; D'Amato et al., 2021, 2022; D'Amato et al., 2023; Del Vitto et al., 2023; Teoh et al., 2019), linear regression (D'Amato et al., 2023; Del Vitto et al., 2023; Dwivedi et al., 2023; Zhang, 2023), and LSTMs (Teoh et al., 2019). Each model provides a different approach to analyzing the data and can yield different accuracies and losses. Furthermore, to determine our model, we must first determine our data.

2.3 Natural Language Processing (NLP) in ESG Risk Assessment

Natural Language Processing is a machine-learning technique that can derive conclusions from a given piece of text in the same way a human analyzes texts. Historically, natural language processing has been utilized to analyze tweets on Twitter and associate them with a movement in the stock market depending on the content of the tweet. In essence, natural language processing is used to take unstructured data, such as public company reports, and make sense of it. Throughout ESG Risk Score prediction research, a vast majority focuses on the analysis of financial statements of companies and trying to derive an ESG Risk Score from the profitability of the company.

Research papers exploring this intersection have been published, yet none focus on the intersection of NLP & ESG in the United States while utilizing multimodal large language models. For example, Zhang (Zhang, 2023) explores the use of NLP on 100 Chinese ESG reports to predict ESG Market Risk. Similarly, Dwivedi et al. (Dwivedi et al., 2023) utilized a sample of 90 companies' publicly released reports, which were included in the National Stock Exchange (NSE), based in India. However, neither utilized American-based companies, such as those listed on the S&P 500. Uniquely, D'Amato et al. (D'Amato et al., 2021) assessed a company's ESG Risk profile utilizing their fundamental ratios found on the balance sheet and other financial statements.

2.4 Data Source Considerations and Methodology

Considering the preceding focus on financial statements and profitability to predict ESG Risk Scores, we decided to explore the relationship between publicly released reports of a company and its ESG Risk Scores. Since we are analyzing reports, which are predominantly text, we will have to use some sort of Natural Language Processing technology. Although we considered other forms of text-based data, such as specific ESG-focused reports, we ultimately decided against them because of the lack of data on the internet, which would have resulted in a poorly trained model.

Dwivedi et al. (Dwivedi et al., 2023) constructed a study with data from the NIFTY100 ESG Index, comprising 84 companies, which was utilized, with ESG risk scores obtained from S&P Global and other corporate attributes from Moody's Orbis. The study applied machine learning techniques, including Gradient Boosting, to develop an ESG risk score model using a dataset with 47 corporate attributes, focusing on analyzing the impact of these attributes on ESG risk scores.

Krappel et al.'s (Krappel et al., 2021) study used a dataset comprising 7413 companies from Refinitiv Eikon, spanning from 2002 to 2019. It included fundamental financial data and ESG ratings, focusing on how a company's fundamental data over time reflects in its ESG ratings. The analysis incorporated a broad range of financial and non-financial data to predict ESG performance.

D'Amato et al.'s (D'Amato et al., 2022) study analyzed the STOXX Europe 600 Index constituents. This study collected ESG risk scores and balance sheet information for 401 companies from Thomson Reuters Refinitiv ESG. The research aimed to understand the relationship between various balance sheet items and ESG risk scores, using a dataset that also detailed the ESG performance across different industry sectors.

T.-T. et al's (Teoh et al., 2019) research utilized Thomson Reuters ESG Scores to assess the CSR efforts of companies. It combined ESG risk scores with financial performance indicators like ROE, analyzing the year-on-year changes in these metrics. Several machine learning models, including SVM, Random Forest, and LSTM, were developed to classify the changes in ESG risk scores and their correlation with financial performance.

Chowdhury et al.'s (Chowdhury et al., 2023) study aims to develop a machine learning-based ESG rating prediction model using firm-specific and macroeconomic predictors, involving steps like feature selection, data cleaning, and model validation. Data sources include Thomson Reuters Datastream and the World Bank, with variables like ESG risk score, company size, and macroeconomic indicators. Various machine learning models, including Neural Networks and Random Forests, were evaluated for predicting ESG ratings, employing cross-validation and ROC curve analysis for model selection.

2.5 Practical Applications and Implications

Our findings will pave the way for investors to derive holistic findings for companies regardless of whether or not they have publicly released ESG risk scores. This approach aligns with the growing trend in finance to leverage machine learning for more accurate predictions and novel measure construction (Ahmed et al., 2022). Not only will this help investors, such as those in corporate finance evaluate smaller companies, whose actions may not be as magnified as larger companies, but it will also hold every single company accountable for their actions in the environmental, social, and governance spaces. Furthermore, the application of machine learning in developing superior measures and reducing prediction errors in finance underscores the potential of these technologies in enhancing corporate governance and accountability (Chen & Liu, 2021).

3 METHODOLOGY

3.1 Dataset Selection

Since our study is targeted to S&P 500 company information, we acquired an open-source Kaggle dataset with S&P 500 companies' ESG metrics (Dugar, n.d.). To accurately account for differences in model input, we kept the following attributes in our data:

- **Symbol:** The unique stock symbol associated with the company.
- **Name:** The official name of the company.
- **Total ESG Risk Score:** An aggregate score evaluating the company's overall ESG risk.
- **Environment Risk Score:** A score indicating the company's environmental sustainability and impact.
- **Social Risk Score:** A score assessing the company's societal and employee-related practices.
- **Governance Risk Score:** A score reflecting the quality of the company's governance structure.
- **ESG Risk Level:** A categorical indication of the company's ESG risk level.

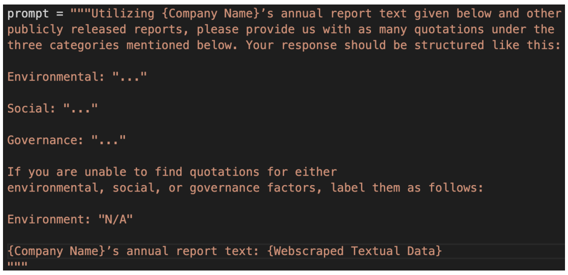
The next step was to add 3 columns to our dataset: "Environmental Description", "Social Description", and "Governance Description". These columns would eventually include direct quotes regarding environmental, social, and governance factors from each company's 2022 annual report and any other publicly released reports. The steps to fill in these columns required a 3 step approach. Firstly, to access each company's annual report we web-scraped text from the URL structured as follows:

https://www.annualreports.com/HostedData/AnnualReports/PDF/{Stock_Exchange}_{Company_Ticker_Symbol}_2022.pdf}

We iterated through a 2D array structured as follows to attribute the Stock Exchange to the Company Ticker Symbol:

```
[(Stock Exchange, Company Ticker Symbol),
 (Stock Exchange, Company Ticker Symbol),
 ...,
 (Stock Exchange, Company Ticker Symbol)]
```

Secondly, after iterating through each company's pdf file, we extracted the respective report's text and



```
prompt = """Utilizing {Company Name}’s annual report text given below and other
publicly released reports, please provide us with as many quotations under the
three categories mentioned below. Your response should be structured like this:

Environmental: "..."

Social: "..."

Governance: "..."

If you are unable to find quotations for either
environmental, social, or governance factors, label them as follows:

Environment: "N/A"

{Company Name}’s annual report text: {Webscraped Textual Data}
"""
```

Figure 1: Prompt Given To ChatGPT-4.

sent a request to OpenAI’s ChatGPT API utilizing the “gpt-4-1106-preview” model. We structured the request input as shown in Figure 1.

Lastly, after formatting the API’s output data into “Environmental Description”, “Social Description”, and “Governance Description” categories, we added all categories to our dataset under each company.

3.2 Pre-Processing Measures

Before our data was ready to use within our models, we had to pre-process our aggregated dataset. Our pre-processing measures are as follows:

- **Missing Values:** In some cases, we had issues with web scraping content and OpenAI API responses. Therefore, we implemented Python’s “.fillna(”)” method to fill in NaN inputs to values that are recognizable by our model.
- **Stemming:** Stemming is a text normalization technique that involves reducing words to their base or root form, which helps minimize the complexity of the textual data by consolidating different forms of a word into a single representation. For example, the words ”connect”, ”connecting”, ”connected”, and “connection” are all reduced to the stem ”connect”. This is particularly beneficial in our analysis as it decreases the variability within the text data, allowing our machine-learning models to focus on the essence of the content rather than getting bogged down by the nuances of language. We utilized “Porter Stemmer”, a widely-used algorithm for stemming English words. The choice of the algorithm was guided by the balance between aggressive stemming, which might oversimplify the text, and gentle stemming, which retains more of the word’s original form.
- **Training and Testing Split:** In our study, we divided our dataset into two distinct sets: the training set and the testing set. This division is a fundamental practice in machine learning to evaluate the performance of our models. Typically, the training set is larger and used to train the model, allowing it to learn the underlying patterns in the data. The testing set, on the other hand, is used to evaluate the model’s performance on unseen data, ensuring that our model can generalize well to new, unobserved data. We adhered to a training-focused ratio, allocating 90% of the data for training and the remaining 10% for testing. This split was performed randomly to ensure a representative and unbiased distribution of data across both sets. We separated the data in this manner to minimize the effects of model underfitting, wherein the model would not learn the inherent patterns within the training data.

3.3 Model Building

Before delving into the specifics of each model utilized in our study, it is crucial to contextualize the application of sentiment analysis within the scope of our research objectives. Sentiment analysis, a computational technique, aims to identify and categorize opinions within the text to ascertain the writer’s attitude towards a particular subject, product, etc., as positive, negative, or neutral. In the context of our research, we seek to gauge the sentiment of company reports regarding Environmental, Social, and Governance (ESG) factors. Our model-building process incorporates four distinct machine learning models: Linear Regression, Logistic Regression, Long Short-Term Memory (LSTM) networks, and Bidirectional Encoder Representations from Transformers (BERT), each offering unique benefits for predicting the ESG Risk Score.

3.3.1 Linear Regression Model

The Linear Regression model is adopted for its simplicity and efficacy in predicting numerical outcomes. Mathematically, it is expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \qquad (1)$$

where y represents the ESG Risk Score, β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficients of the preprocessed text features $x_1, x_2, \ldots, x_n$, and ε is the error term. This model establishes a linear correlation between the textual data from reports and quantifiable ESG risk scores, facilitating straightforward interpretation and efficient prediction.

3.3.2 Logistic Regression Model

Logistic Regression is leveraged for its proficiency in classifying outcomes, ideal for predicting categorical ESG risk levels (negligible risk, low risk, medium risk, high risk, and severe risk). It employs the logistic function to estimate probabilities that then dictate class membership:

$$p(y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}} \tag{2}$$

where $p(y=1)$ represents the probability of the ESG risk score falling into a specific category (e.g., high risk). This model effectively links pre-processed text features from reports to categorical ESG risk scores.

3.3.3 LSTM (RNN) Model

LSTMs, a special kind of RNN, are adept at processing sequence data, making them particularly suitable for text. The core concept of an LSTM is its ability to maintain a cell state and apply gating mechanisms, which include:

- Forget Gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$
- Input Gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$
- Cell State Update: $\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$
- Output Gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$
- Updated Cell State: $C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$
- Output: $h_t = o_t * \tanh(C_t)$

Here, σ denotes the sigmoid function, W and b represent weights and biases for each gate, respectively, and h_t and C_t are the hidden state and cell state at time t. This sophisticated mechanism enables LSTMs to capture long-term dependencies within textual data, essential for understanding the complex nuances associated with ESG factors.

3.3.4 BERT Model

The Bidirectional Encoder Representations from Transformers (BERT) model represents a paradigm shift in how machines understand textual information. Its architecture is grounded in the transformer model, which relies on attention mechanisms to weigh the significance of different words in a sentence. Formally, the transformer uses self-attention mechanisms, which can be described as follows:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Q, K, and V represent the queries, keys, and values matrices, respectively, and d_k is the dimension of the key vectors. This mechanism allows BERT to consider the context of each word in the entire document bidirectionally, as opposed to previous models that processed text in one direction.

BERT's training comprises two main tasks: Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). The MLM task randomly masks words in a sentence and trains the model to predict these masked words, thereby learning context. The NSP task trains the model to predict whether a sentence logically follows another, enhancing its understanding of sentence relationships.

For fine-tuning BERT on specific datasets, such as those related to ESG factors, the pre-trained BERT model is adapted as follows:

1. The final output layer of BERT is replaced with a new layer, tailored to the specific classification task (e.g., predicting ESG risk scores).
2. The entire model is then trained on the domain-specific dataset, allowing the model to adjust its internal weights to better understand and classify the new data.

This fine-tuning process enables BERT to extract meaningful features from ESG-related text, leveraging its deep contextualized representations to understand the nuances and complexities of natural language. The advantage of using BERT lies in its ability to capture both sentiment and thematic content relevant to ESG factors, providing a nuanced analysis of textual data. This approach significantly enhances our methodology, allowing for a more insightful and accurate derivation of ESG Risk Scores.

3.4 Conclusion

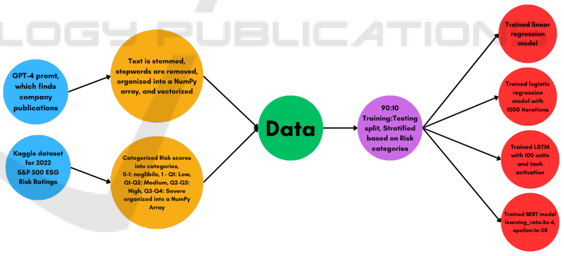

GPT-4 promt, which finds company publications
Kaggle dataset for 2022 S&P 500 ESG Risk Ratings
Text is stemmed, stopwords are removed, organized into a NumPy array, and vectorized
Categorized Risk scores into categories, 0-1: neglibile, 1 - Q1: Low, Q1-Q2: Medium, Q2-Q3: High, Q3-Q4: Severe organized into a NumPy Array
Data
90:10 Training:Testing split, Stratified based on Risk categories
Trained linear regression model
Trained logistic regression model with 1000 iterations
Trained LSTM with 100 units and tanh activation
Trained BERT model learning_rate=2e-6, epsilon=1e-08

Figure 2: Visual Representation of our Methodology Process.

Figure 2 shows a visual representation of our methodology process. The next section will detail the results obtained from applying these methodologies.

4 RESULTS

4.1 Linear Regression Model Results

Considering the nature of traditional linear regression evaluation metrics, we instead decided to interpret our results with a 25 percent tolerance to provide a consistent evaluation approach between models

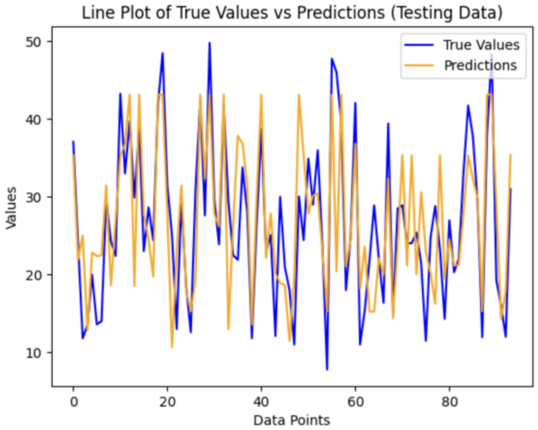


Figure 3: True ESG risk score vs Predicted ESG risk score on Testing Data (Linear Regression).

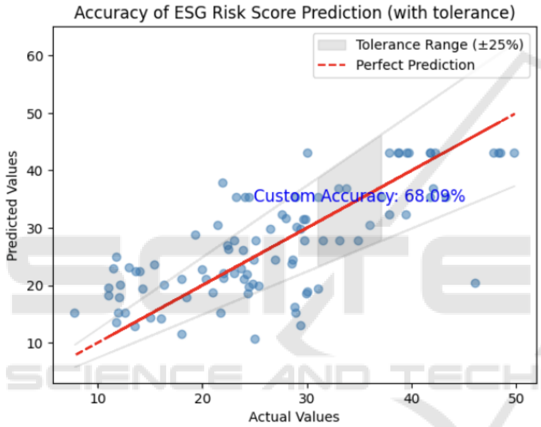


Figure 4: Linear Regression Model Results with ±25% Tolerance.

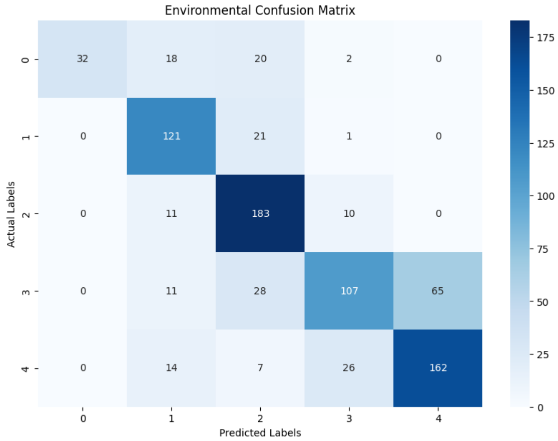


Figure 5: Environmental Model Confusion Matrix.

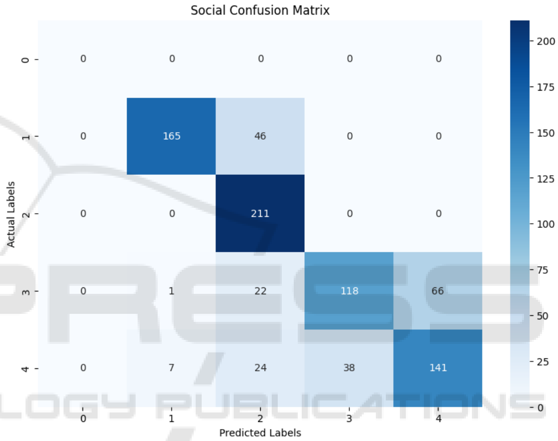


Figure 6: Social Model Confusion Matrix.

(Figure 4). As a result, the Linear Regression model shows a moderately high level of accuracy. This suggests it can relatively effectively predict ESG Risk Scores based on the textual data extracted from company reports. Given the model's simplicity and ease of interpretation, these results are promising, especially for initial assessments or in situations where computational resources are limited. However, the model might not fully capture the complex relationships and nuances inherent in the textual descriptions of ESG factors.

4.2 Logistic Regression Model Results

Figure 5, Figure 6, and Figure 7 are our confusion matrices, which outline the success of the logistic regression model in predicting environmental, social, and governance levels respectively. The varied performance of the Logistic Regression model across different ESG categories (environmental, social, governance) could indicate that the model is more attuned to certain aspects of ESG risk than others. The relatively moderate accuracy in each category, particularly in governance, suggests that the model may struggle with the complexity and variability of language used in ESG reporting. With the similarity in excerpts within the data across ESG risk levels, a logistic regression, which utilized the bag of words NLP technique may struggle.

4.3 LSTM Model Results

The low accuracy of the LSTM model is surprising, given its capability to process sequence data and its effectiveness in natural language processing tasks (Figure 8). This might indicate challenges in the model's configuration or the nature of the data. It could suggest that the LSTM is either overfitting or underfitting the training data or that the sequential

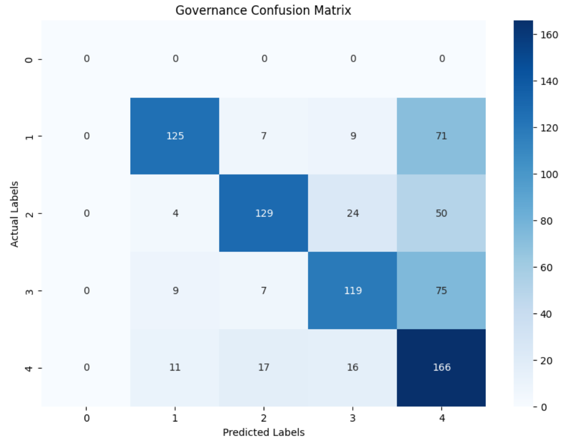


Figure 7: Governance Model Confusion Matrix.

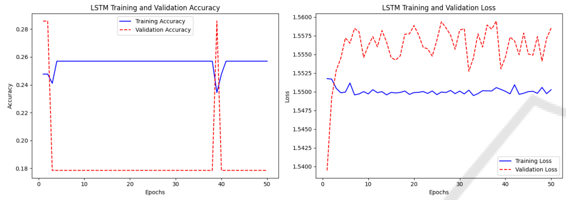


Figure 8: Training & Validation Accuracy Over Epochs and Training & Validation Loss Over Epochs (LSTM Model).

aspects of the text are not as predictive of ESG risk scores as hypothesized. This result might necessitate a review of the model architecture, data preprocessing, or feature selection.

4.4 BERT Model Results

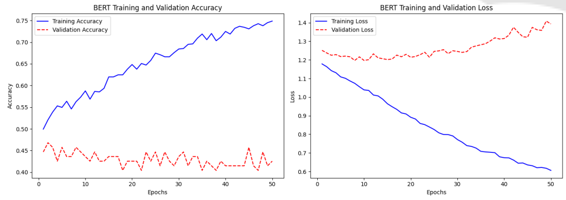


Figure 9: Training & Validation Accuracy Over Epochs and Training & Validation Loss Over Epochs (BERT Model).

While BERT models are proficient in understanding language context, the moderate accuracy in this application suggests that the model may not have been fully optimized for this specific task (Figure 9). The complexity of ESG reporting text and the subtleties of risk assessment might require more fine-tuning of the model, or additional contextual features may be needed to improve its predictive power. This result underscores the challenges of applying advanced NLP models to specialized domains like ESG risk assessment.

4.5 General Results

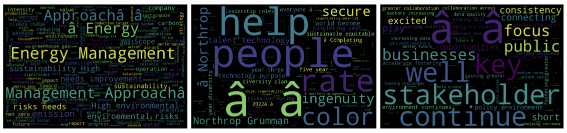


Figure 10: The image on the left is the Environmental Word Cloud. The image in the middle is the Social Word Cloud. The image on the right is the Governance Word Cloud.

Reflecting on the results of our research on ESG risk scores, it becomes evident that each model we employed has its unique strengths and limitations in this task. Interestingly, despite the advanced capabilities of LSTM and BERT models, it is both the Linear and Logistic regression models that stand out with their effectiveness. Achieving an accuracy of 68.09% for Linear Regression and 51.05% for Logistic Regression, this model's simplicity, ease of interpretation, and decent performance make it a surprisingly viable option for predicting ESG Risk Scores. This is particularly significant in scenarios where stakeholders prefer models that are transparent and easy to understand.

5 DISCUSSION AND LIMITATIONS

5.1 Discussion of Data Findings

The lower accuracy rates of the LSTM and BERT models, 25.53% and 46.81% respectively, indicate possible challenges in their configuration and training, or perhaps the inherent complexity of analyzing the nuances in ESG reporting. While powerful in handling sequential and contextual data, these sophisticated models require more in-depth tuning and an enhanced understanding of ESG report subtleties to fully utilize their potential. The varied performance of the Logistic Regression model across different ESG categories also offers valuable insights. It suggests a sensitivity to the specifics of each category, though its overall effectiveness seems to lag behind that of Linear Regression for this particular task. These results collectively underscore the complexity of assessing ESG risks from textual data and the importance of careful model selection and tuning. The success of the Linear Regression model in our study is particularly intriguing, as it suggests that in certain aspects of ESG risk assessment, simpler models can be quite effective, especially for initial screenings or situations where interpretability is crucial. However, there are limitations faced throughout our process.

5.2 Implications for Stakeholders

The upside of our results is undeniable for stakeholders. With a machine learning algorithm that can provide an enterprise risk score on ESG factors efficiently, investors can now conduct a holistic risk analysis of companies that do not already have ESG risk scores/analysis. For a vast majority of companies, especially those with a market cap under $500 million, an ESG risk score is not provided, meaning that our system could provide an analysis within seconds. This form of analysis is especially useful for private equity firms who tend to acquire companies that are not publicly traded, much less assigned an ESG risk score. Considering that ESG initiatives, which comprise ESG risk scores, correlate with market risk and returns as described by Zhang, being able to efficiently analyze ESG risk will help investors differentiate sound companies from unsound companies.

5.3 Data Limitations

Our research, focusing exclusively on S&P 500 companies, as opposed to Teoh who focused on major technological stocks and the NASDAQ index, and Zhang who focused exclusively on Chinese companies, presented both strengths and constraints. While this focus allowed us to work with a consistent and relatively homogenous dataset, it also limited the generalizability of our findings. Expanding our data beyond the S&P 500 could have potentially introduced a more diverse range of ESG practices and reporting standards, reflecting a broader spectrum of corporate behaviors and policies. This expansion could have provided a richer and more nuanced understanding of ESG risk assessments, enabling our models to capture a wider array of ESG factors and their impact. Additionally, including smaller or international firms, which often have different ESG reporting standards and challenges, might have revealed additional insights into the variability and complexity of ESG practices globally. Next, our specific data source, company reports, may have led to poor, monotonous data as companies tend to provide standard responses for certain issues, making it difficult for our models to differentiate companies experiencing high risk from those experiencing low risk. Finally, ChatGPT's response algorithm tends to follow a specific format that may have introduced unintended patterns within our dataset that the models tried to recognize. This may be another reason why our linear and logistic regression models may have performed better than our BERT or LSTM models, as linear regression is more adept at capturing these consistent, systematic patterns in data, while more complex models like BERT might overfit to the nuances in language, missing out on these broader, more uniform trends.

5.4 Enhancing the Process

To enhance the effectiveness of our process, several strategies could be considered. Firstly, our strategy could have focused solely on either annual reports or even sustainability reports as opposed to data from a diverse range of publicly released company reports to enhance and streamline the data retrieval process. Secondly, exploring alternative data sources, like news articles, social media, or consumer reviews, could provide additional context and depth to the ESG assessments. Furthermore, continuously updating the dataset with the latest reports and data would ensure that the models stay relevant and accurate over time. Another aspect to consider is improving data preprocessing techniques, such as more advanced natural language processing methods, to better capture the nuances and subtleties in the textual data. Lastly, as independent researchers, we faced financial constraints that inhibited our data retrieval process and our machine-learning capabilities. Specifically, upgrading the LLM model we used requires more financial flexibility. Our process could incorporate interdisciplinary approaches, such as integrating insights from behavioral economics to understand the impact of corporate governance on ESG performance. This draws inspiration from D'Amato et al.'s exploration of balance sheet items and their correlation with ESG scores, suggesting a nuanced approach to feature selection in our model. Further, Krappel et al.'s work on the temporal dynamics of company fundamentals in reflecting ESG ratings underlines the importance of including longitudinal data analysis in our methodology. This could ensure our model adapts to temporal changes in ESG criteria, much like the dynamic models suggested by T.-T. et al. and Chowdhury et al., who assessed year-on-year changes in ESG risk scores and their correlation with financial performance using various machine learning models.

5.5 Subjectivity in ESG Risk Scores

A limitation in our study, and indeed in the field of ESG risk assessment in general, is the intrinsic subjectivity of ESG risk scores. ESG scoring is an extensive process, often involving qualitative judgments and varying interpretations of what constitutes good environmental, social, and governance practices. This subjectivity can lead to inconsistencies and variability in ESG risk scores, even among similar companies. It also poses a challenge for machine learning models, which rely on consistent and objective data to make accurate predictions. Furthermore, we must take into consideration the uneven distribution of companies within the S&P 500 company portfolio. As shown by Figure 11, in 2022 the Information Technology (IT) sector and the Healthcare sector comprised 25.7% and 15.8% of the S&P 500 companies, respectively. This skew towards IT and Healthcare may raise concerns about the representativeness of the dataset used for machine learning models. Such dominance could lead to models that are inadvertently tailored to the ESG reporting standards, challenges, and practices prevalent in these sectors, potentially overlooking the unique environmental, social, and governance issues pertinent to other industries.

5.6 Ethical Considerations

Critical ethical issues are brought up by using machine learning and natural language processing to predict ESG risk scores. These issues include the need to address biases in data sources and algorithms to guarantee impartial and accurate assessments in all industries. Robust data handling and anonymization protocols are imperative to ensure privacy and protect sensitive information found in analyzed texts. Furthermore, to promote accountability and trust among stakeholders and enable well-informed decision-making based on ESG evaluations, models must remain transparent and interpretable.

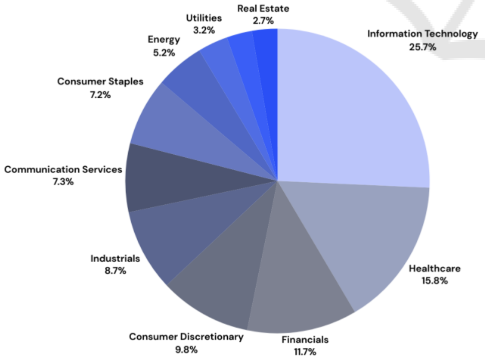


Figure 11: S&P 500 Market Representation by Sector.

6 CONCLUSION AND FUTURE WORKS

6.1 Conclusion

In conclusion, our research explores the use of various machine learning and natural language processing techniques with public company reports to predict their respective ESG risk scores. Specifically, venture capital firms, private equity firms, and relatively modest investors who can not afford the labor and capital-intensive process of doing an in-depth corporate social responsibility analysis on each venture can use our discovery to bridge this gap. Our research introduces a pioneering framework that utilizes a new architecture for analyzing ESG practices, transforming the traditional assessment process for both large and small companies alike.

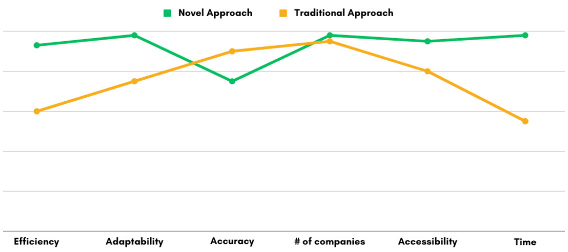


Figure 12: Holistic Representation of our Novel Approach vs Traditional Approaches.

6.2 Future Works

Future research in the area of machine learning models for ESG risk assessment has several promising avenues to pursue. Including companies that are not listed on the S&P 500 in the research is a particularly promising direction. This expansion would improve the findings' inclusivity and suitability for a wider range of businesses, including startups and smaller enterprises, in addition to diversifying the dataset. These businesses frequently face distinct operating constraints and could present particular ESG issues and behaviors, providing a wealth of material for additional research. An important topic for further study is the examination of global ESG risk scores. A company's ESG practices and reporting may be greatly impacted by the differing ESG standards, laws, and cultural viewpoints of various nations and areas. Future research can also offer a more global perspective on ESG risk assessment by incorporating international data, which will help to create a more thorough understanding of global ESG practices and their implications. A more comprehensive understanding of a company's ESG impact could be obtained by incorporating a wider range of data sources, including news articles, social media sentiment, and even regional and political variables. Furthermore, we can also find more relationships in our data and findings by using more machine learning algorithms and techniques like principal component analysis (PCA), support vector machines (SVM), random forests, decision trees, and neural networks.

REFERENCES

Ahmed, S., Alshater, M. M., El Ammari, A., and Hammami, H. (2022). Artificial intelligence and machine learning in finance: A bibliometric review. *Research in International Business and Finance*, 61.

Aithal, P. K., Geetha, M., U, D., Savitha, B., and Menon, P. (2023). Real-time portfolio management system utilizing machine learning techniques. *IEEE Access*, 11:32595–32608.

Bauguess, S. W. (2017). The role of big data, machine learning, and ai in assessing risks: a regulatory perspective. SSRN. Presented at OpRisk North America 2017, New York, NY.

Chen, Q. and Liu, X.-Y. (2021). Quantifying esg alpha using scholar big data: an automated machine learning approach. In *Proceedings of the First ACM International Conference on AI in Finance (ICAIF '20)*, pages 1–8, New York, NY, USA. ACM.

Chowdhury, M. A. F., Abdullah, M., Azad, M. A. K., Sulong, Z., and Islam, M. N. (2023). Environmental social and governance (esg) rating prediction using machine learning approaches. *Annals of Operations Research*.

D'Amato, V., D'Ecclesia, R., and Levantesi, S. (2023). Firms' profitability and esg score: A machine learning approach. *Applied Stochastic Models in Business and Industry*, pages 1–19.

Dastile, X., Celik, T., and Potsane, M. (2020). Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing Journal*, 91.

Del Vitto, A., Marazzina, D., and Stocco, D. (2023). Esg ratings explainability through machine learning techniques. *Annals of Operations Research*.

Dornadula, V. N. and Geetha, S. (2019). Credit card fraud detection using machine learning algorithms. In *Procedia Computer Science*, volume 165, pages 631–641. Elsevier.

Dugar, P. S and p 500 esg risk ratings. Kaggle.

Dwivedi, D., Batra, S., and Pathak, Y. K. (2023). A machine learning based approach to identify key drivers for improving corporate's esg ratings. *Journal of Law and Sustainable Development*, 11(1).

D'Amato, V., D'Ecclesia, R., and Levantesi, S. (2021). Fundamental ratios as predictors of esg scores: a machine learning approach. *Decisions in Economics and Finance*, 44:1087–1110.

D'Amato, V., D'Ecclesia, R., and Levantesi, S. (2022). Esg risk score prediction through random forest algorithm. *Computational Management Science*, 19:347–373.

Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society*, 7(1).

Krappel, T., Bogun, A., and Borth, D. (2021). Heterogeneous ensemble for esg ratings prediction. arXiv. CoRR.

Rundo, F., Trenta, F., di Stallo, A. L., and Battiato, S. (2019). Machine learning for quantitative finance applications: A survey. *Applied Sciences*, 9(24):5574.

Teoh, T.-T., Heng, Q. K. J. J., Chia Shie, J. M., Liaw, S. W., Yang, M., and Nguwi, Y.-Y. (2019). Machine learning-based corporate social responsibility prediction. In *Proc. IEEE Conf. on Cybernetics and Intelligent Systems (CIS) and Robotics, Automation and Mechatronics (RAM)*, pages 501–505. IEEE.

Zhang, Y. (2023). Esg-based market risk prediction and management using machine learning and natural language processing. Bachelor's thesis, Business and Economics Honors Program, NYU Shanghai, May 2023.