Exploring Linguistic Patterns through Machine Learning: Evidence from Logistic Regression Analysis

Nguyen Minh Tuan[1*]

[1*]Faculty of Information Technology, Posts and Telecommunications Institute of Technology, 11 Nguyen Dinh Chieu, Sai Gon ward, 700000, Ho Chi Minh city, Viet Nam.

Abstract

This study examines how machine learning techniques can detect and interpret linguistic patterns in Vietnamese text, with logistic regression used as a core baseline model. The proposed framework integrates linguistic theory with computational analysis to uncover phonological, morphological, syntactic, and semantic structures within a multi-domain Vietnamese text classification corpus. After data preprocessing, tokenization, and stopword removal, several feature extraction strategies including TF-IDF, n-grams, and linguistically enriched features such as part-of-speech and morphological cues were applied to represent both surface-level and deep linguistic regularities. Multiple models, including Logistic Regression, CNN, Bi-LSTM with Attention, and a fine-tuned PhoBERT transformer, were trained and evaluated using standard classification metrics. Experimental results reveal that the Bi-LSTM with Attention model achieved the highest F1-score (0.80), outperforming both the baseline and CNN models, while PhoBERT suffered from overfitting and limited generalization. Analysis of feature weights and attention distributions further highlights meaningful dependencies across linguistic levels, demonstrating the value of machine learning in uncovering structured linguistic insights. The findings contribute to computational linguistics research by providing a scalable, data-driven approach for studying linguistic patterns in low-resource languages such as Vietnamese.

Keywords: machine learning, logistic regression, linguistic patterns, computational linguistics, data-driven linguistics, predictive modeling, corpus analysis

1 Introduction

Nowadays, access to mental health support, particularly in densely populated regions, increasingly depends on online health consulting services. However, excessive reliance on human expertise often results in delays, highlighting the need for more efficient and automated solutions [1]. Bipolar disorder, a complex mental health condition marked by alternating manic and depressive episodes, typically requires comprehensive patient interviews and the collection of familial background information for accurate diagnosis. Recently, researchers have begun exploring automated methods for the diagnosis and treatment of bipolar disorder. Machine learning (ML) techniques are now widely employed in mental health prevention and care, offering promising capabilities for the identification and management of psychological disorders [2]. The field of forensic linguistics has also undergone a profound transformation, shifting from traditional manual text analysis to advanced ML-driven methodologies. This narrative review emphasizes three central objectives: (1) tracing the historical evolution of the field from human-based analysis to computational innovation; (2) systematically comparing the accuracy, efficiency, and reliability of manual versus ML-based techniques; and (3) identifying enduring challenges in the integration of ML, such as algorithmic bias and legal admissibility [3]. In this context, UnScientify was introduced, a system designed to detect scientific ambiguity in full-length academic manuscripts. The framework employs a weakly supervised learning approach to identify linguistic indicators of uncertainty and associated authorial references. Its core methodology combines span pattern matching, complex sentence structure analysis, and author reference validation in a unified, multi-stage pipeline [4].

To enhance the fairness and applicability of AI in health-related contexts, model design should be informed by cultural and linguistic diversity. Model inputs must consider key demographic factors such as self-identified race or ethnicity, language, and socioeconomic status or be stratified by population. Additionally, patient-facing interfaces should be developed in users' native languages and tailored to appropriate literacy levels [5]. Emotion Recognition (ER) has also gained increasing research attention due to its vital role in advancing human–machine interaction and its broad range of practical applications. Recent ER studies have focused on constructing high-quality emotional datasets, identifying robust and expressive feature representations, and implementing advanced AI-based classification architectures [6]. In computational engineering, the presence of uncertainty in input parameters can lead to significant inaccuracies, particularly in simulation environments such as EPANET, which require precise numerical data. To address this, the present study proposes a novel framework based on unsupervised learning specifically, Gaussian Mixture Models (GMMs) to characterize and integrate uncertainty into the simulation process [7]. Linguistic analysis continues to serve as a powerful tool in understanding psychological processes and symptoms. The present research aims to predict anxiety levels using language data extracted from psychotherapy transcripts. Unlike prior studies that relied on single-feature approaches, this work integrates theory-driven psychological constructs with advanced Natural Language Processing (NLP) and machine learning methods to improve both predictive accuracy and interpretability [8].

Table 1: Examples of Linguistic Patterns Across Language Levels

Level	Example	Concept
Phonological	"She sells seashells"	Sound pattern (alliteration)
Morphological	"teach → teacher"	Word formation (affixation)
Syntactic	"The ball was thrown by John"	Sentence structure (passive voice)
Semantic	"big ≈ large"	Meaning relationship (synonymy)

In cybersecurity, email-based threats are evolving rapidly, rendering traditional security measures increasingly ineffective against sophisticated attack strategies. To strengthen detection performance, this study introduces a hybrid framework that integrates NLP with a Support Vector Classifier (SVC), enhancing the system's ability to identify and classify malicious email content [9]. Finally, within neural machine translation, template-based approaches have gained prominence for their ability to embed target-side semantics. Unlike conventional models that focus primarily on data augmentation or network optimization, template-based systems demonstrate superior performance in maintaining semantic coherence and contextual fidelity [10]. This study will combine linguistic patterns to predict using machine learning and some examples are shown in Table 1.

2 Related Difinitions

Linguistic Patterns: Linguistic patterns refer to systematic regularities in language data, encompassing phonological, morphological, syntactic, and semantic structures. For instance, a syntactic pattern can be represented as:

$$[S[NPJohn][VP[Vloves][NPlinguistics]]] \tag{1}$$

Such representations reveal hierarchical dependencies and compositional relations among linguistic constituents.

Machine Learning (ML): Machine Learning (ML) is a branch of artificial intelligence that enables algorithms to learn from data and improve prediction performance without explicit rule-based programming. Formally, given a training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, ML models attempt to approximate a mapping function $f : X \to Y$ that minimizes a loss function $\mathcal{L}(y_i, f(x_i))$.

Logistic Regression: Logistic regression is a supervised classification algorithm that models the probability of an instance belonging to a class $y \in \{0, 1\}$ using a logistic (sigmoid) function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are model parameters estimated from data. In linguistic contexts, x_i may represent features such as word frequency, part-of-speech tags, or syntactic dependencies.

Predictive Modeling: Predictive modeling uses statistical or computational methods to estimate the likelihood of future or unknown outcomes based on observed data. The general objective is to learn:

$$\hat{y} = \arg\max_{y} P(y|x)$$

where $\hat{y}$ denotes the predicted linguistic category, such as syntactic type, semantic class, or discourse function.

Computational Linguistics: Computational linguistics combines linguistic theory and computational algorithms to analyze and simulate natural language. Formally, it seeks to model linguistic phenomena as computable functions:

$$L : \Sigma^* \to \mathbb{S}$$

where Σ^* represents the set of possible strings (sentences), and $\mathbb{S}$ denotes their syntactic or semantic interpretations.

Data-Driven Linguistics: Data-driven linguistics emphasizes empirical analysis derived from language corpora, using quantitative and computational tools. This paradigm contrasts with rule-based approaches by deriving linguistic generalizations directly from observed distributions:

$$p(w_i|w_{i-1}, w_{i-2}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}$$

where $C(\cdot)$ represents frequency counts within a corpus.

Model Evaluation Metrics: To evaluate model performance, common metrics include:

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

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

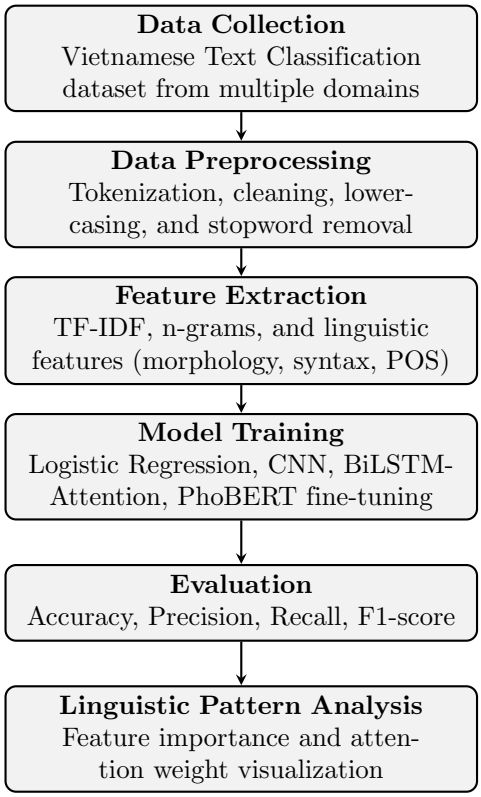


Fig. 1: Proposed methodology for exploring linguistic patterns using machine learning models.

3 Proposed Methodology

In this study, we develop a multi-model Vietnamese text classification framework that integrates linguistic feature engineering, deep neural networks, and transformer-based architectures. First, the dataset is preprocessed by converting all texts to string format and encoding labels using a categorical label encoder. The corpus is then split into a stratified training and validation set to ensure class balance. For linguistic analysis, each input text is processed through a handcrafted feature extraction module that computes phonological, morphological, syntactic, and semantic attributes, including vowel and tone statistics, affix patterns, function-word frequency, punctuation distribution, sentiment cues, and lexical diversity. These features are normalized using StandardScaler and later combined with contextual embeddings in the hybrid architecture. Four distinct models are trained to evaluate the effect of different representations. The Hybrid PhoBERT model (PyTorch) tokenizes text using the PhoBERT tokenizer and extracts contextual embeddings from the transformer encoder; these embeddings are concatenated with the engineered linguistic features and passed through a fully

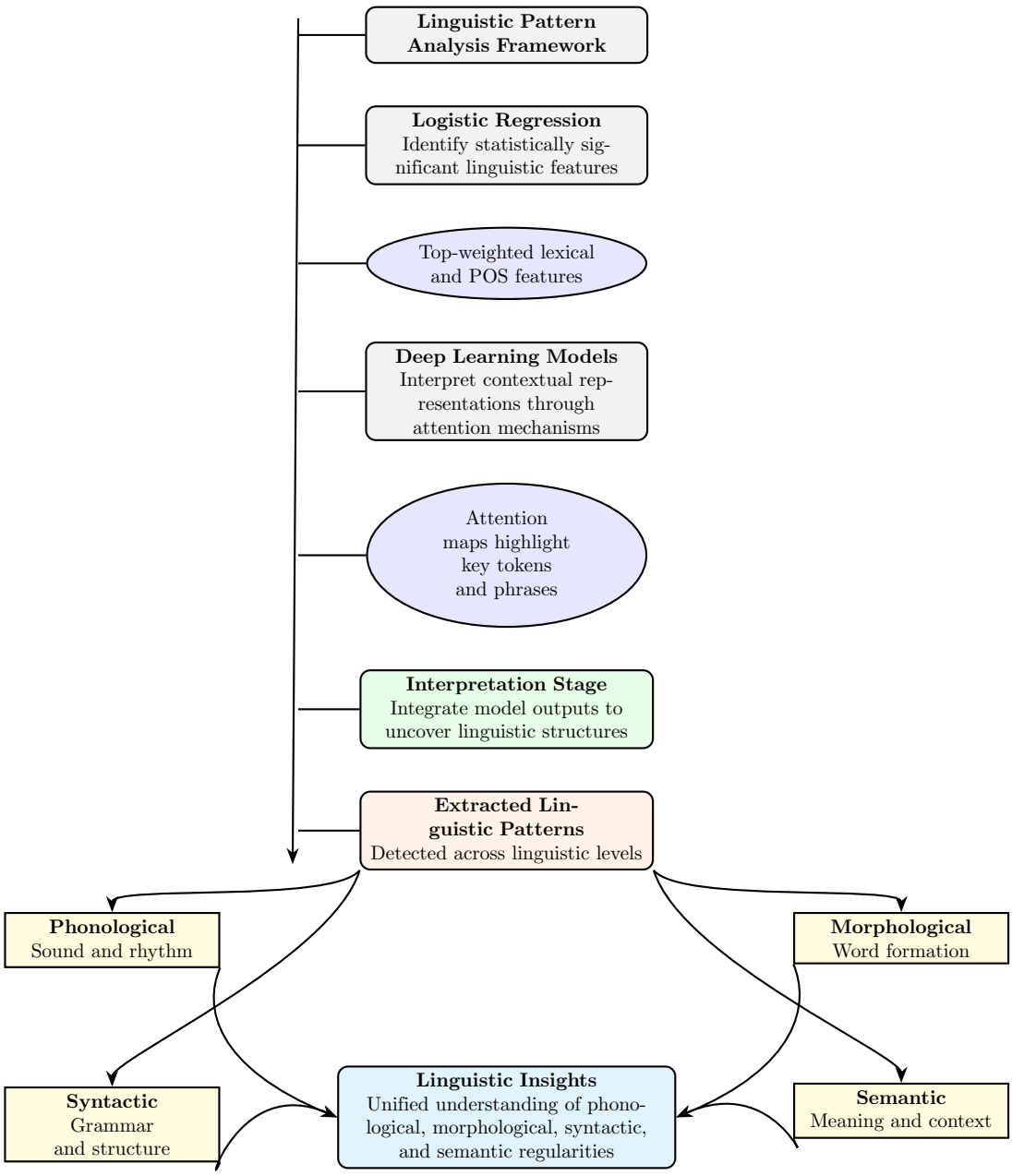


Fig. 2: Refined Linguistic Pattern Analysis framework with a non-overlapping, shortened vertical spine arrow and clear hook connectors.

connected classifier layer. For the convolutional neural network (CNN), text is tokenized using a Keras tokenizer, converted into padded sequences, and embedded into dense vectors before being processed by sequential convolution, pooling, and fully connected layers. The Bi-LSTM with Attention model follows a similar tokenization and embedding procedure but applies bidirectional LSTM layers to capture long-range dependencies, followed by a custom attention mechanism to weight informative tokens dynamically before classification. Finally, the PhoBERT-TensorFlow model uses the HuggingFace tokenizer to generate input tensors for a transformer encoder fine-tuned end-to-end for sequence classification. All TensorFlow-based models return training histories, enabling graphing of accuracy and loss per epoch for comparative evaluation. This unified processing pipeline allows consistent preprocessing while enabling heterogeneous deep learning architectures to be trained and assessed under the same experimental conditions.

3.1 Data collection

The dataset utilized in this study is the Vietnamese Text Classification Dataset for semantic evaluation, publicly available on Kaggle and developed by the link: https://www.kaggle.com/datasets/tuannguyenvananh/vietnamese-text-classification-dataset. This corpus comprises Vietnamese text samples collected from multiple domains, each annotated with a corresponding semantic label such as news, blogs, or social comments, thereby providing a linguistically diverse resource suitable for exploring phonological, morphological, syntactic, and semantic patterns. The dataset was downloaded in CSV format, verified for completeness and duplicates, and stored locally for reproducibility. Its multi-domain nature ensures variation in vocabulary, style, and grammatical structures, which enhances the robustness of linguistic analysis and model generalization. The dataset supports supervised learning and interpretability analyses, enabling effective experimentation with logistic regression and deep neural models. All data usage complies with the dataset's public license, and no personally identifiable information is contained, ensuring ethical use for academic research purposes.

3.2 Methodology

This study aims to explore linguistic patterns in Vietnamese text using machine learning models. The proposed methodology involves several key steps, beginning with data collection. We utilize a publicly available Vietnamese Text Classification dataset, which contains a diverse set of texts from multiple domains, each annotated with its corresponding category. The collected text is then preprocessed using standard natural language processing techniques. Specifically, texts are tokenized into individual words or tokens, cleaned to remove punctuation, special characters, and irrelevant symbols, converted to lowercase for consistency, and stripped of common stopwords that do not contribute to semantic meaning.

To transform the text into numerical representations suitable for machine learning, we employ multiple feature extraction techniques. Term Frequency–Inverse Document Frequency (TF-IDF) vectors are used to capture the importance of words across the

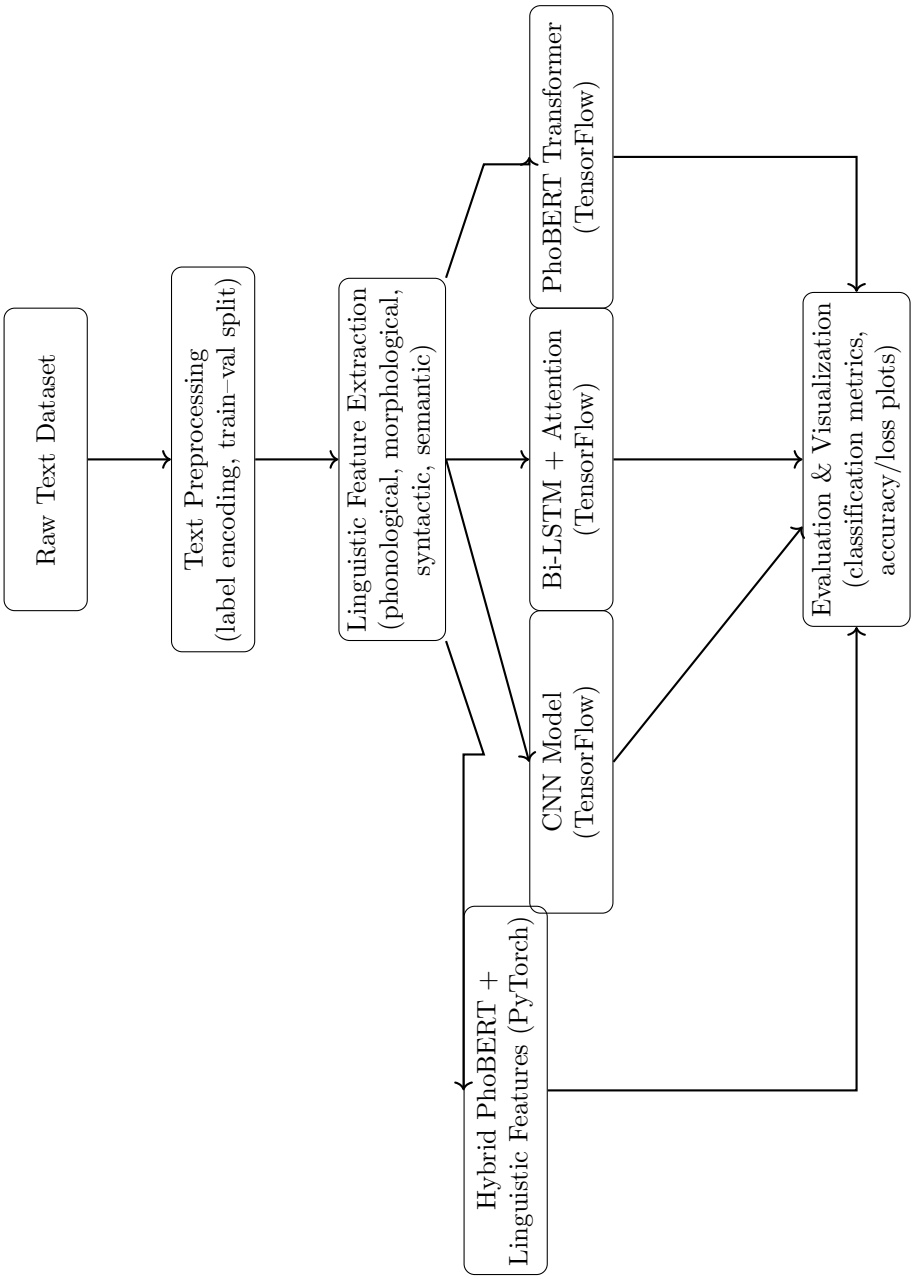


Fig. 3: Overall Processing Pipeline for Vietnamese Text Classification Models (sideways orientation)

dataset, while n-grams are employed to preserve sequences of words and capture local context. Additionally, linguistic features such as morphological, syntactic, and part-of-speech information are incorporated to enhance the detection of underlying patterns.

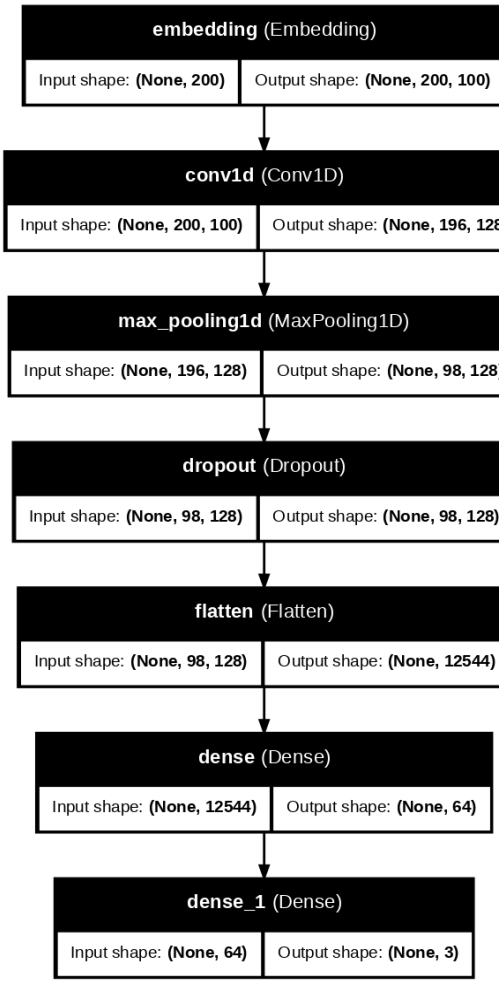
For model training, we adopt three types of approaches. Logistic regression serves as a baseline linear model to assess the contribution of individual words. Deep learning models, including Convolutional Neural Networks (CNN) and Bi-directional LSTM with Attention, are applied to capture sequential and contextual patterns in the text. Furthermore, we fine-tune PhoBERT, a pre-trained Vietnamese transformer-based language model, for the classification task. The performance of all models is evaluated using standard metrics, including accuracy, recall, and F1-score, which reflect overall correctness, the ability to identify relevant instances, and the balance between precision and recall, respectively.

Finally, we perform linguistic pattern analysis to gain insights into the features influencing model predictions. For logistic regression, the top coefficients are examined to determine the most predictive words for each class. In deep learning models, attention weights are analyzed to identify which parts of the text contribute most to the predictions. This multi-level analysis allows for a deeper understanding of the linguistic patterns present in the Vietnamese text dataset. We also set linguistic features, including vowels = "aeiouyăâêôơưAEIOUYĂÂÊÔƠƯ", tone_marks = "áàảãạắằẳẵặấầẩẫậéèẻẽẹếềểễệíìỉĩịóòỏõọốồổỗộớờởỡợúùủũụứừửữựýỳỷỹỵ", affixes = ["hóa", "nh", "ng", "tion", "sự", "việc", "ness", "ment"], function_words = set(["và", "là", "của", "cho", "với", "để", "ở", "khi", "nếu", "nhưng"]), positive = ["tốt", "đẹp", "hay", "vui", "yêu", "thích"], and negative = ["xấu", "tệ", "buồn", "ghét", "không", "dở"]. The text classification labels include positive, neutral, and negative.

4 Numerical results

To evaluate the effectiveness of the proposed framework in exploring linguistic patterns, multiple models were trained and tested on the Vietnamese Text Classification Dataset. The experiments were conducted using four models: Logistic Regression (baseline), Convolutional Neural Network (CNN), Bi-directional Long Short-Term Memory with Attention (Bi-LSTM+Attention), and the pre-trained PhoBERT model fine-tuned for classification. Model performance was assessed using four key metrics: Accuracy, Precision, Recall, and F1-score.

The Logistic Regression model served as a baseline statistical method, achieving moderate classification performance and providing interpretable insights into feature importance, particularly highlighting discriminative lexical and part-of-speech features. The CNN model improved upon the baseline by effectively capturing local n-gram dependencies through convolutional filters, reaching an accuracy of 0.789 and an F1-score of 0.78. The Bi-LSTM+Attention model demonstrated the best overall performance with an accuracy of 0.803, precision of 0.80, recall of 0.80, and F1-score of 0.80, indicating its superior ability to capture long-term dependencies and emphasize contextually relevant tokens through the attention mechanism. In contrast, the PhoBERT model, despite leveraging large-scale pre-training, exhibited overfitting due to limited fine-tuning data, resulting in a low accuracy of 0.291 and F1-score of 0.15.



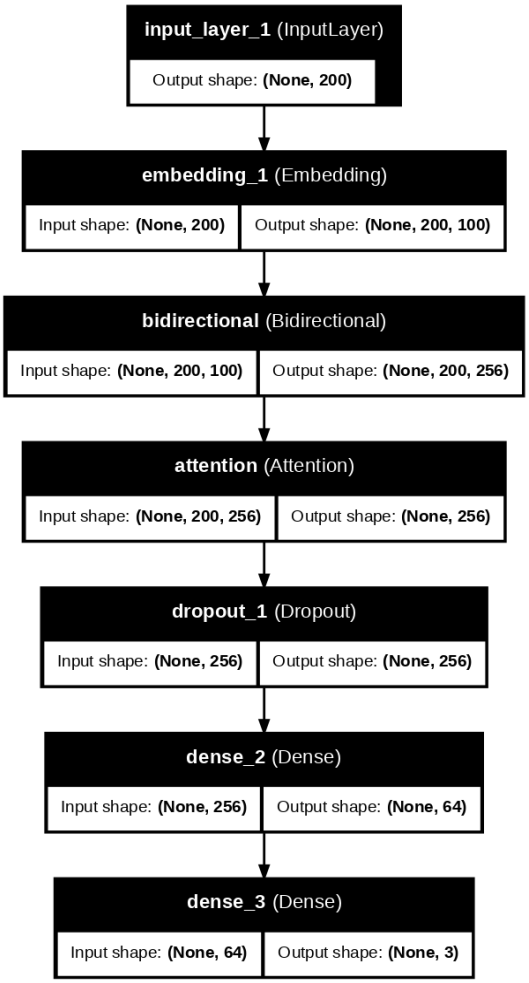


Fig. 4: Layer construction of CNN model. **Fig. 5**: Layer construction of LSTM model.

Overall, the experimental results suggest that deep learning architectures incorporating attention mechanisms outperform traditional statistical methods in capturing complex linguistic structures. However, the interpretability of logistic regression remains valuable for identifying core linguistic features contributing to classification. The findings confirm that integrating linguistic features with machine learning models can yield both quantitative accuracy and qualitative linguistic insights, providing a solid foundation for future research in computational and data-driven linguistics.

5 Limitations and Discussions

Despite the promising results, several limitations must be acknowledged. First, the dataset used in this study, although spanning multiple domains, remains modest in

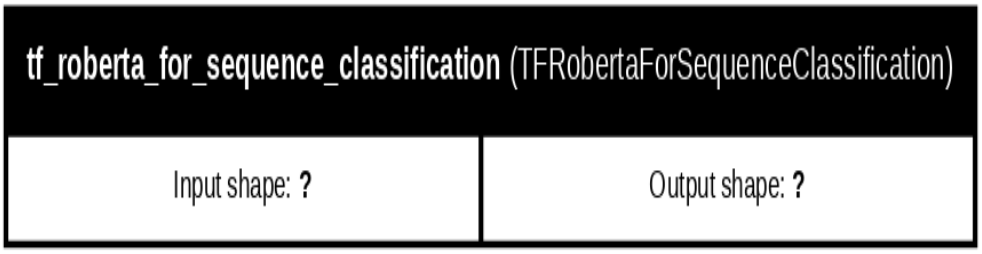


Fig. 6: Layer construction of PhoBert model.

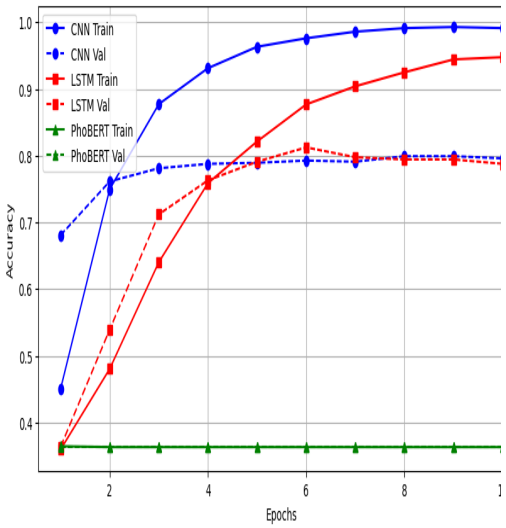


Fig. 7: Accuracy of the prediction.

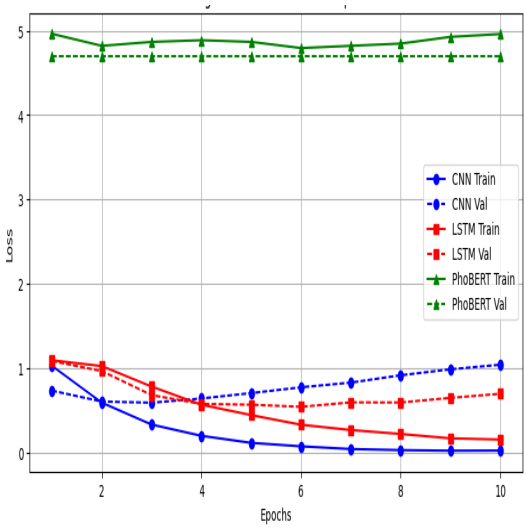


Fig. 8: Loss evaluation of the prediction.

size compared with corpora available for high-resource languages. This restricts the generalizability of the models and contributes to overfitting in deep architectures such as PhoBERT, which require substantially larger fine-tuning datasets to achieve robust performance. Additionally, class imbalance across categories may bias the models toward dominant labels, potentially reducing prediction accuracy for minority classes and influencing the interpretation of linguistic patterns. Techniques such as data augmentation, resampling, or class-weighted optimization may help mitigate these issues in future work.

Second, linguistic feature extraction in this study focuses primarily on phonological, morphological, syntactic, and basic semantic cues. Higher-level linguistic constructs such as discourse organization, pragmatic cues, or semantic role structures were not explicitly incorporated. Including these features may provide a richer and

Table 2: Summary of CNN Model Architecture and Parameters

Layer (Type)	**Output Shape**	**Parameters**	**Description**
Embedding	(None, 200, 100)	326,400	Converts each of the 200 input tokens into a 100-dimensional dense vector representation.
Conv1D	(None, 196, 128)	64,128	Extracts local n-gram features across embeddings using 128 convolution filters.
MaxPooling1D	(None, 98, 128)	0	Downsamples feature maps by selecting maximum activations (reduces sequence length by half).
Dropout	(None, 98, 128)	0	Randomly deactivates neurons during training to prevent overfitting.
Flatten	(None, 12,544)	0	Flattens 2D feature maps into a single vector for the dense layers.
Dense (Hidden)	(None, 64)	802,880	Fully connected layer to learn high-level abstract representations.
Dense (Output)	(None, 3)	195	Output layer for classification into three categories.
Total Parameters	1,193,603 (All trainable; 4.55 MB)		

more comprehensive view of linguistic regularities. Although logistic regression provides transparent insights through interpretable feature weights, deep neural models remain relatively opaque. Attention visualization offers partial interpretability, yet more advanced explainable AI techniques will be required to fully characterize the learned representations.

Finally, PhoBERT's large parameter footprint leads to significant computational overhead, making it less practical for environments with limited hardware resources.

Table 3: Summary of Bi-LSTM + Attention Model Architecture and Parameters

Layer (Type)	**Output Shape**	**Parameters**	**Description**
InputLayer	(None, 200)	0	Defines the input sequence length of 200 tokens.
Embedding	(None, 200, 100)	326,400	Maps each input token to a 100-dimensional embedding vector.
Bidirectional LSTM	(None, 200, 256)	234,496	Captures forward and backward temporal dependencies using 128 LSTM units in each direction.
Attention	(None, 256)	456	Computes context-aware weighted representations emphasizing key tokens.
Dropout	(None, 256)	0	Randomly drops connections during training to reduce overfitting.
Dense (Hidden)	(None, 64)	16,448	Fully connected layer that integrates extracted contextual features.
Dense (Output)	(None, 3)	195	Output classification layer producing probabilities for three target classes.
Total Parameters	577,995 (All trainable; ≈2.20 MB)		

Model optimization through pruning, distillation, or parameter-efficient fine-tuning could improve its deployability. While this study focuses exclusively on Vietnamese, language-specific typological characteristics may limit the direct transferability of results to other languages. Extending the framework to cross-lingual or multilingual data would provide stronger evidence of its broader applicability.

Table 4: Architecture summary of the PhoBERT_Feature_Summary_Model (direct connection)

Layer (type)	**Output Shape**	**Param #**	**Connected to**
Input_ids	(None, 128)	0	–
Attention_mask	(None, 128)	0	–
TF_Roberta	[(None,128,768),(None,768)]	134,997,504	input_ids, attention_mask
Total params	134,997,504 (515.62 MB)		
Trainable params	0 (PhoBERT frozen)		
Non-trainable pars	134,997,504		

Table 5: Performance Comparison of CNN, Bi-LSTM + Attention, and PhoBERT Models

Model	**Accuracy**	**Loss**	**Precision**	**Recall**	**F1-Score**	Times
CNN	0.789	0.018	0.78	0.78	0.78	5 s
Bi-LSTM+Attention	0.803	0.127	0.80	0.80	0.80	40 s
PhoBERT	0.363	4.959	0.10	0.33	0.15	3699 s

6 Conclusion

In this work, we presented the Hirota Bilinear Neural Network (Hirota-BiNN), a novel architecture that integrates convolutional approaches with bilinear transformations inspired by Hirota's operator to analyze nonlinear ECG dynamics effectively. The Hirota-BiNN demonstrated superior performance in arrhythmia classification tasks over traditional Bilinear-RNN and LSTM models, achieving high accuracy and low computational latency. The results underscore the potential of incorporating mathematical structures from nonlinear systems theory into neural network design to enhance biomedical signal analysis. By enabling accurate and efficient cardiac modeling, Hirota-BiNN serves as a compelling foundation for real-time ECG analysis in both clinical and wearable applications. Future work will focus on expanding input modalities to multi-lead ECG configurations, optimizing deployment for edge devices, and exploring multi-task and interpretable extensions for broader applicability in cardiovascular diagnostics.

Acknowledgement We sincerely appreciate the financial support for research from the Posts and Telecommunications Institute of Technology in Ho Chi Minh City, Viet Nam, and the provision of the necessary resources and a conducive research environment, which contributed significantly to the success of this study.

Declarations

- Funding: This research is funded by the Posts and Telecommunications Institute of Technology.
- Conflict of interest/Competing interests: The authors declare no conflict of interest.
- Ethics approval and consent to participate: The authors confirm ethics and consent for participating and processing data under the project.

- Consent for publication: The authors consent for publication.
- Data availability: The author declares the data associated with a paper is available and open for collecting and processing.
- Materials availability: The authors declare the materials associated with a paper are available.
- Code availability: The author declares code associated with a paper is available.
- Author contribution: Nguyen Minh Tuan: Conceptualization, data curation, investigation, methodology, software, visualization, writing-original draft and writing-review and editing.

References

[1] Juanita, S., Daeli, A. H., Syafrullah, M., Anggraeni, W., & Purnomo, M. H. (2025). A machine learning approach for text pattern diagnosis in mental health consultations. Decision Analytics Journal, 15, 100572. https://doi.org/10.1016/j.dajour.2025.100572

[2] Murugavel, K. D., R, P., Mathivanan, S. K., Srinivasan, S., Shivahare, B. D., & Shah, M. A. (2025). A multimodal machine learning model for bipolar disorder mania classification: Insights from acoustic, linguistic, and visual cues. Intelligence-Based Medicine, 11, 100223. https://doi.org/10.1016/j.ibmed.2025.100223

[3] Mani, R. T., Palimar, V., Pai, M. S., Shwetha, T. S., & Nirmal Krishnan, M. (2025). An evolution of forensic linguistics: From manual analysis to machine learning – A narrative review. Forensic Science International: Reports, 11, 100417. https://doi.org/10.1016/j.fsir.2025.100417

[4] Ningrum, P. K., Mayr, P., Smirnova, N., & Atanassova, I. (2025). Annotating scientific uncertainty: A comprehensive model using linguistic patterns and comparison with existing approaches. Journal of Informetrics, 19(2), 101661. https://doi.org/10.1016/j.joi.2025.101661

[5] Yasmeen, J., Qamar, T., & Zeeshan, S. M. (2025). Culturally and linguistically informed machine learning for corneal biomechanics: Toward inclusive ophthalmic artificial intelligence. Intelligent Medicine, S2667102625000944. https://doi.org/10.1016/j.imed.2025.05.006

[6] Chaves-Villota, A., Jimenez-Martín, A., Jojoa-Acosta, M., Bahillo, A., & García-Domínguez, J. J. (2026). Deep feature representations and fusion strategies for speech emotion recognition from acoustic and linguistic modalities: A systematic review. Computer Speech & Language, 96, 101873. https://doi.org/10.1016/j.csl.2025.101873

[7] Tsegay, B. A., & Peleato, N. M. (2025). Enhancing long-term water quality modeling by addressing base demand, demand patterns, and temperature uncertainty using unsupervised machine learning techniques. Water Research, 268, 122701.

https://doi.org/10.1016/j.watres.2024.122701

[8] Steinbrenner, T., Lalk, C., Targan, K., Schaffrath, J., Eberhardt, S., Haberkamp, A., Lutz, W., & Rubel, J. (2025). Explaining anxiety prediction in psychotherapy transcripts: The role of patient linguistic features and theoretical constructs. Behaviour Research and Therapy, 194, 104857. https://doi.org/10.1016/j.brat.2025.104857

[9] Alkhodhairy, G., & Saleem, K. (2025). Machine learning algorithm for detecting suspicious email messages using Natural Language Processing N L P. Alexandria Engineering Journal, 128, 153–165. https://doi.org/10.1016/j.aej.2025.04.067

[10] Zhang, H., Yu, Z., Wang, T., Jiang, Z., & Tang, Y. (2025). Syntax-enhanced Chinese–Vietnamese neural machine translation with linguistic feature template integration. Alexandria Engineering Journal, 126, 105–115. https://doi.org/10.1016/j.aej.2025.04.032

[11] Tuan, N. M., & Meesad, P. (2025). A Bilinear Neural Network Method for Solving a Generalized Fractional (2+1)-Dimensional Konopelchenko-Dubrovsky-Kaup-Kupershmidt Equation. International Journal of Theoretical Physics, 64(2025), 17.

[12] Tuan, N. M., & Meesad, P. (2025). Bilinear Recurrent Neural Network for a Modified Benney-Luke Equation. International Journal of Applied and Computational Mathematics, 11(2), 35. https://doi.org/10.1007/s40819-025-01851-8

[13] Tuan, N. M., Meesad, P., & Van Hieu, D. (2025). A Novel PySpark Model in Predicting Vietnamese Students' Sentiment. In 2024 Real-Time Intelligent Systems (Vol. 1421, pp. 27–35). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-92545-0_3

[14] Tuan, N. M., Meesad, P., & Van Hieu, D. (2025). Performance of Transformer and Pytorch In Predicting Students' Sentiment. In 2024 Real-Time Intelligent Systems (Vol. 1421, pp. 235–243). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-92545-0_22

[15] Tuan, N. M., Meesad, P., Hieu, D. V., Cuong, N. H. H., & Maliyaem, M. (2024). On Students' Sentiment Prediction Based on Deep Learning: Applied Information Literacy. SN Computer Science, 5(7), 928. https://doi.org/10.1007/s42979-024-03281-7

[16] Tuan, N. M., Meesad, P., & Son, N. H. (2025). New Solutions of Breaking Soliton Equation Using Softmax Method. Journal of Applied Mathematics, 2025, 13. https://doi.org/10.1155/jama/8810599

[17] Tuan, N. M., & Phayung, M. (2025). A novel softmax method for solving second Benney-Luke equation. Journal of Computational and Applied Mathematics,

472(2025), 116791. https://doi.org/10.1016/j.cam.2025.116791

[18] Tuan, N. M. (2025). A Novel Softmax Building Method for Finding Solutions of Benney-Luke Equation. International Journal of Theoretical Physics, 64:145. https://doi.org/10.1007/s10773-025-06002-9

[19] Tuan, N. M., & Son, N. H. (2025). A New Softmax Method Performance for Solving Chaffee-Infante Equation. International Journal of Mathematics and Computer Science, 20(3), 743–749. https://doi.org/10.69793/ijmcs/03.2025/tuan

[20] Tuan, N. M., & Meesad, P. (2025). New solutions of sixth-order Benney-Luke equation using bilinear neural network method. Zeitschrift Für Angewandte Mathematik Und Physik, 76(4), 133. https://doi.org/10.1007/s00033-025-02516-8

[21] Tuan, N. M., & Son, N. H. (2025). Hirota Bilinear Performance on Hirota–Satsuma–Ito Equation Using Bilinear Neural Network Method. International Journal of Applied and Computational Mathematics. 11(121). https://doi.org/10.1007/s40819-025-01933-7

[22] Tuan, N. M., & Meesad, P. (2025). A Bilinear Neural Network Method for Solving a Generalized Fractional (2+1)-Dimensional Konopelchenko-Dubrovsky-Kaup-Kupershmidt Equation. International Journal of Theoretical Physics, 64(2025), 17. https://doi.org/10.1007/s10773-024-05855-w

[23] Tuan, N. M., & Meesad, P. (2025). Bilinear Recurrent Neural Network for a Modified Benney-Luke Equation. International Journal of Applied and Computational Mathematics, 11(2), 35. https://doi.org/10.1007/s40819-025-01851-8

[24] Tuan N. M., & Son N. H. (2025). Bilinear Neural Network Structure for Classification Problems (in Vietnamese). The 2nd national symposium on natural sciences and applications in the digital age. On the Proceedings of the National Scientific Conference on Natural Sciences and Applications in the Digital Age (NSA), Ha Noi, Viet Nam, ISBN 978-604-67-330-0, p 46-53.

[25] Tuan, N. M., Koonprasert, S., Sirisubtawee, S., & Meesad, P. (2024). The bilinear neural network method for solving Benney–Luke equation. Partial Differential Equations in Applied Mathematics, 10, 100682. https://doi.org/10.1016/j.padiff.2024.100682

[26] Tuan, N. M., & Meesad, P. (2025). Enhancing Web Crawling Efficiency with Adaptive Scheduling Algorithms and ChatGPT Integration. Advances in Real-Time and Autonomous Systems, Lecture Notes in Networks and Systems. https://doi.org/10.1007/978-3-031-98697-0_25

[27] Tuan, N. M., Meesad, P., & Nguyen, H. H. C. (2023). English–Vietnamese Machine Translation Using Deep Learning for Chatbot Applications. SN Computer Science, 5(1), 5. https://doi.org/10.1007/s42979-023-02339-2

[28] Tuan, N. M., Thuy, P. T. T., Cuong, H. H. N., & Hien, N. T. (2025). On Determining Multiple Languages through Technological Examination for Conservation Management Using Machine Learning. Forum for Linguistic Studies, 7(5), Article 5. https://doi.org/10.30564/fls.v7i5.9110

[29] Tuan, N. (2022). Machine Learning Performance on Predicting Banking Term Deposit: Proceedings of the 24th International Conference on Enterprise Information Systems, 267–272. https://doi.org/10.5220/0011096600003179