

Construction of a Feature Dictionary and Optimization of an Artificial Neural Network for ESG Information Classification: A Case Study in Vietnam

Anh Nguyen-Ngoc-Lan¹², Anh Bui-Tuyet^{3*}, Nhu Hung Duong³, and Van Nguyen-Thi-Thu⁴.

¹Faculty of Business Administration, Van Lang University, Ho Chi Minh City, Vietnam

e-mail: anh.2483401010039@vanlanguni.vn

²Institute of Postgraduate Education, Van Lang University, Ho Chi Minh City, Vietnam

³School of Industrial Management, Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, Ho Chi Minh City, Vietnam.

e-mail: anh.buituyet@hcmut.edu.vn, e-mail: duongnhuhung@hcmut.edu.vn

⁴Faculty of Accounting & Auditing, Van Lang University, Ho Chi Minh City, Vietnam

e-mail: van.nguyen@vlu.edu.vn

Abstract— In the context of the global green transition and Vietnam's commitment to achieving Net Zero by 2050, Environmental, Social, and Governance (ESG) disclosures have become crucial. However, corporate ESG reports in Vietnam often lack quantitative data and transparency. Financial news platforms provide abundant, real-time ESG insights, but processing this massive volume manually is unfeasible. This study pioneers the application of an Artificial Neural Network (ANN) combined with Natural Language Processing (NLP) to automatically classify Vietnamese corporate news into E, S, and G pillars. Using a dataset of 210 scraped articles, we applied TF-IDF for feature extraction and optimized the ANN's hidden layer through grid search. The study successfully established a localized core dictionary of 97 Vietnamese ESG keywords. Performance evaluation showed the Environmental (E) model achieved excellent predictive capacity with 90.48% accuracy and an AUC of 1.000 at 5 hidden nodes. Although the Social and Governance models exhibited instability due to small test sample sizes, the overall framework demonstrates significant potential. This automated approach provides an objective, real-time screening tool, mitigating "greenwashing" practices, assisting investors in directing green capital, and supporting regulatory bodies in promoting sustainable corporate governance

Keywords: *ESG classification, Artificial Neuron Network (ANN), Text mining, Vietnamese financial news, Sustainable investing*

I. INTRODUCTION

In the context of globalization and green transformation is increasingly strong, ESG information (Environmental, Social, Governance) has become a key factor in assessing corporate performance and deciding on sustainable investments. Many international studies have proven that business with high ESG scores not only reduce capital costs, increase access to green capital, but also achieve superior long-term financial performance thanks to better risk management and higher brand reputation. (Kotsantonis et al., 2016), (Verheyden et al., 2016), (Yang et al., 2025).

In Viet Nam, with commitment to achieving Net Zero by 2050 and the rapid increase in sustainable investment capital flows, the need for transparency and reliability in ESG information is becoming increasingly urgent. Since 2022, The State Securities Commission of Vietnam in collaboration with the two Stock Exchanges has issued regulations the disclosure of ESG information, especially VNSI – 30 index of HOSE. Nevertheless, a significant implementation gap persists in practice: about 55% of large listed enterprises publish sustainability reports with average reporting quality remaining low (24,24/36 according to GRI Standards). Many reports are largely ceremonial and lack measurable data.

A rich yet underexploited source of data lies in the daily corporate news published on platforms such as CafeF, Vietstock, and VnExpress Kinh doanh. This is publicly available, continuously updated information source that reflects the actual ESG performance of thousands of enterprises. Manually processing such a massive volume of news is infeasible in terms of both time and cost. Therefore, the application of Artificial Neuron Network (ANN) models for automatic classification ESG news into three pillars Environmental (E), Social (S), Governance (G) has become an urgent necessity. This approach provides timely and objective information to investors, enterprises and regulatory authorities.

This study aims to this research gap by constructing and evaluating the effectiveness of an Artificial Neuron Network (ANN) model for comprehensive classification of Vietnamese language ESG news. The research is expected to enhance the transparency of sustainability information, support investment decision – making and promote the green transformation in alignment with Vietnam's National Strategy on Green Growth for the period 2021–2030.

II. LITERATURE REVIEW

The ESG framework has evolved into an important tool for assessing corporate social responsibility and long-term business performance. (Kotsantonis & Serafeim, 2019) and (Verheyden et al., 2016) demonstrated that firms with strong ESG practices often gain competitive advantage in the capital

market, labor market and product market. More recently, (Xiao & Xiao, 2025) through a survey of 200 managers from Chinese state – owned enterprises, confirmed that AI plays a significant mediating role between technology adoption and improved ESG performance. Similarly, (Zhou et al., 2025) developed deep learning model for the new energy sector and provided evidence that integrating ESG factors enhance both financial performance and corporate governance.

In Europe, (Moreno & Caminero, 2022) analyzed Pillar 3 report of banks and found that awareness of ESG risks has increased, yet the level of detail remains limited. (Busco et al., 2020) evaluated reports following SASB standards, emphasized the ESG financial materiality of ESG issues. These studies indicate that ESG reporting shifting from voluntary to mandatory, however significant gap still exists in term comparability and data quality.

Text mining and Natural Language Processing (NLP) have been effectively applied to extract ESG information from corporate reports and news. For example, Kiri and Nozaki (Kiri & Nozaki, 2020) developed text mining model to visualize ESG activities based on lexical structure within CSR reports in Japan. In the research of (Yoon et al., 2023), the authors utilized bigram, TF-IDF, semantic network and CONCOR and analyses to examine ESG reports in South Korea in comparison with the SDGs. In (Jiang et al., 2022), the authors proposed hybrid text mining approach to simplify the ESG taxonomy, integrating automated and manual methods to enhance comparability. The above studies have demonstrated that text mining can facilitate the rapid identification of ESG trends and topics; however, they have mainly focused on official reports in English.

Artificial Neuron Network (ANN) and machine learning models have demonstrated efficacy in ESG analysis. (Cheong et al., n.d.) employed ANN to explore the non – linear relationship between AI capabilities and ESG scores. (Du et al., 2024) developed an ESG lexicon by utilizing Word2vec from reports and news to measure ESG risks. These studies indicated that while ANN excel in semantic processing, they remain largely untested in context of comprehensive Vietnamese digital news encompassing all three E-S-G pillars.

In Vietnam, while ESG has emerged as a strategic priority, actual disclosure practices remain limited. (Trần Ngọc, 2023) and (Thi Thu Thao, 2023) point out that many listed enterprises engage in merely symbolic ESG disclosure, which lack quantitative data and independent assurance. A survey of 169 enterprises by (Thi Phuong Nguyen et al., 2024) reveals that accounting proficiency, female board representation and information technology systems positive influence ESG reporting readiness, conversely, leadership awareness and government guidelines have not demonstrate a significant impact. (Anh et al., 2024) has examined top 100 listed companies on HOSE and concluded that financial indicators such as ROE and AT exert a positive effect on ESG reporting quality, whereas the Debt – to – Equity (DE) ratio has negative impact.

Although numerous studies have examined ESG text mining (Kiri & Nozaki, 2020), (Yoon et al., 2023) and the application of ANN in ESG analysis (Cheong et al., n.d.), there are remain

lack of comprehensive research applying ANN models to classify Vietnamese ESG news into three pillars Environmental (E), Social (S), Governance (G) based on daily digital news source in Vietnam. This study aims to fill that research gap.

III. RESEARCH DATA AND METHODOLOGY

A. Data Source and data Preprocessing

This study employs two principal data sources covering 2020 – 2025: (1) digital corporate the level of detail remains limited news and (2) sustainable reports of listed enterprises in Vietnam. The news articles were extracted from prominent financial journalism platforms, namely CafeF, Vietstock, VnExpress Kinh doanh, Báo Đầu tư and NDH. The second source comprises the sustainability reports of over 40 companies listed on HOSE. The data collection process yielded a total of 210 documents.

Data were gathered via web scraping techniques complemented by manual verification to ensure data integrity. Each news article and excerpt from the sustainability reports was manually labeled into three distinct classes: E (Environmental), S (Social) and G (Governance). The labeling process was independently executed by three ESG domain experts in accordance with GRI Standard framework. To guarantee the robustness of dataset, inter – rater reliability was assessed and any coding discrepancies were resolved through mutual consensus prior to finalization. The raw text was subsequently preprocessed. The preprocess steps involved the removal of special characters, lowercasing, stop – word elimination and Vietnamese word segmentation. Finally, the text data was transformed into feature vectors utilizing the Term Frequency-Inverse Document Frequency (TF-IDF) technique.

Table 1: Distribution of ESG Labels in the Augmented Dataset

Label	Number of documents	Percentage %
Environmental (E)	19	6%
Social (S)	83	28%
Governance (G)	229	77%

B. Artificial Neuron Network model (ANN)

A neural network model can be illustrated via the following equation

$$y = f^{(1)} \left(\sum_{j=1}^n \alpha_j^{(1)} \cdot f_i^{(0)} \left(\sum_{i=1}^d \alpha_{ij}^{(0)} x_i + \beta_j^{(0)} \right) \right) + \beta^{(1)}$$

where,

x_i : Input representing the TF-IDF features of the news article

$\alpha_{ij}^{(0)}$: Connection weights from the input layer to the hidden layer

$f_j^{(0)}$: Hidden layer activation function (ReLU or tanh)
 $\alpha_j^{(1)}$: Connection weights from the hidden layer to the output layer
 $f^{(1)}$: Sigmoid activation function for the 4 output classes (E, S, G, O)
 $\beta_j^{(0)}, \beta^{(1)}$: Bias terms

The neural network was trained on the labeled dataset, utilizing the categorical cross-entropy loss function and the Adam optimizer

C. Training and Evaluation of the ANN Model

To mitigate overfitting and ensure an objective evaluation, the entire dataset was partitioned into three distinct subsets. The training set (60% of the data) was explicitly utilized to train the neural network to recognize the characteristic patterns of ESG news. The validation set (30% of the data) was employed to fine-tune the model's hyperparameters. The test set (10% of the data) served as a strictly independent hold-out dataset, reserved for the final phase to measure the system's empirical performance.

The ANN models for the E, S, and G pillars were trained over 200 epochs, utilizing a learning rate of 0.01 and a momentum of 0.9. The architecture was configured with a single hidden layer, and the optimal number of hidden nodes was determined via a grid search strategy, exploring a range from 2 to 10 nodes. The optimally configured network for each respective pillar was subsequently evaluated on the 10% test set. The evaluation metrics comprised the Area Under the Curve (AUC) and Accuracy, where values approaching 1.0 and 100%, respectively, indicate superior predictive capacity.

IV. RESULT AND DISCUSSION

A. Results of Data Preprocessing and Feature Dictionary Construction

The Natural Language Processing (NLP) pipeline for Vietnamese financial news data requires careful refinement due to the language's complex compound word structure. We initially performed text segmentation into tri-grams, however, this approach resulted in an excessively large vector space with substantial noise. By applying the Information Gain weighting method to calculate feature importance, we successfully extracted the top of 2000 highest weighted terms. These terms were then manually refined and synthesized into a core feature dictionary consisting of 97 precise keywords that comprehensively represent the three ESG pillars: E (Environmental), S (Social) and G (Governance). The successful construction of this 97 keywords domain – specific lexicon holds significant importance for the Vietnamese market. It enables developing automated ESG assessment based on machine learning and artificial intelligence, rather than manual assessment as earlier.

B. Optimization Results and ANN Evaluation on the Validation Set

To identify the optimal model architecture for ESG news classification, we evaluated the performance of models on the validation set (accounting 30% of the total dataset) to finetune the number of neurons in the single hidden layer, with the search range varying from 2 to 10 neurons. The algorithm was train for 200 epochs using a learning rate of 0.01 and a momentum of 0.9.

Based on the optimization log, the study identified the optimal configuration for each ESG pillar according to the highest AUC value. Specifically, the best – performing model for Environmental (E) news classification utilized 5 neurons in the hidden layer, while optimal model for Social (S) news required 6 neurons and the Governance (G) news model achieved the highest performance with 9 neurons. This differentiation reflects the inherent characteristics of real – word ESG information. Environmental (E) news typically employs direct, highly quantitative terminology (e.g., emissions volumes, renewable energy sources and waste quantities), enabling the algorithm to effectively extract features using a relatively compact network architecture (5 neurons). In contrast, Governance (G) information – such as board structure, business ethics, anti – corruption practices and risk management – generally has a higher degree of abstraction is contextually complex and intertwined with multiple legal dimensions. Consequently, the neural network requires an expanded representational capacity (9 neurons) to adequately capture the complex nonlinear of governance – related language. This phenomenon may be attributed to the small sample size of the test set, which does not yet fully represent the data distribution.

Table 2: Presents the Detailed Variations in the Area under the Curve (AUC) across each of the three ESG pillars.

Hidden Neurons	AUC_E	AUC_S	AUC_G
2	0.844	0.768	0.839
3	0.877	0.794	0.828
4	0.902	0.783	0.806
5	0.910	0.779	0.844
6	0.877	0.817	0.850
7	0.885	0.792	0.806
8	0.885	0.791	0.783
9	0.885	0.764	0.889
10	0.902	0.753	0.844

C. Performance Evaluation on Test Set

To confirm the practical value of the models, the best – performing configurations were independently applied to the test dataset (comprising 10% of the total data), which the neural network had ever been exposed to during the entire training process.

Table 3: Performance of the ANN Model on the Test Set

Models	ACC	AUC
--------	-----	-----

ANN_E (5 neurons)	90.48%	1.000
ANN_S (6 neurons)	57.14%	0.605
ANN_G (9 neurons)	80.95%	0.100

The out – of – sample test results demonstrate that the Environmental model (ANN_E_5) maintains excellent predictive power, achieving an accuracy of 90.48% and a perfect AUC of 1.000. This is an extremely promising signal, indicating that the algorithm has accurately captured the core environmental terminology in corporate media coverage. The Governance model (ANN_G_9) attained a high accuracy of 80.95%. However, its AUC score was only 0.100. The Social model (ANN_S_6) exhibited moderate performance, recording an accuracy of 57.14% and an AUC of 0.605, which reflects the inherently ambiguous and difficult – to – quantify nature of news related to corporate social and community activities. Overall, the test set results indicate that with the exception of the E model, the S and G model exhibit instability. Even the perfect AUC value of 1.000 for the E model raises questions about the model's absolute effectiveness. This phenomenon is primarily attributed to the small sample size of the test set, which limits the quantitative characteristics represented in the data distribution observed in the training and validation sets.

V. Conclusions

This study has pioneered a new approach for the automatic classification of Vietnamese financial news into the three ESG pillars Environmental (E), Social (S), Governance (G) in Vietnamese using text mining and neural network model. Based on dataset of 210 news articles and reports tokenized via TF – IDF, the research successfully constructed a core domain – specific dictionary comprising 97 native Vietnamese keywords that are highly representative of the Vietnam market. The hyperparameters optimization process revealed that the network architecture requires distinct level of structural flexibility: quantitative Environmental information requires only 5 neurons in the hidden layer, whereas the higher degree of abstraction inherent in Social and Governance information necessitates more complex architectures with 6 and 9 neurons, respectively.

Performance evaluation on the independent test set demonstrated that the Environmental news classification modal (ANN_E_5) achieved excellent results, with an accuracy of 90.48% and a perfect AUC scores 1.000. However, the Social (S) and Governance (G) models have not yet attained the desire stability, as indicated by their relatively low AUC values (0.605 and 0.100, respectively). This limitation is primarily attributable to the current small size of the test sample, which fails to fully capture the complex distributional characteristics observed in the training set. Nevertheless, the study has established a robust technology foundation. Future efforts to enrich the dataset are expected to further refine the model, ultimately enabling the development of a fully automated, objective and transparent ESG evaluation framework for the Vietnamese stock market.

ACKNOWLEDGMENT

We acknowledge financial support from the Postgraduate Scholarship of Van Lang University.

REFERENCES

- [1] Anh, N. T. M., An, N. T., Ngoc, N. T. M., & Ngoc Xuan, V. (2024). The Influence of Financial Indicators on Vietnamese Enterprise's Sustainability Reports Disclosing Process. *Journal of Risk and Financial Management*, 17(4), 146. <https://doi.org/10.3390/jrfm17040146>
- [2] Busco, C., Consolandi, C., Eccles, R. G., & Sofra, E. (2020). A Preliminary Analysis of SASB Reporting: Disclosure Topics, Financial Relevance, and the Financial Intensity of ESG Materiality. *Journal of Applied Corporate Finance*, 32(2), 117–125. <https://doi.org/10.1111/jacf.12411>
- [3] Cheong, T. S., Shum, W. Y., & Ma, N. (n.d.). *Harnessing Artificial Neural Networks for ESG Performance Analysis: A Study of AI Capabilities and Moderating Factors*.
- [4] Du, F., Bao, C., Meng, Q., & Hui, Y. (2024). Text-Based Measure of ESG Risk Exposure. *Procedia Computer Science*, 242, 693–700. <https://doi.org/10.1016/j.procs.2024.08.142>
- [5] Jiang, L., Gu, Y., & Dai, J. (2022). ESG Taxonomy Simplification: A Hybrid Text Mining Approach. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4201462>
- [6] Kiri, T., & Nozaki, M. (2020). A Text Mining Model to Evaluate Firms' ESG Activities: An Application for Japanese Firms. *Asia-Pacific Financial Markets*, 27(4), 621–632. <https://doi.org/10.1007/s10690-020-09309-1>
- [7] Kotsantonis, S., Pinney, C., & Serafeim, G. (2016). ESG Integration in Investment Management: Myths and Realities. *Journal of Applied Corporate Finance*, 28(2), 10–16. <https://doi.org/10.1111/jacf.12169>
- [8] Kotsantonis, S., & Serafeim, G. (2019). Four Things No One Will Tell You About ESG Data. *Journal of Applied Corporate Finance*, 31(2), 50–58. <https://doi.org/10.1111/jacf.12346>
- [9] Moreno, Á. I., & Caminero, T. (2022). Analysis of ESG disclosures in pillar 3 reports. A text mining approach. *A Text Mining Approach (March 8, 2022)*. Banco de Espana Occasional Paper, (2204).
- [10] Thi Phuong Nguyen, D., Thi Huong Nguyen, L., Thi Mai Nguyen, A., & Le Thanh Phan, L. (2024). Factors affecting the readiness for ESG reporting in Vietnamese enterprises. *Problems and Perspectives in Management*, 22(3), 263–275. [https://doi.org/10.21511/ppm.22\(3\).2024.21](https://doi.org/10.21511/ppm.22(3).2024.21)
- [11] Thi Thu Thao T. (2023). ESG Trend—The History of ESG Concept Formation and the Current Trends in the Goal of Business Sustainable Development. *VNU Journal of Science: Policy and Management Studies*, 39(2). <https://doi.org/10.25073/2588-1116/vnupam.4385>
- [12] Trần Ngọc H. (2023). ESG trong môi trường bất định Covid-19: Nghiên cứu thực nghiệm tại các doanh nghiệp Việt Nam. *Tạp chí Kinh tế và Phát triển*, 44–53. <https://doi.org/10.33301/JED.V1900>
- [13] Verheyden, T., G. Eccles, R., Feiner, A., & Partners, A. (2016). Governance and Climate Change: A Success Story in Mobilizing Investor Support for Corporate Responses to Climate Change. *Journal of Applied Corporate Finance*, 28(2), 29–33. <https://doi.org/10.1111/jacf.12171>
- [14] Xiao, Y., & Xiao, L. (2025). The impact of artificial intelligence-driven ESG performance on sustainable development of central state-owned enterprises listed companies. *Scientific Reports*, 15(1), 8548. <https://doi.org/10.1038/s41598-025-93694-y>
- [15] Yang, S. S., Huang, J.-W., Chen, H.-Y., & Tsay, M.-H. (2025). Detecting corporate ESG performance: The role of ESG materiality in corporate financial performance and risks. *The North American Journal of Economics and Finance*, 76, 102370.
- [16] Yoon, J., Han, S., Lee, Y., & Hwang, H. (2023). Text Mining Analysis of ESG Management Reports in South Korea: Comparison With Sustainable Development Goals. *Sage Open*, 13(4), 21582440231202896. <https://doi.org/10.1177/21582440231202896>
- [17] Zhou, X., Peng, Y., Sun, X., Cao, X., Wang, Z., & Zhang, J. (2025). Advancing new energy industry quality via artificial intelligence-driven integration of ESG principles. *Humanities and Social Sciences Communications*, 12(1), 1491. <https://doi.org/10.1057/s41599-025-05800-0>