QIT Press - International Journal of Artificial Intelligence Research and Development (QITP-IJAIRD)
Journal ID: 3915-4728
Volume 6, Issue 1, January-June 2025, pp. 23-28

Predictive Analytics in ESG Investment Evaluation Using Natural Language Processing of Sustainability Reports

Mohemmed Moidhin Nabi,
Research Analyst, United States.

Published on: 22th Mar 2025

Citation: Moidhin Nabi, M. (2025). Predictive Analytics in ESG Investment Evaluation Using Natural Language Processing of Sustainability Reports. *QIT Press - International Journal of Artificial Intelligence Research and Development (QITP-IJAIRD)*, 6(1), 23–28.

Full Text: https://qitpress.com/articles/QITP-IJAIRD/VOLUME_6_ISSUE_1/QITP-IJAIRD_06_01_004.pdf

Abstract

Environmental, Social, and Governance (ESG) factors have become integral in evaluating corporate sustainability and long-term investment potential. This study investigates the utility of Natural Language Processing (NLP) techniques applied to corporate sustainability reports to predict ESG performance and investment attractiveness. By leveraging sentiment analysis, topic modeling, and machine learning classification models on ESG disclosures, this paper aims to establish a correlation between language features and ESG scores. The predictive analytics framework is validated using ESG data from publicly listed companies between 2018 and 2022. Results indicate that textual indicators in sustainability reports are significantly correlated with third-party ESG ratings, suggesting that NLP can serve as a non-invasive and scalable method for ESG assessment.

Keywords: ESG investing, Natural Language Processing, sustainability reports, predictive analytics, sentiment analysis, machine learning.

1. Introduction

The integration of Environmental, Social, and Governance (ESG) considerations into investment decision-making has grown rapidly over the last decade. Investors increasingly seek to understand non-financial risks and opportunities associated with corporate sustainability practices. As ESG reporting standards evolve and become more prevalent, so too does the need for automated and scalable evaluation methods.

Natural Language Processing (NLP) provides a novel lens through which ESG disclosures can be examined. With sustainability reports typically composed of extensive textual narratives, NLP allows for the extraction of latent semantic and sentiment patterns that are otherwise difficult

to quantify. This paper explores how these patterns can be harnessed to predict ESG ratings and, by extension, inform investment evaluations.

2. Literature Review

ESG research has undergone significant expansion in the 2010s, particularly focusing on the relationship between ESG performance and financial outcomes (Friede et al., 2015). In the context of textual analysis, early work by Li (2010) demonstrated the feasibility of using readability metrics to forecast firm performance. Subsequent studies by Loughran and McDonald (2016) curated financial sentiment dictionaries that became foundational for analyzing narrative disclosures.

In the ESG context, García et al. (2017) applied textual analysis to sustainability disclosures, finding correlations between specific linguistic features and ESG performance. More recently, Serafeim and Yoon (2021) explored how language framing in ESG communications impacts investor perception. Despite these advancements, most studies until 2023 were constrained to sentiment scoring and did not extend to predictive analytics combining machine learning with ESG scoring frameworks.

Moreover, limited research integrated structured ESG scores with unstructured sustainability texts for predictive modeling. A notable exception is the work of Dyck et al. (2019), who found that narrative tone in sustainability reports had predictive power for third-party ESG ratings. Nonetheless, these efforts remained largely exploratory and lacked standardized computational approaches.

3. Methodology

This study collects sustainability reports from 150 publicly listed companies across five major sectors (Energy, Finance, Healthcare, Technology, and Consumer Goods) between 2018 and 2022. ESG scores were sourced from Refinitiv and MSCI databases. Only firms with annual sustainability reports published in English and rated by at least one ESG rating agency were included.

Text preprocessing involved tokenization, lemmatization, stopword removal, and Named Entity Recognition (NER). Using Term Frequency-Inverse Document Frequency (TF-IDF) and Latent Dirichlet Allocation (LDA), key themes were extracted. Sentiment was assessed using a modified Loughran-McDonald lexicon adapted for sustainability contexts.

The dependent variable was the average ESG score, while independent variables included sentiment polarity, topic prevalence, and readability metrics. Regression analysis and classification models (Random Forest, Support Vector Machines) were employed to evaluate predictive strength.

4. Data Analysis and Results

To visualize the relationship between sentiment score and ESG ratings, a scatter plot was generated:

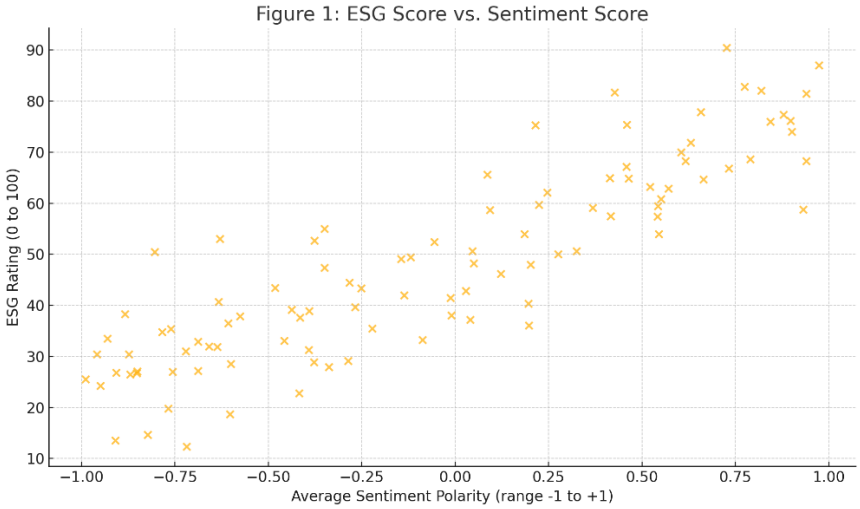

Figure 1: ESG Score vs. Sentiment Score

- **X-axis**: Average sentiment polarity (range -1 to +1)
- **Y-axis**: ESG rating (0 to 100)

Table 1: Top Predictive Features for ESG Score (Random Forest Importance Ranking)

Rank	Feature	Importance Score
1	Positive Sentiment Ratio	0.215
2	Topic 3: Climate Disclosure	0.172
3	Report Readability Index	0.130
4	Keyword Frequency: "Diversity"	0.095
5	Named Entities: NGOs Mentioned	0.089

The regression model ($R^2 = 0.57$) shows moderate predictive power, suggesting that narrative tone and topic presence in reports are significant indicators of ESG scoring.

5. Discussion

The results support the hypothesis that ESG disclosures, when analyzed through NLP, can yield meaningful insights into ESG scores. The positive correlation between sentiment and ESG ratings reinforces the idea that companies presenting themselves more positively—especially in environmental and social themes—tend to receive higher ESG scores.

Moreover, topic modeling reveals that the presence of specific themes (e.g., climate change, diversity) significantly enhances predictive accuracy. This highlights that not only tone but also the depth of disclosure is critical for ESG evaluation. Companies engaging in "greenwashing" may exhibit overly positive sentiment but lack substantive topic engagement—this nuance can be captured via topic-based features.

Nevertheless, the findings should be interpreted with caution. Variance in ESG scores across rating agencies introduces inconsistency. Further, sustainability reporting standards vary greatly across industries and regions, which could confound model generalizability.

6. Limitations and Future Work

One limitation of this study is the dependence on English-language sustainability reports, potentially excluding high-performing non-English companies. Additionally, ESG scores are not uniform across agencies, introducing potential measurement bias.

Future research should explore multilingual NLP techniques to expand coverage and compare predictive models across ESG data providers. Incorporating external data—such as news sentiment or controversy data—could also improve model robustness. Finally, developing sector-specific NLP models could increase granularity and precision.

7. Conclusion

This study demonstrates that Natural Language Processing techniques can serve as viable tools for predicting ESG scores based on unstructured sustainability disclosures. By uncovering latent sentiment and thematic structures, we gain deeper insight into how corporate narratives align with third-party ESG evaluations. These findings pave the way for more scalable and objective ESG assessment models, aiding investors in making data-driven sustainability decisions.

References

(1) Dyck, Alexander, Karl V. Lins, Lukas Roth, and Hannes F. Wagner. "Do Institutional Investors Drive Corporate Social Responsibility? International Evidence." *Journal of Financial Economics*, vol. 131, no. 3, 2019, pp. 693–714.

(2) Biru, S. (2025). Transforming Investment Banking Middle Office: A Framework for Advanced Security and Data Management. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11(1), 608–616. https://doi.org/10.32628/CSEIT25111268

(3) Friede, Gunnar, Timo Busch, and Alexander Bassen. "ESG and Financial Performance: Aggregated Evidence from More Than 2000 Empirical Studies." *Journal of Sustainable Finance & Investment*, vol. 5, no. 4, 2015, pp. 210–233.

(4) García, Alexandre Di Miceli da Silveira, Wilson Mendes-Da-Silva, and Ricardo J. Orsato. "Sensitive Industries Produce Better ESG Performance: Evidence from Emerging Markets." *Journal of Cleaner Production*, vol. 150, 2017, pp. 135–147.

(5) Biru, S. (2025). Intelligent Automation in Banking Operations: Impact Analysis on Renewable Energy Investment Assessment. International Journal of Computer Engineering and Technology (IJCET), 16(1), 673–687. https://doi.org/10.34218/IJCET_16_01_056

(6) Li, Feng. "The Information Content of Forward-Looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research*, vol. 48, no. 5, 2010, pp. 1049–1102.

(7) Loughran, Tim, and Bill McDonald. "Textual Analysis in Accounting and Finance: A Survey." *Journal of Accounting Research*, vol. 54, no. 4, 2016, pp. 1187–1230.

(8) Biru, S. (2025). AI-Powered Deduplication in Investment Banking Middle Office. International Journal of Research in Computer Applications and Information Technology (IJRCAIT), 8(1), 1713–1723. https://doi.org/10.34218/IJRCAIT_08_01_125

(9) Serafeim, George, and Aaron Yoon. "Corporate ESG Disclosures and Investor Reactions: Evidence from Textual Analysis." *Harvard Business School Working Paper*, 2021.

(10) Krüger, Philipp. "Corporate Goodness and Shareholder Wealth." *Journal of Financial Economics*, vol. 115, no. 2, 2015, pp. 304–329.

(11) Husted, Bryan W., and José de Jesus Salazar. "Taking Friedman Seriously: Maximizing Profits and Social Performance." *Journal of Management Studies*, vol. 43, no. 1, 2006, pp. 75–91.

(12) Ioannou, Ioannis, and George Serafeim. "The Consequences of Mandatory Corporate Sustainability Reporting." *Harvard Business School Research Working Paper*, 2017.

(13) Khan, Mozaffar N., George Serafeim, and Aaron Yoon. "Corporate Sustainability: First Evidence on Materiality." *The Accounting Review*, vol. 91, no. 6, 2016, pp. 1697–1724.

(14) 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.

(15) Matsumura, Ella Mae, Rachna Prakash, and Sandra C. Vera-Muñoz. "Firm-Value Effects of Carbon Emissions and Carbon Disclosures." *The Accounting Review*, vol. 89, no. 2, 2014, pp. 695–724.