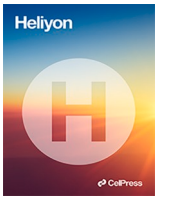
Heliyon 10 (2024) e26404

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models

Haein Lee [a], Seon Hong Lee [a], Heungju Park [b], Jang Hyun Kim [c], Hae Sun Jung [d,*]

[a] Department of Applied Artificial Intelligence/ Department of Human Artificial Intelligence Interaction, Sungkyunkwan University, 03063, Seoul, South Korea
[b] SKK Business School, Sungkyunkwan University, 03063, Seoul, South Korea
[c] Department of Interaction Science/ Department of Human Artificial Intelligence Interaction, Sungkyunkwan University, 03063, Seoul, South Korea
[d] Department of Applied Artificial Intelligence, Sungkyunkwan University, 03063, Seoul, South Korea

ARTICLE INFO

Keywords:
ESG
Natural language processing (NLP)
Ensemble
Pretrained language model
BERT

ABSTRACT

Incorporating environmental, social, and governance (ESG) criteria is essential for promoting sustainability in business and is considered a set of principles that can increase a firm's value. This research proposes a strategy using text-based automated techniques to rate ESG. For autonomous classification, data were collected from the news archive LexisNexis and classified as E, S, or G based on the ESG materials provided by the Refinitiv-Sustainable Leadership Monitor, which has over 450 metrics. In addition, Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), and A Lite BERT (ALBERT) models were trained to accurately categorize preprocessed ESG documents using a voting ensemble model, and their performances were measured. The accuracy of the ensemble model utilizing BERT and ALBERT was found to be 80.79% with batch size 20. Additionally, this research validated the performance of the framework for companies included in the Dow Jones Industrial Average (DJIA) and compared it with the grade provided by Morgan Stanley Capital International (MSCI), a globally renowned ESG rating agency known for having the highest creditworthiness. This study supports the use of sophisticated natural language processing (NLP) techniques to attain important knowledge from large amounts of text-based data to improve ESG assessment criteria established by different rating agencies.

1. Introduction

The term environmental, social, and governance (ESG) was initially presented in a report "Who Cares Wins" by the United Nations Global Compact (UNGC) in 2004 [1]. In 2006, the United Nations Principles for Responsible Investment (UNPRI) reinforced ESG as a financial investment principle and established the framework for ESG [2]. Since then, the importance of ESGs has grown significantly, with major global agendas, including international cooperation, to combat climate change by reducing carbon emissions [3]. Although the significance of ESG is growing steadily, the impartiality of ESG ratings remains debated. According to Ref. [4], indicators such as bond ratings showed a similar trend regardless of the rating agency, but ESG ratings were highly inconsistent. Additionally [5], highlighted the absence of standards for reporting ESG information. Furthermore, the lack of transparency regarding the evaluation

* Corresponding author.
E-mail address: jestiriel@g.skku.edu (H.S. Jung).

https://doi.org/10.1016/j.heliyon.2024.e26404
Received 26 August 2023; Received in revised form 20 December 2023; Accepted 13 February 2024
Available online 14 February 2024

criteria by ESG rating agencies made it even considerably harder to ensure objectivity [6]. As the significance of ESG continuously grows, there is an increasing volume of data being produced in this area. Consequently, it has become essential to employ advanced natural language processing (NLP) techniques to effectively manage and analyze unstructured textual data. Furthermore, optimal data usage has become increasingly important [7,8].

In this study, data were collected from the news archive LexisNexis and labeled E, S, or G using ESG documents provided by the Refinitiv-Sustainable Leadership Monitor, which employs over 450 scales. Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), and A Lite BERT (ALBERT) models were subsequently trained to accurately classify preprocessed ESG documents to E, S, and G using a voting ensemble model, and their performance was evaluated.

To validate the effect of the proposed approach, this study further collected news data from LexisNexis for 10 enterprises listed in the Dow Jones Industrial Average (DJIA) [9]. The collected text data were then categorized as E, S, or G using the proposed framework. Additionally, sentiment analysis was conducted on the labeled data utilizing FinBERT [10]. Subsequently, the resulting outputs were weighted for each category (E, S, G) and the ESG grade was determined by integrating the results of sentiment analysis. The extracted ESG grades were compared with those provided by Morgan Stanley Capital International Index (MSCI), a prominent organization specializing in ESG ratings [11]. MSCI assigns grades to registered companies based on three categories: laggard, average, and leader. However, the specific rating standards or criteria employed by MSCI are not disclosed. Therefore, the comparison between the grades obtained from MSCI and those derived from the proposed framework indicates the potential of this approach in extracting meaningful objective ESG grades. If the evaluation results of the framework align closely with the grades assigned by MSCI which incorporates diverse information sources, it would substantiate the validity of this approach. Such validation process is essential for resolving real-world problems [12,13].

This study is a novel study that addresses the reliability problem deriving from the lack of disclosure regarding evaluation methodologies and processes by the majority of ESG rating companies. In addition, this approach indicates the possibility of achieving objective ESG grading using corporate open data, such as news data. In conclusion, this paper aims to achieve the following objectives:

- Designing an ensemble model capable of automatically classifying unstructured textdata from news articles into E, S, and G.
- Comparing the extracted grades with the grades from MSCI after conducting weightedsummation operations.

The organization of the research is as follows. In the subsequent section, reviews of related works are performed. In the Method section, descriptions of the data collection and analytical techniques are presented. Finally, the Results section presents the findings, and the Discussion and Conclusion section presents the implications of this study.

2. Related works

2.1. Study on intrinsic value of ESG

As stated in an earlier section, the phrase ESG first officially appeared in a report “Who Cares Wins” published by the UNGC [1]. ESG was designed to develop investment knowhow in reflection of ESG factors and emphasized the importance of sustainable development. Owing to the growing importance of ESG, related studies are emerging. According to Ref. [14], companies that overlook their social responsibilities or have ineffective governance face substantial “hidden” risks of severe economic losses or payment of costs due to environmental litigation. Additionally [15], affirms the significance of ESG metrics in Socially Responsible Investment (SRI).

2.2. Study on positive correlation between ESG and company performance

As the importance of ESG becomes increasingly significant, research has illustrated a positive correlation between ESG and corporate performance indicators. Reference [16] extracted data and aggregated the findings from 2,200 studies to figure out the relations between Corporate Financial Performance (CFP) and ESG. The consequences showed that the business case of ESG investments is well-established. Most studies found nonnegative ESG-CFP relationships, with most studies reporting positive findings.

The positive affection of ESG on CFP appeared to be reliable over time, and the ESG criteria were positively correlated with CFPs on average. According to Ref. [17], it is possible to significantly reduce costs through ESG management. ESG was found to be effective in reducing operational costs, such as raw material or carbon costs, which could affect an enterprise’s operating income by up to 60%. In addition, the authors discovered a meaningful relationship between resource capability and financial achievement. According to Refs. [18,19], strong ESG within an enterprise made it possible to attract competent employees, enhance motivation, and eventually increase overall productivity. Furthermore, the social responsibility of these companies showed a positive correlation with shareholder returns.

2.3. Study on objectivity of ESG rating

However, the impartiality of ESG indicators remains a contentious issue. According to4, although the bond ratings of certain issuers tended to be similar across rating agencies, ESG ratings were often inconsistent. In other words, a company that scored high on one evaluator may score low on another. According to Ref. [5], the biggest challenge investors faced when combining ESG information into their investment strategy was the absence of standards for managing ESG information reporting. Reference [6] compared the most illustrative ESG ratings and information providers and highlighted the limitations of identifying objectivity or equity owing to the

absence of disclosed information provided by ESG rating companies on their evaluation standard. Reference [20] provided evidence of the relationship between an enterprise's credit rating and Corporate Social Responsibility (CSR). Each ESG rating agency considered different individual elements of CSR to be pertinent, leading to different ESG ratings. To overcome the problems identified in previous research, a framework and standard are required that can objectively measure ESG ratings regardless of the data employed.

2.4. Research on ESG using natural language processing

The goal of NLP is to gain insights into the information extracted from text. This section reviews existing research that applied NLP to gain insights into ESG.

Reference [21] introduced the ESG2 Risk framework, a model that predicts future stock profit volatility using transformer-based language models. By extracting and utilizing ESG-related news data from 2003 to 2019, the authors confirmed that ESG news flows have a significant impact on companies' future returns and risks. The authors concluded that ESG factors are relevant for investors when making investment decisions. Reference [22] employed NLP to convert unstructured text data extracted from social media into ESG scores. To enhance the accuracy of evaluating the correlation of documents in ESG contexts, the authors utilized BERT and devised ESG score automation to identify ESG risks. Reference [23] used big data from news articles and academic papers to analyze ESG discourse. Using Bidirectional Encoder Representations from the Transformers topic (BERTopic), the authors found major keywords and topics associated with ESG and observed the changes in topic patterns over time. Based on these findings, the authors proposed a strategic direction for successful ESG management. Reference [24] analyzed past tendencies in ESG discourses by examining the records of corporate earning calls. The authors used a pre-trained transformer-based model to classify the correlation of text sentences to ESG and applied it to sentences in conference transcripts. The authors found that over the past five years, 15 percent of statements made during earning calls were related to ESG, indicating the growing importance of ESG in business strategy. Reference [25] explored the influence of news sentiment related to ESG on the achievement of the DJIA. The authors extracted the sentiments of news articles related to ESG topics between 2010 and 2018 and calculated a polarity-based sentiment index. The authors found that changes in news sentiment related to ESG impact the profits of most DJIA constituents.

2.5. Studies employed machine learning and ensemble learning

Machine learning and ensemble methods are being actively applied in various fields to derive insights or perform classification tasks.

Reference [26] employed deep learning and pre-trained language models for classifying user sentiments in metaverse services. The authors utilized soft voting mechanism and achieved 88.57% for the sentiment classification task. Reference [27] utilized convolutional neural network (CNN) for automatically detecting neonatal quiet sleep (QS). As a result, the authors achieved 94.07% accuracy with high computational efficiency. Reference [28] applied diverse machine learning algorithms for predicting upward/downward trends of Bitcoin price with sentiment analysis and technical indicators. Consequently, the authors achieved an accuracy of 90.57%. Reference [29] focused on predicting user satisfaction with mobile healthcare services with five machine learning classifiers. Consequently, logistic regression with TF-IDF showed superior performances on prediction tasks. Reference [30] investigated an automatic sleep stage classification system utilizing internet of things (IoT) and ensemble techniques to replace time-consuming visual annotation in neonatal sleep analysis. The authors utilized multiple classifiers such as CNN, support vector machine, and multi-layer perceptron. Then ensemble algorithms were used to combine the outputs for final classification.

However, throughout the related works, several technical gaps were identified, which have led to the design of the proposed methodology. The first gap is the limited focus on automated classification. Previous studies have relied on manual analysis or subjective methods to evaluate ESG performance. There is a lack of research focusing on automated techniques, such as NLP, for classifying unstructured text data from news articles into E, S, and G categories. This gap suggests a need for a more objective and efficient approach to extract relevant information from textual data. The second gap is that there is insufficient comparison with external ratings. Previous studies lacked a comprehensive evaluation of their results by comparing them with well-known ESG rating agencies, such as those from MSCI (Morgan Stanley Capital International). This gap indicates the importance of comparing the proposed methodology with results from widely accepted institutions to assess reliability. By addressing these technical gaps, the proposed methodology aims to contribute to the existing body of knowledge by providing an automated, ensemble-based classification model for ESG analysis. Additionally, this study aims to enhance the evaluation process by performing weighted summation operations on the extracted results and comparing them with established ratings.

3. Method

The framework proposed in this paper consists of two main phases. The first process is the textual ESG classification model with a pre-trained language model. The second stage involves utilizing labeled data and sentiment analysis to extract the ESG grades on given companies.

3.1. Textual ESG classification model

In this first process, the authors examined a robust deep learning model on data from LexisNexis, the most utilized news archive in the social sciences [31]. Previous research has pointed out that the E, S, and G assessment methodologies were not merged [32]. To

ensure consistency in the ESG assessment methodologies, the authors labeled the LexisNexis data using the ESG document provided by the Refinitiv- Sustainable Leadership Monitor, which employs over 450 ESG scales [33]. The constructed data were used to train a pre-trained language model. The system architecture is illustrated in Fig. 1.

3.1.1. Data collection

To build a dataset for ESG scoring, news data from LexisNexis was collected using the data queries "ESG" and "esg" between January 1, 2016, and March 13, 2023, resulting in 16,735 data. The collected data were labeled using the ESG pillar scores of the Refinitiv-Sustainable Leadership Monitor, a metric designed to calculate the relative environmental, social, and governance performance of companies within a sector [34,35]. The scores were structured to measure ESG achievement and efficiency. Using these scores, the authors assigned labels to each article based on the highest frequency among E, S, and G appearances.

3.1.2. Preprocessing

The E, S, and G labels obtained from the LexisNexis data showed an imbalance, with 8,876 articles labeled as "E", 3,135 as "S", and 1,784 as "G" (Fig. 2). This imbalance can negatively impact the model's performance. To solve this issue, the authors applied down sampling to balance the number of classes by resolving this disproportion [36]. Afterwards, tokenization and lemmatization were performed using the Spacy Python library [37]. Finally, regular expressions and removing duplication tasks were performed. The authors randomly divided the entire dataset into training and test sets in an 8:2 ratio to train the deep learning model. To prevent overfitting, 20% of the training data was allocated as the validation set.

The authors utilized histograms to examine the distribution of the preprocessed dataset (Fig. 3). The average sentence length for the articles was determined to be 976.47. Furthermore, the first quartile length was 527.0 and the third quartile length had a value of 1104.0, and the most of data was concentrated within this range. Additionally, the authors employed statistical analysis, utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) score, to assess word importance [38]. From each category, the top 10 words with the highest TF-IDF scores were extracted (Table 1). In the "E" category, specific words like energy, climate, carbon, and sustainability were regarded as crucial, highlighting their relevance to the environment. However, discerning notable distinctions between words in the S and G categories proved challenging.

Nevertheless, a distinct semantic disparity exists between "S," which denotes social aspects like CSR and influence, and "G", which indicates the governance structure representing a company's internal governance [39]. The verifiable fact through TF-IDF analysis is that conventional statistical approaches fail to grasp the implied meanings in ESG domain documents, underscoring the importance of adopting advanced NLP techniques to accomplish the objective.

3.2. System architecture: pre-trained language model

The proposed NLP approach involves a transformer architecture-based language modeling methodology that learns general language properties and can be used to identify linguistic characteristics. This approach includes pretraining and downstream steps, in which a large corpus is pre-trained to capture the general attributes of the language. Pretraining helps generalize downstream tasks that work with small datasets. Although pretraining is costly, it improves the performance and convergence of the downstream model. However, the data distribution of a specific domain may differ from that of the pretraining data. Thus, fine-tuning was employed to

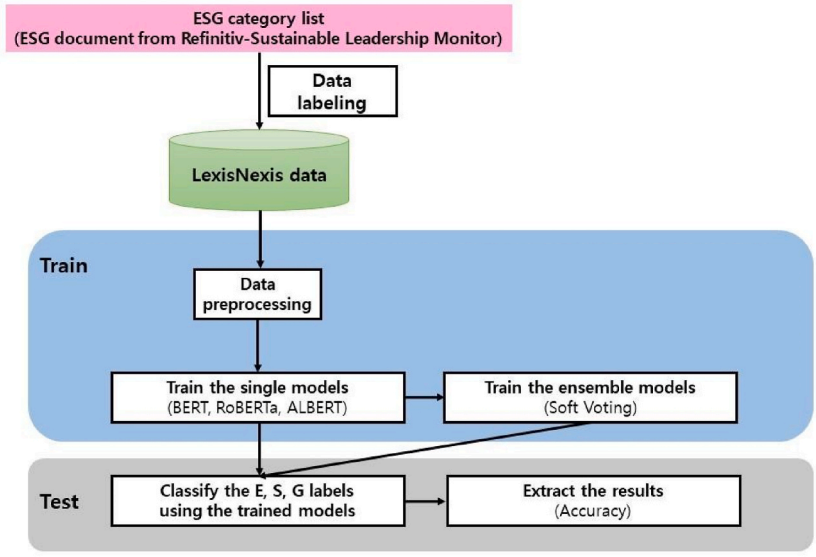


Fig. 1. Visual representation of the ESG classification flowchart.

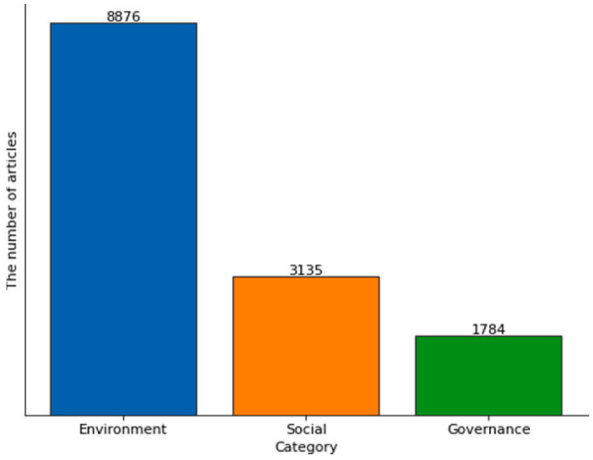


Fig. 2. Description of the label distributions.

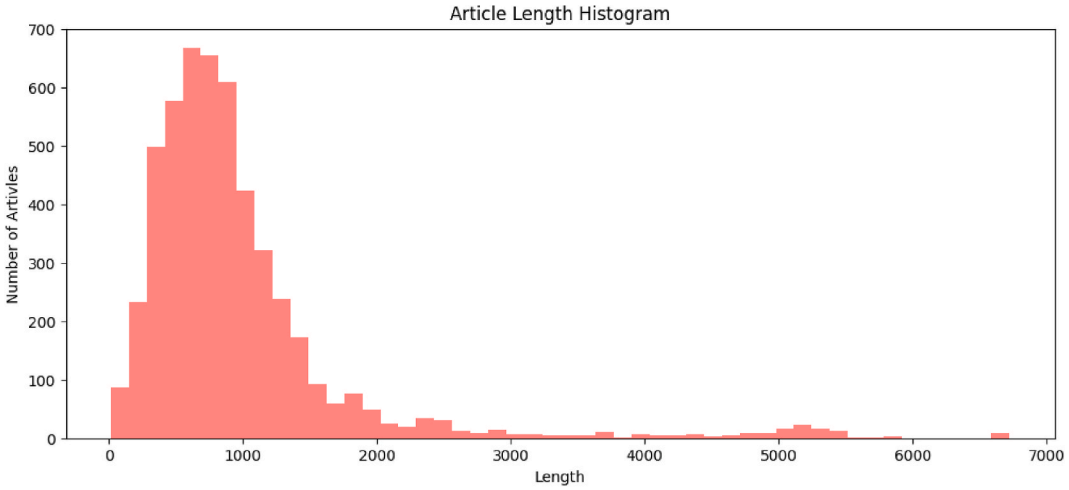


Fig. 3. Histogram of the number of words in articles.

Table 1
Top 10 words with the highest TF-IDF scores by category.

Environment		Social		Governance	
Word	TF-IDF	Word	TF-IDF	Word	TF-IDF
energy	167.1550	business	164.7125	investment	191.8154
climate	159.3469	investment	159.3763	fund	155.1029
investment	138.4658	fund	142.9709	management	146.9943
carbon	133.4575	says	140.5733	business	144.7922
business	127.0346	social	117.9323	financial	138.246
financial	125.0324	investing	115.1547	market	135.1126
green	124.7973	financial	113.092	climate	134.5719
sustainability	123.6673	over	111.8117	asset	120.4749
global	118.4634	market	110.5638	global	120.0967
market	117.9908	people	105.2516	social	116.1851

tune the Language Model (LM) for the data to be handled in the downstream stage. Given a pre-trained general language model, this step can achieve faster convergence and robust performance even on small datasets by learning distinct data specialties [40]. Specifically, by fine-tuning the model using our news dataset, the model reflected the subjectivity and biases inherent in the news itself. The authors experimented with various models, such as BERT, RoBERTa, and ALBERT, which are applicable to the ESG domain. Compared to other LMs, BERT-based language models have the advantage of facilitating optimization during fine-tuning [41].

Moreover, through sequential learning of multiple tasks, BERT improves the overall performance in the downstream stage [42].

The encoder component of the transformer was used to develop the BERT model [43]. In the pretraining process, two unsupervised learning tasks were performed using unlabeled data. The first is Masked LM. In this process, a certain percentage of the input tokens were masked, and the masked tokens were predicted. The second is the Next Sentence Prediction (NSP), which learns whether two sentences are connected or not. The data used in these processes are from BookCorpus and English Wikipedia [44]. Subsequently, in the fine-tuning step, the parameters were tuned using the labeled data. Ultimately, BERT presents strength in language understanding tasks. The entire structure is illustrated in Fig. 4.

RoBERTa has emerged to complement the lack of BERT training in various validation experiments. BERT uses static masks, whereas RoBERTa learns different masked sentences in each epoch. Additionally, when the NSP task is excluded, the model performance is similar or slightly better. Finally, for robustness and optimization, the RoBERTa model in Fig. 5 uses a 160 GB dataset collected from five domains and a large batch size [45].

Since the introduction of the BERT model, it has become common practice to train large structures and fine-tune them using a distillation model to improve network performance. However, large model architectures can cause memory limitation problems or significantly reduce the speed. Therefore, ALBERT separates a huge word embedding matrix into two smaller metrics with factorized embedding parameterization. Additionally, cross layer parameter sharing prevents the number of parameters from increasing as the network depth increases. In addition, ALBERT performance is enhanced by reducing NSP inefficiency and boosting inter-sentence coherence with sentence order prediction [46]. The overall flow is illustrated in Fig. 6.

A soft-voting classifier is a meta classifier that uses single predictions as features to produce a final prediction. It uses the classes predicted by several classifiers and selects the final class as the desired result [47]. Because it aggregates the predictions of various models, the voting model exhibits better outcomes than the other baseline models [48]. After training the single models, the voting ensemble models were set. Consequently, this study employed deep learning models to create multiple ensemble models for all possible outcomes: BERT, ALBERT, and RoBERTa (Fig. 7).

3.3. ESG assessment framework with pre-trained ensemble model: ESG2PreEM

The second process focused on validating the proposed framework: ESG2PreEM. For this purpose, news data for 10 companies belonging to the DJIA were collected. The proposed model was then utilized to label the collected data into E, S, and G categories. Afterwards, sentiment analysis was performed on the classified text data, and through calculations considering the frequency-based weights, the ESG grade for each company was obtained as the result. The system architecture is illustrated in Fig. 8.

3.3.1. Data collection

To assess the ESG ratings of 10 companies; Coca, 3 M, Cisco, Home Depot, McDonald, Walmart, American Express, JPMorgan Chase, Apple, and Nike listed on the DJIA, news article data containing the query "ESG" along with the names of each company was collected from LexisNexis for the period between January 1, 2016, and March 13, 2023. A total of 32,784 data were collected from LexisNexis on the selected companies.

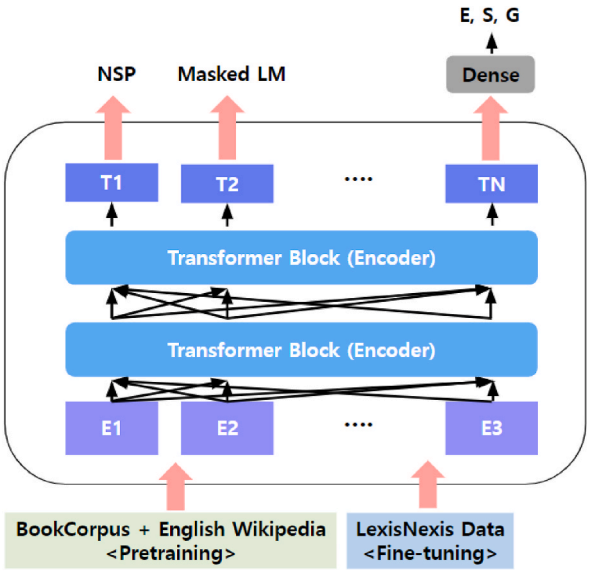


Fig. 4. Structure of the BERT model.

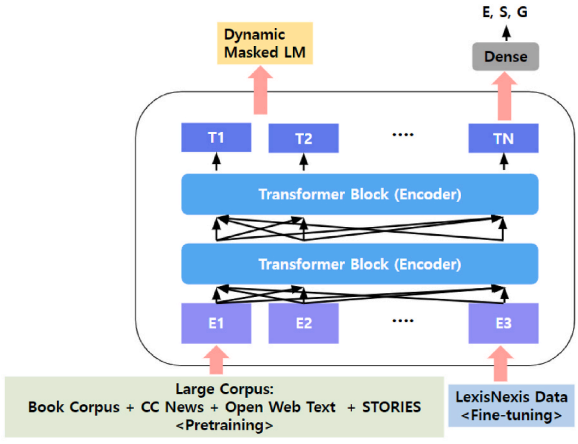


Fig. 5. Structure of RoBERTa model.

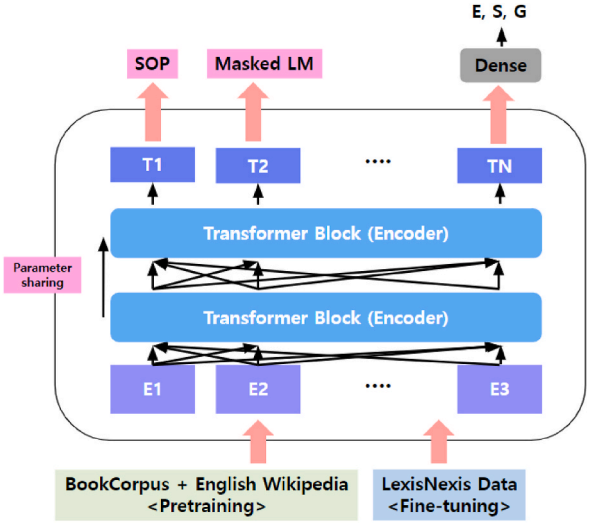


Fig. 6. Structure of ALBERT model.

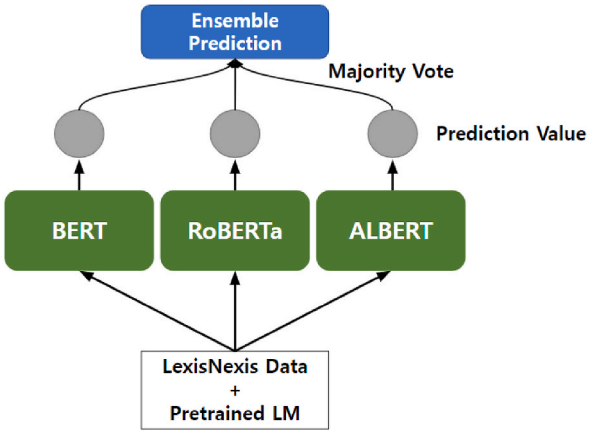


Fig. 7. Structure of soft voting classifier.

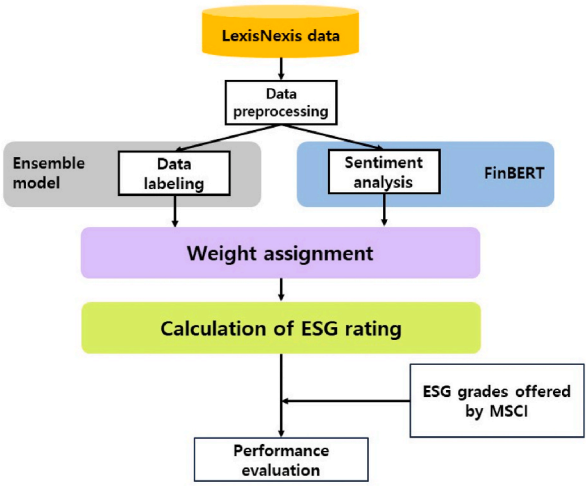


Fig. 8. Visual representation of the ESG rating framework (ESG2PreEM).

3.3.2. Assigning weights to each category (E, S, G)

The collected data were labeled using the proposed ensemble framework. Among the entire dataset, 8,217 instances were labeled as "E," 5,901 as "S," and 18,666 as "G" (Fig. 9). Under the assumption that categories appearing frequently in the overall data are more important, weights were assigned to each category based on their respective proportions [49]. The weights were determined as follows: W_G (3.1631), W_E (1.3925), and W_S (1), with G being the most frequently mentioned category.

3.3.3. Sentiment analysis with FinBERT

The data collected was labeled as E, S, or G, and sentiment analysis was conducted using FinBERT. FinBERT is a specialized model trained on financial news classification using the Reuters Thomson Reuters Text Research Collection subset dataset, which consists of 1.8 million news articles published by Reuters between 2008 and 2010. Through FinBERT, newspaper articles were labeled as positive, neutral, or negative. As a result, each newspaper article was assigned a label from E, S, or G category, along with a label of positive, neutral, or negative sentiment.

3.3.4. ESG rating calculation

Finally, to extract ESG ratings, the difference between the number of positive and negative classifications for each company's E, S, and G data was calculated. The calculated weights for E, S, and G for all companies were multiplied by the difference values and then divided by the sum of positive and negative texts. This approach considering the weights of categories has also been applied to the concept of CSR for producing objective ratings [49]. The equation for calculating each company's ESG rating is as follows:

$$ESG\ rating_{company} = \frac{W_E \times \left(P_{E\_company} - N_{E\_company}\right) + W_S \times \left(P_{S\_company} - N_{S\_company}\right) + W_G \times \left(P_{G\_company} - N_{G\_company}\right)}{T_{company}} \quad (1)$$

where W_E, W_S, and W_G is the weight calculated by the ratio of each category (based on the data of the entire companies, $P_{i\_company}$ is the number of positive data on category i, and $N_{i\_company}$ is the number of negative data on category i, respectively. $T_{company}$ is the sum of positive and negative data on the given company.

4. Results

The result section consists of a process that evaluates an ESG2PreEM that classifies the input text data into the categories of E, S, and G. Afterwards, the authors compare the calculated ESG rating with MSCI's ESG grade.

4.1. Performance of textual ESG classification model

In the current section, the authors compared the performance of the single and ensemble models that utilize language models to classify ESG. The hyperparameters used in the experiments are described in Table 2. The results of the various fine-tuned transformer models (i.e., base learners) and combinations of classifiers on the test dataset are presented in Table 3. Based on the results obtained from the experiments conducted with different batch sizes (i.e., 8, 16, 20, 24, 32), it can be concluded that the ensemble models generally perform better than the individual classifiers. Specifically, the accuracy of the ensemble model using BERT and ALBERT was found to be 80.79% with batch size 20 (Fig. 10). This suggests that combining the strengths of multiple classifiers can lead to improved performance in certain scenarios [50,51].

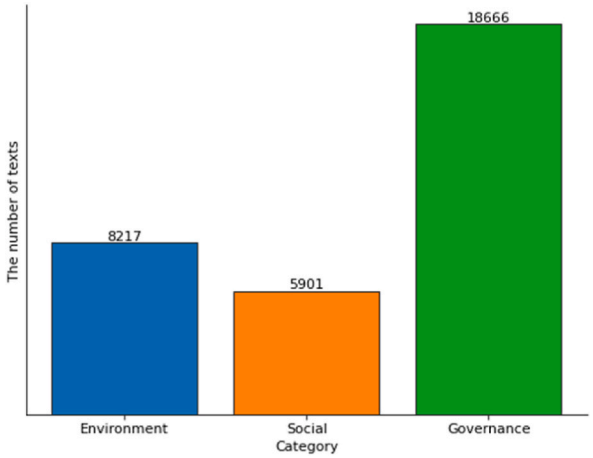


Fig. 9. Total distribution of labels for the selected companies.

Table 2
Hyperparameters used for the experiment.

Hyperparameters	Value
Batch size	8, 16, 20, 24, 32
Epoch	5
Learning rate	Warm up method
Warm up decay	0.2
Weight decay	0.01
Activation function	Gaussian Error Linear Unit (GELU)
Dropout Rate	0.1
Optimizer	AdamW

Table 3
Classification performance comparison based on different batch sizes (unit: %).

Model \ Metrics	Accuracy	Recall	Precision	F1-score
Batch size: 8				
BERT	74.58	75.00	75.00	74.00
RoBERTa	75.86	76.00	76.00	76.00
ALBERT	76.45	77.00	76.00	76.00
BERT+RoBERTa	76.55	75.00	75.00	74.00
BERT+ALBERT	77.04	75.00	75.00	74.00
RoBERTa+ALBERT	78.03	77.00	76.00	76.00
BERT+RoBERTa+ALBERT	78.23	75.00	75.00	74.00
Batch size: 20				
BERT	78.62	79.00	79.00	79.00
RoBERTa	75.37	76.00	75.00	75.00
ALBERT	76.75	77.00	77.00	77.00
BERT+RoBERTa	78.82	79.00	79.00	79.00
BERT+ALBERT	80.79	79.00	79.00	79.00
RoBERTa+ALBERT	77.83	77.00	77.00	77.00
BERT+RoBERTa+ALBERT	80.30	79.00	79.00	79.00
Batch size: 32				
BERT	76.85	79.00	77.00	77.00
RoBERTa	77.54	77.00	78.00	77.00
ALBERT	75.37	76.00	75.00	75.00
BERT+RoBERTa	78.72	79.00	77.00	77.00
BERT+ALBERT	78.13	79.00	77.00	77.00
RoBERTa+ALBERT	78.23	76.00	75.00	75.00
BERT+RoBERTa+ALBERT	79.11	79.00	77.00	77.00

4.2. Comparison of ratings attained from ESG2PreEM with MSCI

In the second process, the authors compared the ratings of 10 companies from the framework with the classifications provided by MSCI, which categorizes companies into three groups: Leader, Average, and Laggard (Table 4). Suggested framework's ESG ratings demonstrated values higher than 0.80 when the MSCI results placed the companies in the Leader group (AAA or AA). Conversely, for companies in the Average group, the rating values ranged from 0.20 to less than 0.80. This variation occurred because the Average group encompasses a broader range of ratings, such as BB, BBB, and A, which have a wider range compared to other grades.

To measure the relevance of the results, the authors calculated the proportion of results that are related to MSCI. This is described as *Precision@n* where $N_{rs@}n$ means the number of mutually related elements and n is the number of total items [52].

$$Precision@n = \frac{N_{rs@}n}{n} \tag{2}$$

This metric estimates how many items out of *n* are relevant. A high *Precision@n* indicates that the top *n* results have been accurately selected. This value ranges from 0 to 1, with 1 being the best. The results showed a high *Precision@n* of 0.9.

5. Discussion

This section discusses aspects of the analysis that overcome the challenges presented in previous studies. The limitations of existing research are the absence of ESG rating's uniformity among companies and a lack of standards for handling ESG information reporting [5]. To resolve the obstacles identified in prior research, a structure and criterion are necessary to evaluate ESG ratings in a neutral system, regardless of the data utilized. For this purpose, the authors collected news articles from LexisNexis and the performance was improved by adjusting the transformer-based language model and combining ensemble models. Overall, employing advanced NLP techniques to extract pertinent information from extensive textual data can supplement the ESG evaluation criteria established by various rating companies. This study presents an approach for utilizing and embracing text-based automated methodologies for ESG ratings. Suggested workflow aimed to make analyses related to sustainable and responsible investing more accessible and complementary.

6. Conclusion

Since the introduction of ESG at the UNGC, it has received substantial attention and emerged as a crucial assessment standard for corporations. Particularly, owing to the Covid-19 pandemic, concerns regarding ESG-related matters, including environmental and climate changes, have escalated globally. Moreover, from a long-term investment perspective, an ESG analysis has become indispensable for businesses. However, the objectivity of ESG ratings remains a topic of discussion even with the progressively increasing importance of ESG.

The authors independently categorized LexisNexis data within the ESG sector using the Refinitiv-Sustainable Leadership Monitor, demonstrating a methodology for building a dataset that enabled the suggested language model to assess ESG ratings, which was confirmed to be effective. To further validate the suggested approach, the authors utilized the proposed model for labeling and performed sentiment analysis using FinBERT on news data for 10 selected companies from the DJIA. After calculating incorporate weights for each category, the authors calculated the ESG rating and ultimately obtained results that closely resemble MSCI's Grade. Therefore, the suggested approach implements cost-effective data labeling and can be applied impartially to other scoring systems.

This study presents several limitations and directions for future research. First, ESG ratings can have a meaningful influence on corporate value and perception, leading news articles to be reluctant in using negative words or tones. In other words, it can be

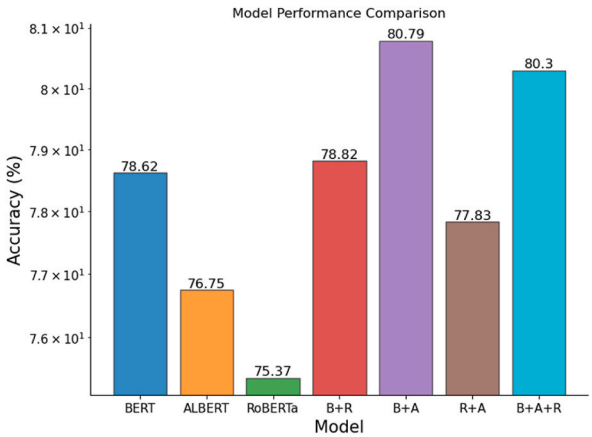


Fig. 10. Results of classifying ESG from LexisNexis data (B: BERT, R: RoBERTa, A: ALBERT).

Table 4
MSCI grades and ESG ratings obtained through ESG2PreEM.

Company	ESG rating (ESG2PreEM)	ESG grade (MSCI)
Coca	1.108038	Leader
3 M	0.962214	Leader
Cisco	0.958109	Leader
Home Depot	0.854606	Leader
McDonald	0.738103	Average
Walmart	0.685224	Average
American Express	0.497870	Leader
JPMorgan Chase	0.479188	Average
Apple	0.403684	Average
Nike	0.235148	Average

challenging to classify companies that have poor ESG practices. Hence, for future research, it is necessary to categorize ESG information by integrating data sources like social media or online forums that encompass adverse information regarding companies. Second, to strengthen the model's robustness, future research should extend the validation process to include ESG data collected from a variety of sources, ensuring a more comprehensive evaluation.

Funding

This study was supported by a National Research Foundation of Korea (NRF) (http://nrf.re.kr/eng/index) grant funded by the Korean government (RS-2023-00208278).

Data availability statement

The data associated with this study has not been deposited into a publicly available repository. However, the data supporting the findings will be made available on request.

CRediT authorship contribution statement

Haein Lee: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Seon Hong Lee:** Methodology. **Heungju Park:** Conceptualization, Data curation. **Jang Hyun Kim:** Conceptualization, Data curation. **Hae Sun Jung:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgements

We would like to thank Editage (www.editage.co.kr) for the English language editing.

References

[1] G. Hagart, I. Knoepfel, Communicating ESG Value Drivers at the Company-Investor Interface: Who Cares Wins Annual Event 2006-conference Report, World Bank Group, 2006.
[2] A.G. Hoepner, A.A. Majoch, X.Y. Zhou, Does an asset owner's institutional setting influence its decision to sign the principles for responsible investment? J. Bus. Ethics 168 (2) (2021) 389–414.
[3] S.N. Seo, Beyond the Paris agreement: climate change policy negotiations and future directions, Regional Science Policy & Practice 9 (2) (2017) 121–140.
[4] E. Dimson, P. Marsh, M. Staunton, Divergent ESG ratings, J. Portfolio Manag. 47 (1) (2020) 75–87.
[5] A. Amel-Zadeh, G. Serafeim, Why and how investors use ESG information: evidence from a global survey, Financ. Anal. J. 74 (3) (2018) 87–103.
[6] E. Escrig-Olmedo, M.Á. Fernández-Izquierdo, I. Ferrero-Ferrero, J.M. Rivera-Lirio, M.J. Muñoz-Torres, Rating the raters: evaluating how ESG rating agencies integrate sustainability principles, Sustainability 11 (3) (2019) 915.
[7] M.A. Almessiere, Y. Slimani, N.A. Algarou, M.G. Vakhitov, D.S. Klygach, A. Baykal, T.I. Zubar, S.V. Trukhanov, A.V. Trukhanov, H. Attia, M. Sertkol, I.A. Auwal, Tuning the structure, magnetic and high frequency properties of Sc-doped Sr0.5Ba0.5ScxFe12-xO19/NiFe2O4 hard/soft nanocomposites, Adv. Electr. Mater. (2022) 01124, https://doi.org/10.1002/aelm.202101124.
[8] G.Zh Moldabayeva, G.M. Efendiyev, A.L. Kozlovskiy, N.S. Buktukov, S.V. Abbasova, Modeling and adoption of technological solutions in order to enhance the effectiveness of measures to limit water inflows into oil wells under conditions of uncertainty, ChemEngineering 7 (2023) 89, https://doi.org/10.3390/chemengineering7050089.
[9] Y. Tse, Price discovery and volatility spillovers in the DJIA index and futures markets, J. Futures Mark. 19 (8) (1999) 911–930.
[10] D. Araci, Finbert: Financial Sentiment Analysis with Pre-trained Language Models, 2019 [Online]. Available: https://arxiv.org/abs/1908.10063.
[11] B. Bruder, Y. Cheikh, F. Deixonne, B. Zheng, Integration of ESG in Asset Allocation, 2019 [Online]. Available: SSRN-id3473874.

[12] D.A. Vinnik, A.Yu Starikov, V.E. Zhivulin, K.A. Astapovich, V.A. Turchenko, T.I. Zubar, S.V. Trukhanov, J. Kohout, T. Kmječ, O. Yakovenko, L. Matzui, A.S. B. Sombra, D. Zhou, R.B. Jotania, C. Singh, Y. Yang, A.V. Trukhanov, Changes in structure, magnetization and resistivity of BaFe12-xTixO19, ACS Appl. Electron. Mater. 3 (2021) 1583–1593, https://doi.org/10.1021/acsaelm.0c01081.

[13] G.Z. Moldabayeva, G.M. Efendiyev, A.L. Kozlovskiy, S.R. Tuzelbayeva, Z.B. Imansakipova, Study of the rheological characteristics of sediment-gelling compositions for limiting water inflows, Appl. Sci. 13 (2023) 10473, https://doi.org/10.3390/app131810473.

[14] V. Díaz, D. Ibrushi and J. Zhao, "Reconsidering systematic factors during the COVID-19 pandemic–The rising importance of ESG," Finance Res. Lett., vol. 38, pp. 101870. 102021.

[15] L. Widyawati, A systematic literature review of socially responsible investment and environmental social governance metrics, Bus. Strat. Environ. 29 (2) (2020) 619–637.

[16] G. Friede, T. Busch, A. Bassen, ESG and financial performance: aggregated evidence from more than 2000 empirical studies, Journal of sustainable finance & investment 5 (4) (2015) 210–233.

[17] W. Henisz, T. Koller, R. Nuttall, Five Ways that ESG Creates Value, McKinsey, 2019.

[18] A. Edmans, Does the stock market fully value intangibles? Employee satisfaction and equity prices, J. Financ. Econ. 101 (3) (2011) 621–640.

[19] A. Edmans, The link between job satisfaction and firm value, with implications for corporate social responsibility, Acad. Manag. Perspect. 26 (4) (2012) 1–19.

[20] N. Attig, S. El Ghoul, O. Guedhami, J. Suh, Corporate social responsibility and credit ratings, J. Bus. Ethics 117 (2013) 679–694.

[21] T. Guo, N. Jamet, V. Betrix, L.A. Piquet, E. Hauptmann, Esg2risk: A Deep Learning Framework from Esg News to Stock Volatility Prediction, 2020, https://doi.org/10.48550/arXiv.2005.02527 [Online]. Available:.

[22] A. Sokolov, J. Mostovoy, J. Ding, L. Seco, Building machine learning systems for automated ESG scoring, The Journal of Impact and ESG Investing 1 (3) (2021) 39–50.

[23] H. Lee, S.H. Lee, K.R. Lee, J.H. Kim, Esg discourse analysis through bertopic: comparing news articles and academic papers, Computers, Materials & Continua 75 (3) (2023) 6023–6037.

[24] N. Raman, G. Bang, A. Nourbakhsh, Mapping ESG trends by distant supervision of neural language models, Machine Learning and Knowledge Extraction 2 (4) (2020) 453–468.

[25] A. Schmidt, Sustainable News–A Sentiment Analysis of the Effect of ESG Information on Stock Prices, 2019 [Online]. Available: SSRN-id3809657.

[26] H. Lee, H.S. Jung, S.H. Lee, J.H. Kim, Robust sentiment classification of metaverse services using a pre-trained Language Model with soft voting, KSII Transactions on Internet & Information Systems 17 (9) (2023).

[27] S.F. Abbasi, Q.H. Abbasi, F. Saeed, N.S. Alghamdi, A convolutional neural network-based decision support system for neonatal quiet sleep detection, Math. Biosci. Eng. 20 (9) (2023) 17018–17036.

[28] H.S. Jung, S.H. Lee, H. Lee, J.H. Kim, Predicting Bitcoin trends through machine learning using sentiment analysis with technical indicators, Comput. Syst. Sci. Eng. 46 (2) (2023).

[29] H. Lee, S.H. Lee, D. Nan, J.H. Kim, Predicting user satisfaction of mobile healthcare services using machine learning: confronting the COVID-19 pandemic, J. Organ. End User Comput. 34 (6) (2022) 1–17.

[30] S.F. Abbasi, H. Jamil, W. Chen, EEG-based neonatal sleep stage classification using ensemble learning, Computers, Materials & Continua 70 (3) (2022).

[31] D.A. Weaver, B. Bimber, Finding news stories: a comparison of searches using LexisNexis and Google News, Journal. Mass Commun. Q. 85 (3) (2008) 515–530.

[32] G. Dorfleitner, G. Halbritter, M. Nguyen, Measuring the level and risk of corporate responsibility–An empirical comparison of different ESG rating approaches, J. Asset Manag. 16 (2015) 450–466.

[33] I.S. Popescu, C. Hitaj, E. Benetto, Measuring the sustainability of investment funds: a critical review of methods and frameworks in sustainable finance, J. Clean. Prod. 314 (2021) 128016.

[34] F. Berg, K. Fabisik, Z. Sautner, "Is History Repeating Itself? the (Un) Predictable Past of ESG Ratings," the (Un) Predictable Past of ESG Ratings (August 24, 2021), vol. 708, European Corporate Governance Institute–Finance Working Paper, 2020.

[35] Refinitiv, [Online]. Available: https://www.refinitiv.com/en/products/sustainability-reporting-on-leadership.

[36] T. Zhou, H. Jiao, Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment, Educ. Psychol. Meas. 83 (4) (2022) 00131644221117193.

[37] M. Honnibal, I. Montani, spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, To Appear 7 (1) (2017) 411–420.

[38] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (5) (1988) 513–523.

[39] L. Pérez, V. Hunt, H. Samandari, R. Nuttall, K. Biniek, Does ESG Really Matter—And Why, McKinsey Quarterly, 2022.

[40] J. Howard, S. Ruder, Universal Language Model Fine-Tuning for Text Classification, 2018, https://doi.org/10.48550/arXiv.1801.06146 [Online]. Available:.

[41] Y. Hao, L. Dong, F. Wei, K. Xu, Visualizing and Understanding the Effectiveness of BERT, 2019, https://doi.org/10.48550/arXiv.1908.05620 [Online]. Available:.

[42] A. Gillioz, J. Casas, E. Mugellini, O. Abou Khaled, Overview of the transformer-based models for NLP tasks, in: In Proc. 15th Conference on Computer Science and Information Systems, (FedCSIS), Sofia, Bulgaria, 2020, pp. 179–183.

[43] L. Diao, Z. Tang, X. Guo, Z. Bai, S. Lu, et al., Weibo disaster rumor recognition method based on adversarial training and stacked structure, KSII Transactions on Internet & Information Systems 16 (10) (2022) 3211–3229.

[44] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018, https://doi.org/10.48550/arXiv.1810.04805 [Online]. Available:.

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, et al., Roberta: A Robustly Optimized Bert Pretraining Approach, 2019, https://doi.org/10.48550/arXiv.1907.11692 [Online]. Available:.

[46] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, et al., Albert: A Lite Bert for Self-Supervised Learning of Language Representations, 2019, https://doi.org/10.48550/arXiv.1909.11942 [Online]. Available:.

[47] H. Wang, Y. Yang, H. Wang, D. Chen, Soft-voting clustering ensemble, in: Proc. Multiple Classifier Systems: 11th International Workshop, vol. 11, MCS, Nanjing, China, 2013, pp. 307–318, 2013.

[48] S. Kumari, D. Kumar, M. Mittal, An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier, International Journal of Cognitive Computing in Engineering 2 (2021) 40–46.

[49] J.S. Choi, Y.M. Kwak, C. Choe, Corporate social responsibility and corporate financial performance: evidence from Korea, Aust. J. Manag. 35 (3) (2010) 291–311.

[50] M. Jeong, N. Lee, B.S. Ko, I. Joe, Ensemble deep learning model using random forest for patient shock detection, KSII Transactions on Internet and Information Systems 17 (4) (2023) 1080–1099.

[51] M. Sevri, H. Karacan, Two stage deep learning based stacked ensemble model for web application security, KSII Transactions on Internet and Information Systems 16 (2) (2022) 632–657.

[52] M. Chen, P. Liu, Performance evaluation of recommender systems, Int. J. Perform. Eng. 13 (8) (2017) 1246.