A Hybrid NLP - Ensemble Learning Architecture for Automated ESG Report based Company Classification

Shoeb Sutar[1,2*], Kupakwashe Mapuranga[2,3†] and Chintan Shah[1,2†]

[1*]AIML Department, Symbiosis Institute of Technology, Pune, Maharashtra, India.
[2]AIML Department, Symbiosis Institute of Technology, Pune, Maharashtra, India.
[3]AIML Department, Symbiosis Institute of Technology, Pune, Maharashtra, India.

*Corresponding author(s). E-mail(s): shoeb.sutar.mtech2025@sitpune.edu.in;
Contributing authors: kupakwashe.mapuranga.mtech2025@sitpune.edu.in; chintan.shah.mtech2025@sitpune.edu.in;
†These authors contributed equally to this work.

Abstract

Environmental, Social, and Governance (ESG) analysis has become a crucial factor in the performance of any company in sustainability and ethical practices. However, the process of manually reviewing hundreds of ESG reports is both time-consuming and can be subjective. This study proposes an approach of automated classification that uses Natural Language Processing and Ensemble Learning to classify companies as Green or Non-Green companies based on textual data of ESG reports. The methodology of the study involves text processing, vectorization of texts using Bag of Words (BoW), TF-IDF, and Word2Vec, which is followed by ensemble learning with a Bagging classifier. The base estimators used are a Decision Tree Classifier paired with Word2Vec and a Complement Naive Bayes Classifier with Bag of Words and TF-IDF. The findings from the study suggest that ESG-based company classification can help investors and stakeholders make informed decisions by offering a scalable and data-driven assessment.

Keywords: ESG analysis, Natural Language Processing, Ensemble Learning, Bao of Words, TF-IDF, Word2Vec, Classification

1 Introduction

Environmental, social, and governance (ESG) principles have become a global standard for assessing an organization's overall sustainability performance in recent years. ESG indicators evaluate an organization's ethical governance practices, social engagement, and environmental footprint management. ESG factors are now essential for investors, regulators, and stakeholders who want to look beyond short-term financial gains and concentrate on long-term responsible growth as global markets move towards sustainable development.[11] Conventional financial analysis models prioritise economic performance and profitability while frequently ignoring the wider sustainability consequences of business operations. Furthermore, it takes a lot of time, is inconsistent, and is subject to subjective interpretation to manually review ESG reports, which are usually long, unstructured textual documents. The need for automated, dependable, and scalable techniques to analyse ESG-related text is increasing due to the abundance of publicly available sustainability reports.[12] The need for automated, dependable, and scalable techniques to analyse ESG-related text is increasing due to the abundance of publicly available sustainability reports. Recent work has applied NLP and ensemble learning for exactly this purpose, leveraging models like BERT, XLM-R, and Word2Vec for scalable ESG classification [1] [2] [7]

An effective answer to this problem is Natural Language Processing (NLP), a branch of artificial intelligence that makes it possible for machines to comprehend and interpret human language. It is feasible to extract significant patterns, sentiments, and indicators from ESG disclosures by utilising NLP techniques. When combined with ensemble learning an advanced machine learning approach that integrates multiple models to improve prediction accuracy and robustness automated ESG classification systems can achieve significantly better performance than traditional single-model approaches.

Globally, organizations like Bloomberg, Refinitiv, CRISIL, Sustainalytics, and S&P Global evaluate companies and provide ESG scores based on a variety of textual and numerical data. These scores are vital for investment decisions, checking regulatory compliance, shaping brand image, and affecting overall market perception. However, since many scoring methods are proprietary and require a lot of resources, there is growing interest in clear and automated tools that can aid in making sustainability-focused decisions.[13] .

This study addresses this need by proposing an NLP-based classification approach to categorize companies as Green or Non-Green using their ESG-related textual data. The objective is to provide an efficient, data-driven alternative to traditional ESG scoring systems, enabling quicker analysis while reducing human bias and error.

1.1 Contribution

1. Compares multiple vectorization methods : BoW, TF-IDF, and Word2Vec
2. Uses Bagging ensemble learning with Decision Tree and Complement Naive Bayes
3. Compares model performance using accuracy, precision, recall, and F1-score
4. Demonstrates practical performance on a real ESG report dataset

2 Related Work

Recent research in ESG text classification has increasingly incorporated ensemble learning techniques combined with advanced natural language processing (NLP) models. Several studies have focused on classifying company ESG behavior or sustainability signals from unstructured text sources such as corporate reports, news articles, and regulatory filings. Lee et al. (2024) used an ensemble of fine - tuned BERT and ALBERT models to classify ESG news articles into environmental, social, and governance topics, achieving an accuracy of 80.79 % [1]. Similarly, Qiu et al. (2023) and Veeramani et al. (2023) participated in the ML-ESG shared task and used ensembles of multilingual transformer models and feature-based methods (e.g., TF - IDF, LSA) to classify news into opportunity / risk labels, reporting high macro and weighted F1-scores above 90 % [2] [3].

Other works have explored hybrid architectures combining transformer models with traditional neural networks or ensemble decision strategies. Abburi et al. (2024) proposed a stacked ensemble of XLM-RoBERTa and CNN models for multilingual ESG news classification [4], while Dakle et al. (2024) achieved top-ranked performance using ensembles of XLM-R and DeBERTa models across English, French, Japanese, and Korean ESG tasks [5]. Luo et al. (2024) took a different approach by using Word2Vec and TF-IDF features from ESG reports and classifying them with Random Forest and XGBoost, where Random Forest achieved 87 % accuracy [7]. These studies demonstrate the growing interest in blending semantic embeddings with ensemble models for improved generalization and interpretability.

Our work aligns with this emerging direction by combining classic vectorization techniques such as Bag of Words, TF - IDF, and Word2Vec with ensemble classifiers like Decision Tree and Complement Naive Bayes under a Bagging framework. While recent literature often emphasizes transformer based models, our results highlight that simpler NLP methods, when paired with robust ensemble techniques, can still deliver highly competitive performance [2]. This also supports scalable and interpretable solutions, particularly valuable in practical ESG analysis scenarios where computational cost and transparency are critical.[20]

3 Methodology

The proposed project employs Natural Language Processing (NLP) techniques combined with Ensemble Machine Learning models to classify companies as Green or Non-Green based solely on textual information extracted from ESG reports. The methodology follows a structured workflow beginning with data preprocessing and culminating in model development and evaluation using standard classification metrics. This systematic approach ensures that the resulting model is both accurate and interpretable.

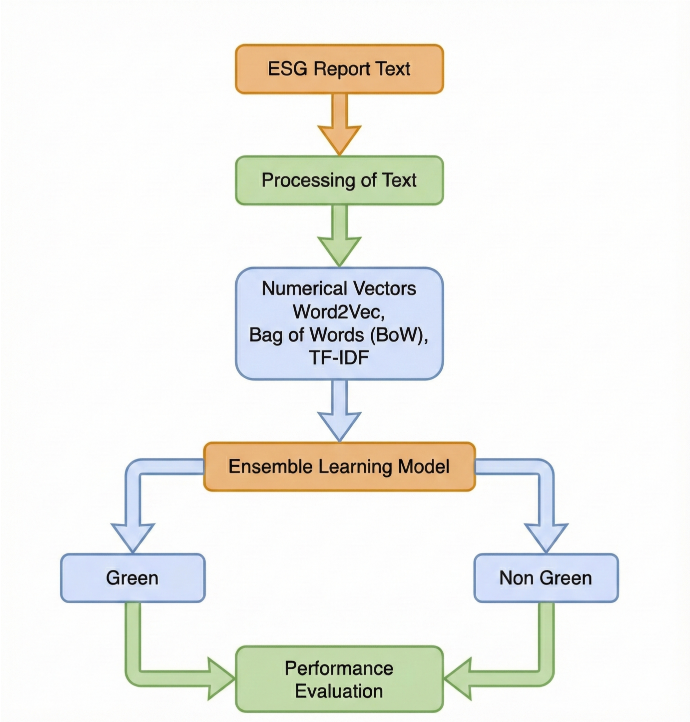

Fig. 1: Methodology

3.1 Dataset Description

The dataset used in the study is titled ESG Sustainability Reports of S & P 500 Companies, was obtained from Kaggle. It includes ESG text extracted from corporate sustainability and annual reports. The ESG ratings provided in the dataset and given by Sustainalytics.[19]

The Target feature is derived based on the threshold value. If the ESG score exceeds the threshold value, the company will be classified as Non-Green, else Green. The threshold ESG score chosen is 29.9 as per documentation of Sustainalytics.

3.2 Data Preprocessing

The raw ESG text data underwent a series of preprocessing steps to prepare it for machine learning analysis. First, tokenization was applied to split the text into smaller units initially into sentences and then into individual words forming the foundation

Table 1: Dataset Structure

Feature	Description
Content	Extracted ESG text data of each company
E-Score	Environmental performance rating
S-Score	Social responsibility rating
G-Score	Governance score
ESG Score	Combined sustainability metric
Target	Class label : Green / Non-Green

for further processing.
Common stopwords such as "is," "the," "and," and "at" were removed to eliminate wordsthat do not add information and reduce noise. Next, lemmatization was performed to convert words to their base forms, ensuring that variations like "running," "runs," and "ran" were standardized to "run."Finally, the text was cleaned by removing unnecessary elements such as punctuation, symbols, numbers, and extra whitespace, resulting in a more refined and consistent dataset suitable for feature extraction and modeling.[16]

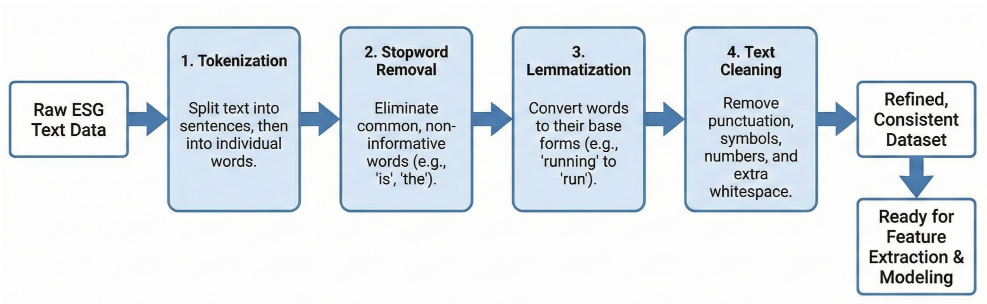

Fig. 2: ESG Text Preprocessing

3.3 Text Vectorization

Text vectorization is the process of converting unstructured textual data into numerical representations that machine learning models can understand and process. Since algorithms cannot interpret raw text directly, vectorization transforms words, sentences, or documents into numerical vectors by capturing features such as word frequency, importance, or semantic meaning. These numerical vectors then serve as inputs for classification, clustering, or other machine learning tasks.[21] Recent ESG classification studies have also used a combination of TF-IDF, Word2Vec, or deep contextual embeddings for this purpose [4] [7].

Bag of Words (BoW):

$$\mathrm{BoW}(d, t_i) = f_i \tag{1}$$

where f_i is the frequency of term t_i in document d. Each document is represented as a vector of term frequencies.

Term Frequency-Inverse Document Frequency (TF-IDF):

$$\text{TF-IDF}(t, d) = \mathrm{TF}(t, d) \times \log\left(\frac{N}{\mathrm{DF}(t)}\right) \tag{2}$$

where: - $\mathrm{TF}(t, d)$ is the frequency of term t in document d - $\mathrm{DF}(t)$ is the number of documents containing term t - N is the total number of documents in the corpus.

Word2Vec (Averaged Word Embeddings):

$$\vec{v}_{\text{sentence}} = \frac{1}{n}\sum_{i=1}^{n} \vec{w}_i \tag{3}$$

where $\vec{w}_i$ is the vector representation of the i^{th} word in the sentence, and n is the total number of words in the sentence.[22]

Table 2: Text Vectorization Techniques

Technique	Concept
Bag of Words	Frequency-based Word Counts
TF-IDF	Score based on uniqueness across corpus
Word2Vec	Semantic embeddings learned using Gensim

3.4 Model Building

Decision Tree Classifier

A Decision Tree is a supervised learning algorithm that splits the input space into distinct regions based on feature values. It uses a tree - like structure of decisions and conditions to classify data points.

At each internal node, the model selects the best feature and threshold to split the data using an impurity measure such as Gini Index or Entropy.

1. Gini Impurity:

$$Gini(D) = 1 - \sum_{i=1}^{C} p_i^2 \tag{4}$$

2. Entropy:

$$Entropy(D) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{5}$$

where p_i is the proportion of samples belonging to class i in dataset D, and C is the number of classes.

The goal is to minimize impurity at each split and construct a tree where each leaf node corresponds to a class label. Decision Trees are interpretable and work well with dense, continuous, or vectorized input like Word2Vec embeddings.[23]

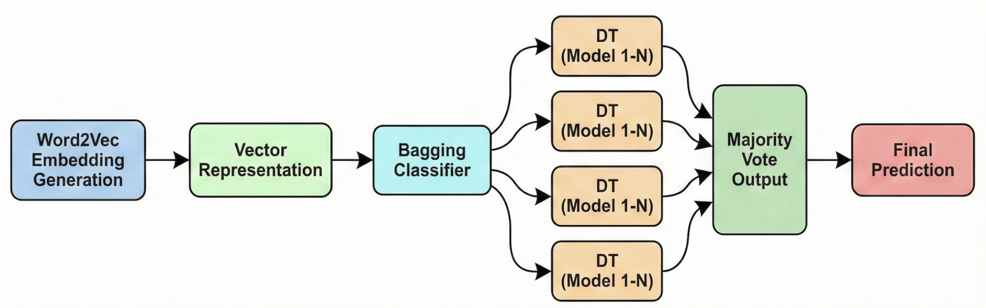

Fig. 3: Word2Vec Model Architecture

Complement Naive Bayes Classifier

Complement Naive Bayes (CNB) is a variant of the Multinomial Naive Bayes algorithm designed to address imbalanced data and improve performance in text classification tasks. It works well with sparse document - term matrices generated by methods like Bag of Words and TF-IDF.

Unlike standard Naive Bayes, CNB is able to estimate the likelihood of a word given the *complement* of each class, which reduces bias in cases where one class dominates. The posterior probability for a class c given a document d with features $x_1, x_2, ..., x_n$ is computed as:

$$P(c \mid d) \propto P(c) \prod_{i=1}^{n} P(x_i \mid \bar{c}) \tag{6}$$

where $P(x_i \mid \bar{c})$ represents the conditional probability of feature x_i given all classes other than c, and $P(c)$ is the prior probability of class c.

CNB is particularly effective when dealing with imbalanced class distributions in high-dimensional spaces, such as text data.

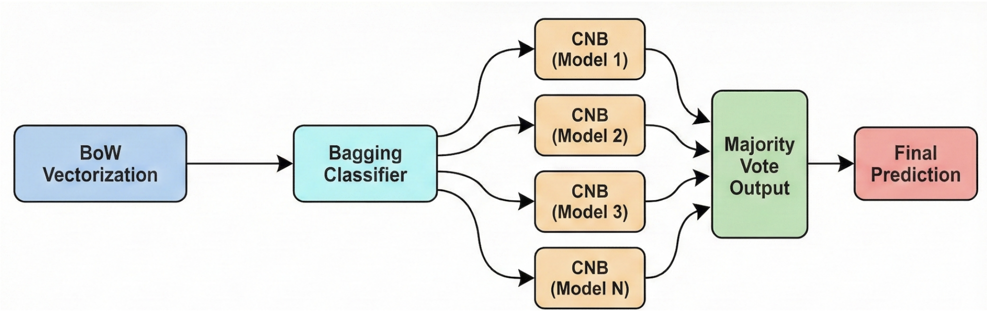

Fig. 4: BoW Model Architecture

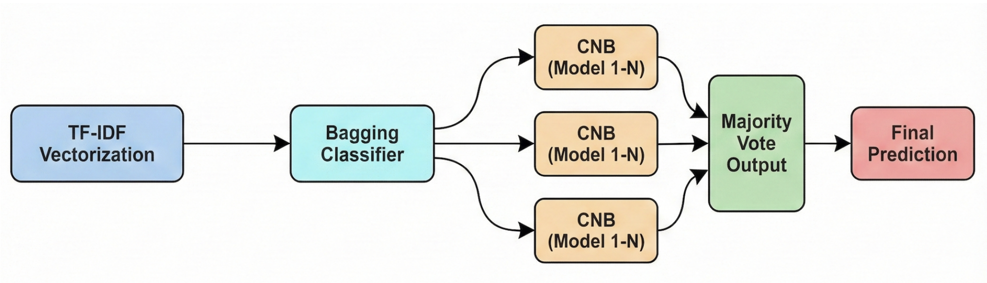

Fig. 5: TF-IDF Model Architecture

3.5 Model Evaluation

Once the training phase was complete, we evaluated each model using standard metrics to understand its effectiveness. We calculated Accuracy to determine the overall percentage of correct predictions and Precision to see how many of the predicted positive cases were actually valid. We also measured Recall to check how well the model identified the actual positive samples, while the F1-Score gave us a balanced view by combining precision and recall.To make these results easier to compare, we plotted performance graphs using the Matplotlib and Seaborn libraries, visualizing how the different vectorization methods performed against each other.[18]

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

These standard classification metrics are calculated as shown in Equations (7) to (10), where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives respectively.

4 Result and Analysis

After building all three models, BoW, TF - IDF, and Word2Vec. We trained them using bagging-based ensemble learning. Then the results were analyzed. How effectively each method classified companies into Green and Non-Green categories was checked. The evaluation used Accuracy, Precision, Recall, F1-Score, and graphical visualizations like Confusion Matrix, Optimal Number of Estimators, and ROC-AUC curve. The results show that every model learned ESG-based patterns from text, but their performance changed depending on how the text was vectorized and which base estimator was used.[15]

Table 3: Performance Metrics Comparison

	Accuracy	**Precision**	**Recall**	**F1-Score**
Bag of Words	0.9249	0.88	0.89	0.8850
TF-IDF	0.87	0.85	0.86	0.8550
Word2Vec	0.885	0.88	0.89	0.8850

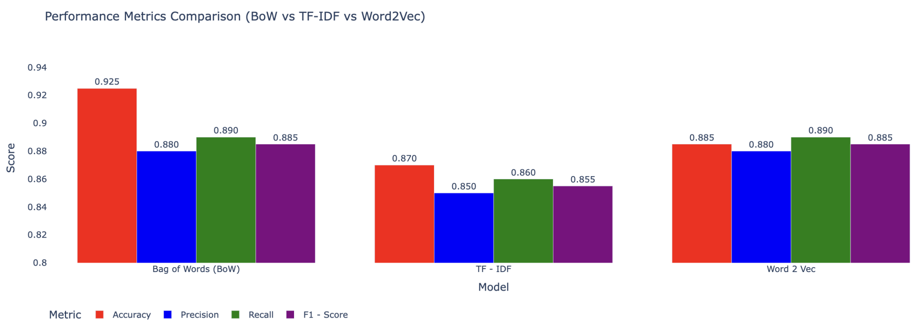

Fig. 6: Performance Metrics Comparison (BoW vs TF-IDF vs Word2Vec)

The results show that every model learned ESG-based patterns from text, but their performance changed depending on how the text was vectorized and which base estimator was used.

4.1 Word2Vec with Decision Tree

For the Word2Vec model with a Decision Tree as base estimator, words from ESG reports were turned into vector representations using the Gensim Word2Vec library. The following hyperparameters were chosen after testing,

Table 4: Parameter Description (Word2Vec model)

Parameter	Description	Value
min_count	Ignores words that occur less than this number of times	2
window	Context window size	7
vector_size	Dimension of word vector	300
sg	skip-gram model	1

Word2Vec will convert each word into a 300-dimensional vector, so to convert each sentence into a vector, first each word should be converted to a vector, and the average of all these words should be taken to get the final sentence vector.

$$\vec{v}_{\text{sentence}} = \frac{1}{n}\sum_{i=1}^{n} \vec{w}_i \tag{11}$$

The generated word vectors were then input into a bagging Ensemble Model.

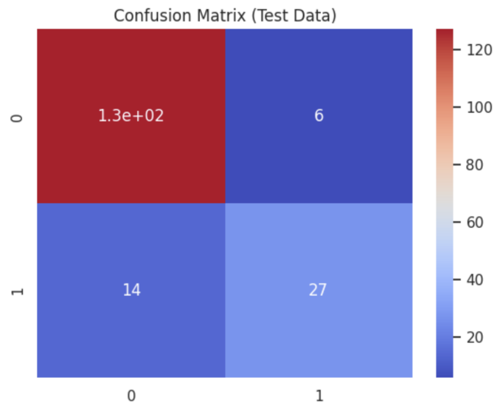

Fig. 7: Confusion Matrix (Word2Vec)

The highest accuracy occurred when the number of base estimators was 120. However, the model displayed signs of overfitting, with high training accuracy but slightly lower generalization.

As a result, n equal to 110 was chosen as the best number of estimators, achieving a

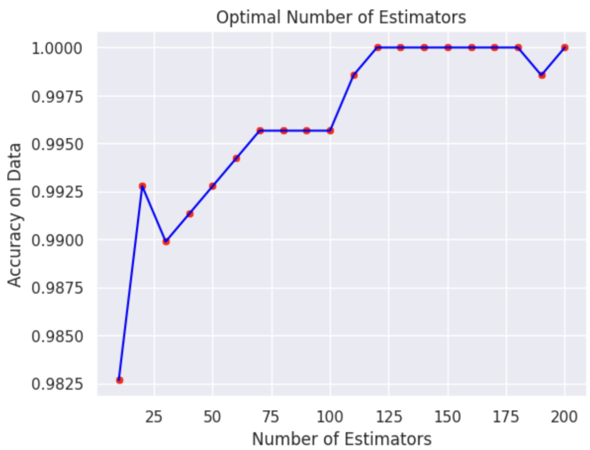

Fig. 8: Optimal Number of estimators (Word2Vec)

balance between performance and stability. This showed strong interpretability and solid classification ability on vectorized data.

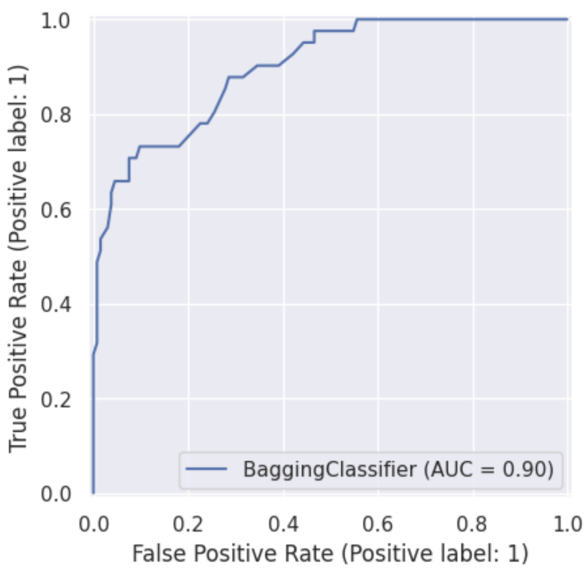

Fig. 9: ROC - AUC Curve (Word2Vec)

4.2 Bag of Words with CNB

In this method, text from the ESG reports was converted into vectors using a BoW model. The generated word vectors were then passed into a bagging ensemble model, using a Naive Bayes classifier as a base estimator.

The number of base estimators was adjusted repeatedly, from 10 to 200, to determine where accuracy performs best. The highest accuracy occurred when n = 70. The following table shows the details of parameter values for training the ensemble model. The performance of the BoW model is illustrated with visual indicators. The confu-

Table 5: Model Parameters (BoW)

Parameter	Description	Value
estimator	Base classifier used inside Bagging	Complement Naive Bayes
n_estimators	Number of models combined inside the ensemble	70
max_samples	Fraction of training data sampled for each base model	0.8
bootstrap	Whether sampling is done with replacement	True
random_state	Ensure reproducibility of results	42

sion matrix shows that the model can classify the majority of data points correctly with minimal misclassification. The optimal estimator plot confirms that the model is achieving the highest accuracy around n = 70. A high ROC-AUC value further supports the model's strong classification ability.

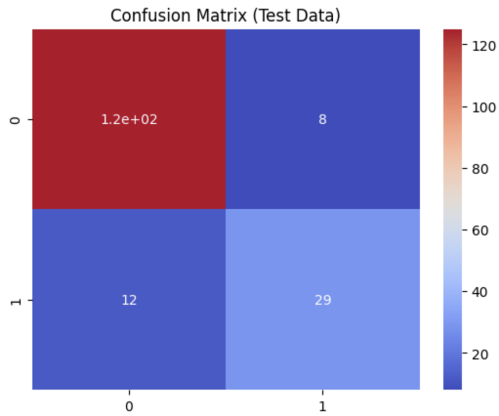

Fig. 10: Confusion Matrix (BoW)

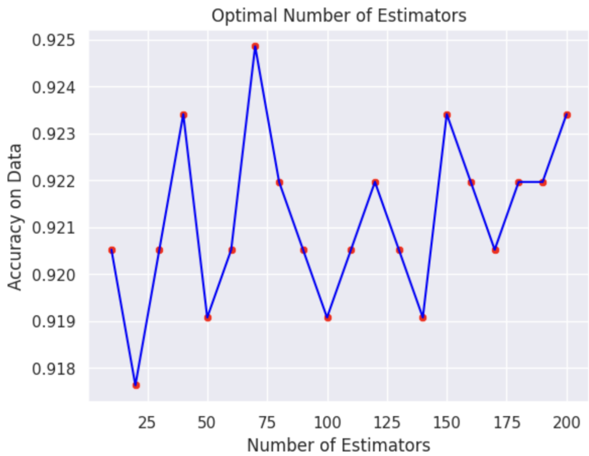

Fig. 11: Optimal number of estimators (BoW)

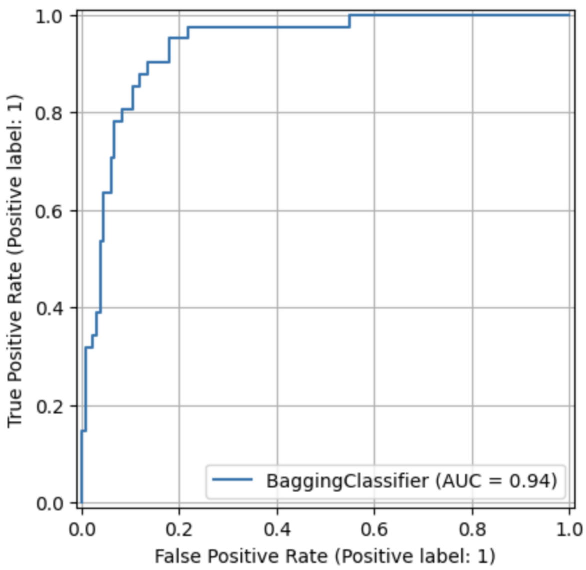

Fig. 12: ROC - AUC (BoW)

4.3 TF-IDF with CNB

TF-IDF uses the uniqueness of a word in the corpus to assign a weight to it. The words that appear too frequently across the corpus will receive lower weights. In contrast, specific ESG-related keywords will get higher relevance. This helps the model focus on meaningful terms instead of common filler words.

The Complement Naive Bayes Classifier was used as it works well with sparse vector data. The generated word vectors were then used as input into a Bagging Ensemble model. It used the Complement Naive Bayes Classifier as the base estimator.The number of base estimators was iterated from 10 to 200. The highest accuracy occurred when the number of estimators reached 90.The following table shows the details of parameter values for training the ensemble model.

Table 6: Model Parameters (TF-IDF)

Parameter	Description	Value
estimator	Base classifier used inside Bagging	Complement Naive Bayes
n_estimators	Number of models combined inside the ensemble	90
max_samples	Fraction of training data sampled for each base model	0.8
bootstrap	Whether sampling is done with replacement	True
random_state	Ensure reproducibility of results	42

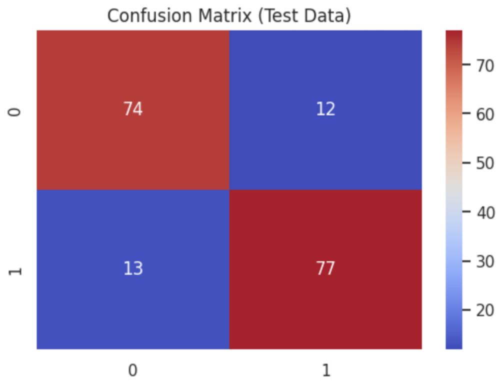

Fig. 13: Confusion Matrix (TF-IDF)

4.4 Comparative Analysis with related work

To further understand the effectiveness of our proposed method, we compared its performance with recent ESG classification models that also rely on textual data from company reports. The following table highlights studies that used traditional machine learning or hybrid NLP methods similar to ours. As shown, our model achieved higher accuracy than comparable methods, demonstrating that well - optimized classical techniques can perform better than more complex or resource - intensive alternatives in ESG text classification tasks.[10]

Table 7: Comparison with Recent ESG Classification Models

Study	NLP Technique(s)	Result (Accuracy / F1)
Luo & Zhou (2024)	Word2Vec + TF-IDF	87%, AUC: 0.94
Raman et al. (2023)	XLNet (fine-tuned)	F1 ≈ 78%
Schimanski et al. (2024)	ESG-BERT (pretrained)	Not reported
Leone et al. (2025)	ESG Sentiment + Transformer Models	F1 ≈ 82%
Seow & Tan (2025)	Deep ESG Analytics (Hybrid NLP)	Accuracy ≈ 89%
This Study (2025)	BoW, TF-IDF, Word2Vec	**92.49%**, F1: 0.8850

5 Conclusion and Future Direction

This study successfully demonstrated that combining Natural Language Processing (NLP) with ensemble learning techniques is a highly effective way to classify companies as either "Green" or "Non-Green" based on the text found in their ESG disclosures. Among the various approaches tested, the Bag of Words (BoW) model, when used with a Complement Naive Bayes (CNB) learner, achieved the highest accuracy of 92.49%. This performance notably surpassed other vectorization methods like TF-IDF and Word2Vec, suggesting that even basic textual features, when paired with bagging strategies, can deliver robust and consistent results [7] [8]. Overall, the method proposed in this research helps solve the problem of slow manual analysis, offering a faster and more reliable metric that can significantly aid analysts in making data-driven decisions.[14]

Looking ahead, there are several key opportunities to expand the scope and practical value of this project. While the current model offers a simple binary classification, future versions could be designed to provide more detailed insights by assigning specific ESG grades such as A, B, C, or D or by scoring the Environmental, Social, and Governance pillars independently. To make the model more adaptable to the real world, the dataset should also be grown to include companies from diverse global markets and industries, rather than relying solely on the current S & P 500 reports. Finally, a major practical goal is to deploy this trained model as a user-friendly web or cloud application. This would allow investors to upload company documents and instantly receive classification results with confidence scores, transforming this research into a powerful everyday tool for the industry.

Limitations

While the proposed model showed strong performance, there are some limitations. The dataset only includes ESG reports from S & P 500 companies, so the results may not generalize to other regions or industries. The classification is limited to two classes (Green and Non - Green), which does not fully capture the complexity of ESG factors. Finally, cross-validation and statistical tests were not performed in this version but are important for improving reliability of model.

Practical Implications

This study shows that simple and easy-to-understand machine learning models can be used to classify companies based on ESG text data. The approach is useful for investors, analysts, or regulators who need a fast and affordable way to assess sustainability without relying on ESG rating agencies. Since it uses open-source tools and public data, the model can be applied by organizations with limited resources. Its transparency and simplicity also make it easy to add into ESG tools or decision-making systems.[17]

Declarations

Data Availability

The dataset used in this study, titled *"ESG Sustainability Reports of S&P 500 Companies"*, was obtained from the public Kaggle repository. It is available at: https://www.kaggle.com/datasets/jaidityachopra/esg-sustainability-reports-of-s-and-p-500-companies.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] J. Lee, M. Kim, and S. Cho, "ESG Topic Classification Using BERT and ALBERT Based Ensemble," *Heliyon*, vol. 10, no. 2, pp. e19785, 2024.

[2] Y. Qiu, H. Song, Z. Deng, and Z. Yu, "Ensemble Learning for ESG Opportunity and Risk Detection in Multilingual News Articles," in *Proc. FinNLP Workshop at IJCAI*, 2023.

[3] S. Veeramani and M.-F. Moens, "Multilingual ESG Opportunity-Risk Detection Using Transformer Ensembles," in *Proc. FinNLP Workshop at IJCAI*, 2023.

[4] S. K. Abburi, Y. Patel, and U. Singh, "Stacked Ensemble of XLM-RoBERTa and CNN for Multilingual ESG Classification," in *Proc. FinNLP Shared Task*, 2024.

[5] H. Dakle, S. Venkatesan, and A. Raj, "Multilingual ESG News Classification with Transformer Ensembles," in *Proc. FinNLP Shared Task*, 2024.

[6] H. Rong and L. Yang, "Averaging and Random Forest Ensembles for ESG Impact Duration Classification," in *Proc. FinNLP Shared Task*, 2024.

[7] Y. Luo and Z. Zhou, "Detecting ESG Report Exaggeration Using Ensemble Machine Learning," *Data and Information Management*, vol. 8, no. 1, pp. 1–13, 2024.

[8] T. Schimanski, A. Reding, J. Bingler, and M. Leippold, "Bridging the Gap in ESG Measurement: Using NLP to Quantify Environmental, Social, and Governance Communication," *Finance Research Letters*, vol. 56, p. 104227, 2024.

[9] M. Raman, R. Agarwal, and S. Rajput, "ESG Signal Detection in Earnings Calls via Weakly Supervised Learning," *Neural Computing and Applications*, vol. 35, pp. 15027–15039, 2023.

[10] G. Dorfleitner and R. Zhang, "ESG News Sentiment and Stock Price Reactions: A BERT-Based Approach," *International Review of Financial Analysis*, 2024.

[11] K. Du and H. Zhang, "Natural Language Processing in Finance: A Comprehensive Survey," *Decision Support Systems*, 2025.

[12] B. Hamnache Arias, "Predicting ESG Scores Using Numerical and Textual Data: Combining ML Techniques," Master's Thesis, Tilburg University, 2024.

[13] N. Hammadi, "NLP-Driven ESG Disclosure Analysis in Banking Annual Reports," *International Journal of Corporate Governance*, 2024.

[14] V. Lagasio and M. Musella, "Detecting ESG-Washing Using Natural Language Processing in Sustainability Reports," *International Review of Financial Analysis*, 2024.

[15] F. Leone, A. Fiordelisi, and R. Ricci, "Price Forecasting with Global and ESG Sentiment Scores," *Finance Research Letters*, 2025.

[16] S. Katz, Y. Gu, and L. Jiang, "Extracting ESG Assurance and Framework Information from Sustainability Reports Using NLP," SSRN Working Paper 4836432, 2024.

[17] C.-C. Chen et al., "Multilingual ESG Impact Duration Inference Using Transformer Models," in *FinNLP Workshop*, 2024.

[18] K. Tian and H. Chen, "Automatic ESG Taxonomy Classification Using Augmented Term Embeddings," in *FinNLP Workshop*, 2022.

[19] R. Seow and L. Tan, "AI-Driven Sustainable Finance: ESG Analytics, Governance, and Risk," *Frontiers in Business Research*, 2025.

[20] S. Singh and P. Kumar, "Novel Data Mining Methodologies for ESG Assessment," *Preprints*, 2025.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL-HLT*, 2019.

[22] T. Mikolov et al., "Distributed Representations of Words and Phrases and Their Compositionality," in *NeurIPS*, 2013.

[23] L. Breiman, "Bagging Predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.