ADVANCING EMOTION RECOGNITION IN VIETNAMESE: A PHOBERT-BASED APPROACH FOR ENHANCED INTERACTION

Huynh Thi Ngoc Tram[1,3], Pham Minh Dzuy[2,3*], Pham Duc Dat[1,3], Le Duy Tan[1,3], Huynh Kha Tu[1,3]

[1]*International University - Vietnam National University Ho Chi Minh City*
[2]*University of Information Technology - Vietnam National University Ho Chi Minh City*
[3]*Vietnam National University Ho Chi Minh City*

ARTICLE INFO

Received: 26/5/2025
Revised: 29/6/2025
Published: 29/6/2025

KEYWORDS

Emotion recognition
Vietnamese natural language processing
Deep learning models
Sentiment analysis
Artificial intelligence

ABSTRACT

Emotion recognition using artificial intelligence is essential for improving human-machine interactions in healthcare, education, and smart homes. Addressing Vietnamese - specific challenges such as tonality and context-dependent meanings, we developed a high-quality dataset from social media, product reviews, and conversational dialogs. Rigorous preprocessing (cleaning, normalization, tokenization) and oversampling addressed class imbalance, enhancing data reliability. PhoBERT-base-v2, a Vietnamese-optimized Transformer, achieved state-of-the-art accuracy (94.22%) and macro metrics (> 94%), significantly outperforming traditional machine-learning and other deep-learning methods. Analysis revealed strong differentiation of nuanced emotions, though confusion persisted between semantically similar feelings (e.g., Anger vs. Disgust). We demonstrated practical deployment via a Gradio interface for real-time sentiment analysis, illustrating potential applications like social media monitoring, customer feedback analysis, and mental health support. Future work includes multimodal approaches combining text and speech for enhanced accuracy.

NGHIÊN CỨU NÂNG CAO NHẬN DẠNG CẢM XÚC TIẾNG VIỆT: PHƯƠNG PHÁP DỰA TRÊN PHOBERT PHỤC VỤ TƯƠNG TÁC HIỆU QUẢ

Huỳnh Thị Ngọc Trâm[1,3], Phạm Minh Dzuy[2,3*], Phạm Đức Đạt[1,3], Lê Duy Tân[1,3], Huỳnh Khả Tú[1,3]

[1]*Trường Đại học Quốc tế - Đại học Quốc gia Thành phố Hồ Chí Minh*
[2]*Trường Đại học Công nghệ Thông tin - Đại học Quốc gia Thành phố Hồ Chí Minh*
[3]*Đại học Quốc gia Thành phố Hồ Chí Minh*

THÔNG TIN BÀI BÁO

Ngày nhận bài: 26/5/2025
Ngày hoàn thiện: 29/6/2025
Ngày đăng: 29/6/2025

TỪ KHÓA

Nhận dạng cảm xúc
Xử lý ngôn ngữ tự nhiên tiếng Việt
Mô hình học sâu
Phân tích cảm xúc
Trí tuệ nhân tạo

TÓM TẮT

Nhận diện cảm xúc bằng trí tuệ nhân tạo đóng vai trò quan trọng trong việc cải thiện tương tác người – máy. Nghiên cứu này giải quyết các thách thức đặc thù của tiếng Việt bằng cách xây dựng bộ dữ liệu chất lượng cao từ bài đăng trên mạng xã hội, đánh giá sản phẩm và hội thoại tự nhiên. Dữ liệu được tiền xử lý nghiêm ngặt (làm sạch, chuẩn hóa, tách từ) và sử dụng kỹ thuật lấy mẫu bổ sung để cân bằng các lớp dữ liệu thiếu hụt, nâng cao độ tin cậy của mô hình. PhoBERT-base-v2, một mô hình Transformer tối ưu cho tiếng Việt, đạt độ chính xác hiện đại (94,22%) và các chỉ số macro trên 94%, vượt trội rõ rệt so với các phương pháp học máy truyền thống và các mô hình học sâu khác. Phân tích cho thấy mô hình phân biệt tốt các cảm xúc tinh tế, dù vẫn còn nhầm lẫn giữa các cảm xúc gần nhau (ví dụ: Giận dữ và Ghê tởm). Chúng tôi triển khai giao diện Gradio để minh họa ứng dụng thực tế trong phân tích thời gian thực, giám sát mạng xã hội, đánh giá phản hồi khách hàng và hỗ trợ sức khỏe tinh thần. Hướng nghiên cứu tương lai đề xuất tích hợp đa phương thức (văn bản và giọng nói) nhằm nâng cao hơn nữa độ chính xác nhận diện.

DOI: https://doi.org/10.34238/tnu-jst.12889

* Corresponding author. *Email: dzuypm.18@grad.uit.edu.vn*

1. Introduction

Emotion recognition systems utilizing artificial intelligence (AI) are progressively emerging as a pivotal research domain in computer science and intelligent technology applications, with swift advancements enabling these systems to transcend simple emotional state identification by capturing nuanced subtleties and enhancing human-machine interactions across various contexts. These systems have been extensively utilized in sectors such as healthcare, education, and human-machine interactions, thus enhancing quality of life and user experience [1].

Nevertheless, emotion recognition in the Vietnamese language presents several unique challenges. As a monosyllabic and tonal language, Vietnamese heavily depends on both tone variations and contextual information, thereby complicating emotional interpretation compared to many other languages [2]. Furthermore, the prevalent use of implicit or sarcastic expressions within Vietnamese discourse further exacerbates the difficulty of accurately analysing and identifying the intended emotional nuances [3].

This research aims to create a high-quality Vietnamese emotion dataset sourced from platforms like social media, product reviews, and everyday conversations, providing a solid foundation for training emotion recognition systems tailored to the linguistic challenges of Vietnamese. The study also emphasizes data preprocessing to optimize alignment with emotion recognition models. A comparative analysis of emotion recognition methods, including traditional machine learning and advanced deep learning models like PhoBERT-base-v2 - a BERT variant optimized for Vietnamese [4]. The developed emotion recognition system has broad applications, enhancing user interaction in virtual assistants and smart home devices, and supporting early detection of emotional issues in healthcare.

Emotion recognition is an important field of study in natural language processing (NLP). Text-based emotion detection systems typically examine semantics and context to determine emotional states from text content. Traditional approaches such as Naive Bayes and Support Vector Machines (SVM) use hand-crafted features, including Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF), to classify emotions [5].

Deep learning has transformed the field of sentiment classification, allowing models to learn and understand semantic and contextual correlations in data. Recurrent Neural Networks (RNNs) have played an essential role in processing sequential data but are hampered by the vanishing gradient problem [6], [7]. To counter this, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models were created, which help store long-term information and enhance accuracy in sentiment classification [8].

The introduction of Transformer models, especially BERT, marked a significant advancement in NLP by enabling parallel analysis of entire phrases and capturing global dependencies, improving performance on complex sentences [9]. PhoBERT-base-v2, a variation of BERT designed for Vietnamese, has exhibited higher performance in sentiment classification and semantic analysis tasks, thanks to its capacity to handle compound words and tonal shifts peculiar to Vietnamese [4]. PhoBERT-base, built on the Transformer architecture and optimized for the Vietnamese language, has demonstrated exceptional performance in sentiment classification tasks. PhoBERT-base-v2 uses Byte Pair Encoding (BPE) to effectively handle unknown words or those with complex structures, enabling accurate sentiment classification in Vietnamese texts. Recent study indicates that PhoBERT-base-v2 surpasses conventional models like Naive Bayes and SVM, in the analysis of Vietnamese data. Experimental findings indicate that PhoBERT-base-v2 attains superior accuracy in tasks including product review analysis and social media sentiment recognition [10].

However, emotion recognition in Vietnamese faces challenges due to the lack of high-quality, diverse datasets. While languages like English benefit from extensive labelled emotion data, Vietnamese datasets are limited, mainly focusing on social media and product reviews, hindering models' ability to fully understand and classify emotions [3], [11]. This limitation reduces the ability of models to understand and classify emotions accurately and comprehensively.

The recognition of emotions in Vietnamese is a promising domain yet continues to encounter numerous hurdles. The development of multimodal systems and the utilization of deep learning models like PhoBERT-base-v2 are expected to enhance the precision and practical application of these systems. Investment in data development and research is essential to address the language and cultural particularities of Vietnamese.

2. Methodology

Before delving into the methodology, Figure 1 presents an overview of the end-to-end process for building and deploying a Vietnamese emotion recognition system, spanning data collection, preprocessing, label balancing, model training, and real-world deployment.

2.1. Dataset acquisition

This study emphasizes dataset creation to enhance accuracy and efficiency of Vietnamese sentiment classification models, addressing the original UIT-VSMEC [3] dataset's limitations of small size and severe imbalance by integrating diverse sources like social media and everyday conversations.

Initially, comment data from social networks is gathered from prominent sites like Twitter and Facebook. These provide abundant sources of emotional expressiveness, encompassing intricate emotions such as sarcasm, irony, or distinctly positive or negative reactions. The data from these platforms accurately reflect the natural and informal linguistic nuances of users, assisting the sentiment classification model in identifying significant aspects in real-world scenarios [12].

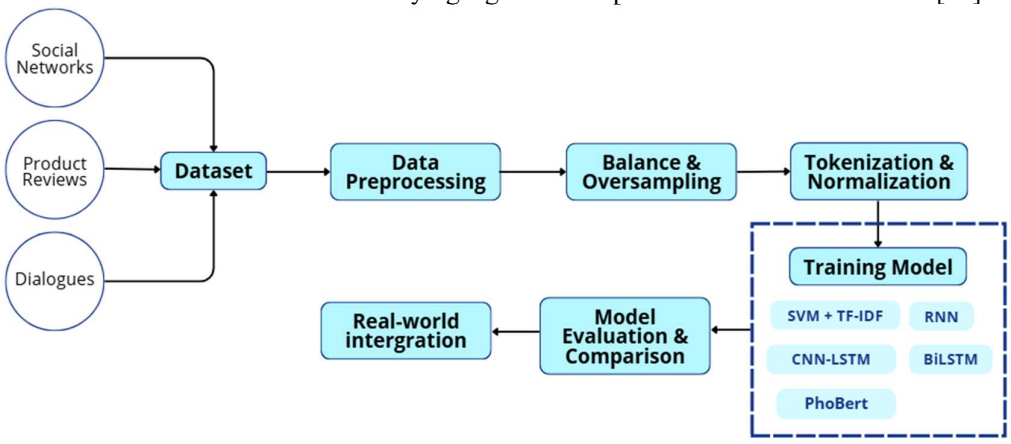

Figure 1. *Overview of the process of Vietnamese emotion recognition system*

Secondly, the data from quotidian conversational texts are generated by modelling communicative scenarios in domestic or virtual assistant contexts. This not only guarantees the dataset's appropriateness for practical applications but also enhances the performance of the sentiment classification model in human-machine interaction contexts.

The dataset is structured as a CSV file with two columns: "sentence" for textual data and "emotion" for corresponding sentiment labels, both stored as strings. Dataset statistics and source distribution details are summarized in Table 1.

However, a major challenge in building the dataset is the imbalance among emotion labels, with "enjoyment" significantly dominating and labels such as "anger" or "sadness" underrepresented in the original UIT-VSMEC dataset [3]. The study uses RandomOverSampler to balance each class at 4,956 instances, mitigating bias and enhancing generalization, resulting in 34,692 training samples per fold.

The data undergoes a comprehensive cleaning process to ensure consistency and quality.

Emoticons are replaced with emotion descriptors like "happy," while non-standard abbreviations and profanities are removed or substituted with appropriate terms. Extraneous spaces and special characters are also eliminated to reduce noise. In normalization, numerical values are replaced with the keyword [number] for consistency. The UnderThesea [13] tool is used for tokenizing the sentences, which is particularly important for Vietnamese due to its complex word structure and contextual dependencies.

Table 1. *Dataset statistics and source distribution*

Label	Number of Rows	Source information (%)
Enjoyment	4956	Social networks: 70%
Anger	1845	product reviews: 5%
Sadness	1778	dialogues: 25%
Disgust	1658	
Fear	1636	
Surprise	1630	
Other	2015	

The approach not only cleanses the dataset but also ensures balance among sentiment labels, creating a strong foundation for training and evaluating sentiment classification models. This data processing method allows the model to understand both clear emotional expressions and complex semantic nuances, improving its effectiveness in real-world applications such as virtual assistants and user feedback analysis.

2.2. Models and architectures

This study employs a range of traditional machine learning and contemporary deep learning models to assess their efficacy in classifying sentiments from Vietnamese text. Each model offers distinct viewpoints and capabilities, ranging from conventional feature-based techniques to contemporary deep learning methods that can capture linguistic context. The utilized models comprise:

Support Vector Machine (SVM) is a traditional machine learning method that uses optimized hyperplanes to classify data. In this study, SVM with TF-IDF vectorization transforms textual data into numerical vectors for classification. Effective in simple feature scenarios, SVM serves as a baseline for comparing advanced models [14]. SVMs are particularly effective in scenarios where the data can be linearly separated, as they aim to find the optimal hyperplane that maximizes the margin between classes.

To apply **SVM** to text classification tasks, the text data needs to be converted into numerical feature vectors. One common technique for this conversion is **TF-IDF** (Term Frequency-Inverse Document Frequency). TF-IDF measures the importance of a word within a document, calculated as the product of two components: term frequency (TF) and inverse document frequency (IDF). The TF-IDF score for a term t in document d is given by:

$$TF - IDF(t,d) = TF(t,d) \times log\ log\ \left(\frac{N}{DF(t)}\right) \quad (1)$$

Where $TF(t,d)$ is the frequency of term t in document d, and $DF(t)$ is the number of documents containing t [15]. This TF-IDF representation is then used as input to SVM for emotion classification, where it helps the model identify and classify emotional nuances in text based on the frequency and significance of words [16].

Recurrent Neural Networks (RNNs) are designed to process sequential data like text and audio. Unlike feedforward networks, RNNs retain information from previous steps, enabling them to capture context and temporal dependencies, making them ideal for tasks like emotion classification and speech recognition [17].

Despite their effectiveness in capturing short-term dependencies, RNNs face the vanishing gradient problem, which hampers learning long-term dependencies, limiting their performance in

tasks requiring long-term context [6]. As a result, RNNs may fail to capture important contextual information when processing long sequences.

The **CNN-LSTM** model combines **Convolutional Neural Networks (CNN)** and **Long Short-Term Memory (LSTM)** networks to analyse complex texts. CNN extracts local semantic features, such as n-grams or grammatical structures, but is limited in capturing long-term dependencies. LSTM addresses this by storing and analysing long-term information, enabling the model to capture relationships between words, sentences, or paragraphs. The architecture includes an embedding layer for word vectors, a CNN layer for local feature extraction, and an LSTM layer for learning long-term dependencies. The final fully connected layer performs classification, with LSTM's gates controlling the retention or discarding of information during learning.

In summary, the CNN-LSTM model not only achieves high performance in text classification tasks but also has broader applications in natural language processing, such as sentiment analysis and named entity recognition.

Bidirectional Long Short-Term Memory (BiLSTM) is an enhancement of the standard LSTM architecture that processes sequential data in both forward and backward directions. This bidirectional approach enables BiLSTM to effectively capture contextual information from both past (preceding words) and future (following words) contexts, which is particularly beneficial in sentiment classification tasks, where the meaning of a word often depends heavily on its surrounding context—such as negation words altering the sentiment expressed earlier in a sentence.

Each directional hidden state is computed through input, forget, and output gates, using sigmoid and hyperbolic tangent activation functions inherent in the LSTM structure. Due to this capability, BiLSTM has emerged as a prominent choice in sentiment analysis applications, especially effective for capturing subtle linguistic nuances and complex semantic structures [18].

In this study, sentiment classification was conducted using **PhoBERT-base-v2**, a Transformer-based Vietnamese language model [4]. PhoBERT-base-v2 utilizes the Transformer encoder architecture [9], such as dynamic masking and enhanced hyperparameter tuning. Specifically tailored for Vietnamese, PhoBERT-base-v2 employs a customized tokenizer combining Vietnamese word segmentation via the RDRSegmenter and Byte-Pair Encoding (BPE) for optimal sub word representation [4]. Importantly, PhoBERT-base-v2 was selected for this research because it has been trained on an expanded and more diverse corpus—approximately 140 GB, including additional data from OSCAR-2301—which significantly improves its capacity to capture linguistic nuances and contextual meanings specific to Vietnamese. Furthermore, previous studies have demonstrated PhoBERT-base-v2's superior performance over its predecessor, particularly in sentiment analysis tasks [10].

2.3. Configuration and hyperparameters

The dataset is split into 70% for training, 15% for validation, and 15% for testing, with a stratified split to maintain label distribution. We use 5-fold cross-validation for training and hyperparameter optimization, applying oversampling only on the training set to avoid data leakage. All procedures are set with a seed of 42 for reproducibility.

PhoBERT-base-v2 is trained with 10 epochs, a learning rate of 2e-5, and a batch size of 16 using the vinai/phobert-base-v2 checkpoint. RNN and BiLSTM use a learning rate of 0.001, a batch size of 32, and 20-100 epochs with Dropout. CNN-LSTM incorporates kernels of dimensions 3 and 5 with Dropout. SVM uses a linear kernel and TF-IDF for text vectorization.

These models, chosen for their balance between traditional and modern techniques, provide insights into the best approach for Vietnamese emotion classification, with PhoBERT-base-v2 representing the state-of-the-art, while SVM serves as a benchmark. RNN, BiLSTM, and CNN-LSTM highlight the effectiveness of sequential and local feature extraction methods.

3. Result and Discussion

The experimental results, which are shown in Table 2, indicate a distinct disparity in performance across the models when assessed on the test dataset, with PhoBERT-base-v2 surpassing the other models. PhoBERT-base-v2 attained the maximum accuracy of 94.22%, with the Precision (Macro), Recall (Macro), and F1-score (Macro) metrics all over 94%. This validates PhoBERT's capacity to comprehend intricate context and differentiate nuanced emotional states in Vietnamese text.

Table 2. *Performance comparison of emotion recognition models on the custom dataset*

Models	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)
PhoBERT-base-v2	**94.22%**	**94.20%**	**94.22%**	**94.16%**
SVM	**78.69%**	78.65%	78.68%	78.65%
CNN-LSTM	62.47%	59.19%	58.09%	58.40%
BiLSTM	59.56%	56.41%	54.77%	55.39%
RNN	**30.02%**	4.29%	14.29%	**6.60%**

The quality improvements in dataset construction and class balancing notably supported PhoBERT's high performance, further underscoring the importance of comprehensive data preprocessing. Enhanced data cleaning, normalization, and tokenization procedures ensured data integrity, allowing PhoBERT-base-v2 to capture more nuanced contextual features effectively.

In comparison to PhoBERT-base-v2, SVM demonstrated superior performance relative to conventional deep learning models, attaining an accuracy of 78.69%. Despite its inability to manage context as neural network-based models do, SVM proved useful by employing TF-IDF for feature extraction. Nevertheless, models like CNN-LSTM (62.47%) and BiLSTM (59.56%) exhibit constraints in attaining comparably high performance, while they still surpass RNN (30.02%), the least effective model in these evaluations. Figure 2 shows PhoBERT's confusion matrix, highlighting classification accuracy across emotional categories.

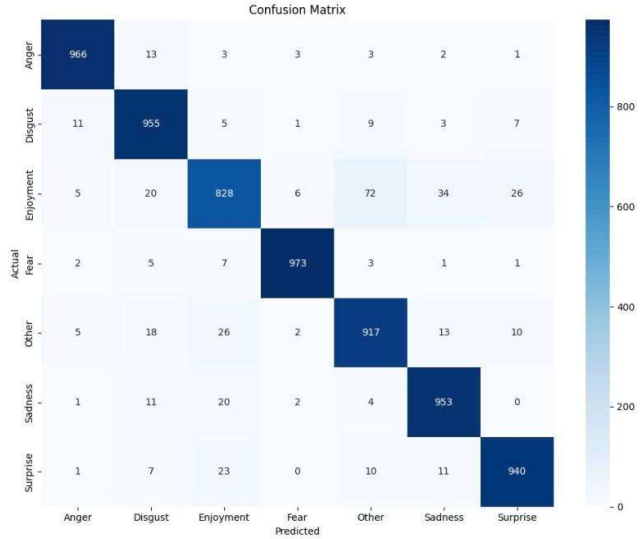

Figure 2. *Confusion Matrix of the PhoBERT-base-v2 Model*

In addition to the evaluation on the custom dataset, a comparison of model performance on the original UITVSMEC dataset is presented, revealing a noticeable performance disparity. Table 3 illustrates the results of PhoBERT-base-v2, SVM, CNN-LSTM, BiLSTM, and RNN models trained and evaluated on UITVSMEC. PhoBERT-base-v2 outperforms all other models on both datasets, achieving an accuracy of 94.22% on the Custom Dataset and 63.31% on UITVSMEC. However, the model demonstrates significantly better performance on the Custom Dataset,

highlighting the substantial impact of improved data quality and preprocessing techniques. Similarly, SVM, CNN-LSTM, and BiLSTM show enhanced performance on the Custom Dataset relative to UITVSMEC, emphasizing the value of the newly constructed dataset in boosting model efficacy. In contrast, RNN consistently exhibits the lowest accuracy across both datasets, reflecting its limitations, particularly when applied to the UITVSMEC dataset.

Table 3. *Performance comparison of emotion recognition models on UIT-VSMEC*

Models	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)
PhoBERT-base-v2	**63.31%**	**63.31%**	**63.08%**	**63.53%**
SVM	**58.59%**	58.62%	52.76%	54.65%
CNN-LSTM	52.81%	59.55%	41.11%	41.33%
BiLSTM	50.94%	53.84%	45.43%	47.65%
RNN	**27.85%**	3.98%	14.29%	**6.22%**

PhoBERT-base-v2 has demonstrated outstanding performance in Vietnamese emotion classification, thanks to its Transformer-based architecture and pretraining on a large Vietnamese corpus. This enables it to capture complex contextual nuances, such as sarcasm or implicit meanings, and understand tonal or grammatical subtleties—surpassing traditional models like SVM and neural networks (CNN-LSTM, BiLSTM). These results emphasize the importance of language-specific optimization, especially for morphologically rich and under-resourced languages like Vietnamese.

PhoBERT's practical value is evident in applications such as empathetic virtual assistants, emotion-aware smart devices, and responsive customer service systems. However, effective deployment depends heavily on dataset quality and diversity. While this study improved label balance, informal and regional variations are still underrepresented. Furthermore, PhoBERT's high computational demands present obstacles for real-time, multimodal systems combining text, speech, and other inputs. Additionally, transformer-based models like PhoBERT typically require extensive data and computational resources, potentially limiting their effectiveness in resource-constrained environments.

In mental healthcare, emotion-recognition technology goes beyond the early detection of anxiety, depression and mood swings; it also enables the deployment of therapeutic chatbots. When trained to understand the nuances, idioms and regional variants of Vietnamese, such chatbots can analyse both text and speech in real time, recognize the patient's emotional state, and respond empathetically. By accurately interpreting culturally specific expressions and hidden psychological cues, the system can automatically log affective trends, flag high-risk situations and provide clinicians with timely alerts, thereby improving treatment decisions [19]. Around-the-clock availability further reduces stigma and gives users a safe space to share feelings in their native language, turning the chatbot into a trusted digital assistant that standardizes emotional monitoring and optimizes care pathways for Vietnamese patients.

4. Conclusion and Future Work

This study advances Vietnamese emotion recognition by building a comprehensive, balanced dataset from varied sources and addressing key challenges like class imbalance and linguistic complexity. PhoBERT-base-v2 outperformed traditional and deep learning models in capturing tonal and contextual nuances, proving effective for real-world applications such as empathetic virtual assistants, smart homes, and customer service.

Future work should focus on multimodal systems combining text and speech for better emotional understanding. This approach can significantly improve emotion recognition accuracy, especially for languages like Vietnamese, where tone, context, and subtleties in speech are crucial for conveying emotional meaning. While text-based models like PhoBERT are effective at capturing semantic content, they may struggle with spoken language nuances such as intonation, pitch, and prosody. By integrating speech features, multimodal systems can leverage both the

linguistic information from text and emotional cues from speech, providing a more comprehensive and accurate emotional analysis [20].

Emotion detection also holds promise in healthcare and education - supporting early mental health monitoring and enabling teachers to adapt to students' emotional states. Despite encouraging results, further work is needed to expand dataset diversity, especially with regional and informal language, and to optimize PhoBERT for real-time use by reducing latency and computational load.

Acknowledgments

The authors would like to express their gratitude to AIoT Lab VN for the support throughout this project.

REFERENCES

[1] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah, and M. Hamdi, "Emotion recognition for healthcare surveillance systems using neural networks: A survey," in *Proceedings of the 2021 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Harbin City, China, 2021, pp. 681–687, doi: 10.1109/IWCMC51323.2021.9498861.

[2] X. T. Le, T. T. Dao, V. L. Trinh, and H. Q. Nguyen, "Speech Emotions and Statistical Analysis for Vietnamese Emotion Corpus," *Journal on Information Technologies & Communications,* vol. V-1, no. 35, pp. 86-98, 2022, doi: 10.32913/mic-ict-research-vn.v1.n35.233.

[3] V. A. Ho, D. H.-C. Nguyen, D. H. Nguyen, L. T.-V. Pham, D.-V. Nguyen, K. V. Nguyen, and N. L.-T. Nguyen, "Emotion Recognition for Vietnamese Social Media Text," *CoRR,* 2019, doi: 10.48550/arXiv.1911.09339.

[4] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 1037–1042, doi: 10.18653/v1/2020.findings-emnlp.92.

[5] A. F. A. Nasir, E. Nee, C. S. Choong, A. S. A. Ghani, A. P. P. A. Majeed, A. Adam, and M. Furqan, "Text-based emotion prediction system using machine learning approach," in *IOP Conference Series: Materials Science and Engineering*, vol. 769, Jun. 2020, Art. no. 012022.

[6] R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," in *Proceedings of the 30th International Conference on Machine Learning (ICML)*, vol. 28, no. 3, pp. 1310–1318, 2013.

[7] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," in *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157-166, March 1994, doi: 10.1109/72.279181.

[8] S.-H. Noh, "Analysis of Gradient Vanishing of RNNs and Performance Comparison," *Information,* vol. 12, vol. 12, no. 11, 2021, Art. no. 442, doi: 10.3390/info12110442.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[10] M. T. Ngo, B. H. Ngo, and V. V. Stuchilin, "Fine-tuned PhoBERT for sentiment analysis of Vietnamese phone reviews," *CTU Journal of Innovation & Sustainable Development*, vol. 16, no. Special issue: ISDS, pp. 52-57, 2024.

[11] H. T. T. Thieu, "Challenges in Classification of Vietnamese Sentiment," *International Journal of Scientific and Technical Research in Engineering (IJSTRE),* vol. 6, no. 5, pp. 1–6, 2021.

[12] N. D. Q. Anh, M.-H. Ha, Q. C. Nguyen, T. H. T. Nguyen, Q. Vu, D. X. Minh-Duc, D.-C. Nguyen, and T. K. Dinh, "VNEMOS: Vietnamese Speech Emotion Inference Using Deep Neural Networks," in *2024 9th International Conference on Integrated Circuits, Design, and Verification (ICDV)*, Hanoi, Vietnam, 2024, pp. 97-101, doi: 10.1109/ICDV61346.2024.10616411.

[13] undertheseanlp, "undertheseanlp/underthesea: Underthesea - Vietnamese NLP Toolkit," 2017, [Online]. Available: https://github.com/undertheseanlp/underthesea. [Accessed 11 May 2025].

[14] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[15] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of Documentation,* vol. 60, no. 5, pp. 503-520, 2004.
[16] N. S. M. Nafis and S. Awang, "An Enhanced Hybrid Feature Selection Technique Using Term Frequency-Inverse Document Frequency and Support Vector Machine-Recursive Feature Elimination for Sentiment Classification," *IEEE Access,* vol. 9, pp. 52177-52192, 2021.
[17] E. Gkintoni, A. Aroutzidis, H. Antonopoulou, and C. Halkiopoulos, "From Neural Networks to Emotional Networks: A Systematic Review of EEG-Based Emotion Recognition in Cognitive Neuroscience and Real-World Applications," *Brain Sciences,* vol. 15, no. 3, 2025, Art. no. 220.
[18] Z. Hameed and B. Garcia-Zapirain, "Sentiment Classification Using a Single-Layered BiLSTM Model," *IEEE Access,* vol. 8, pp. 73992-74001, 2020.
[19] M. Samaneh, P. David, A. Olayinka, P. Christian, M. Farhaan, M. Shilpa, and S. Sandra, "Automatic Speech Emotion Recognition Using Machine Learning: Digital Transformation of Mental Health," in *PACIS 2022 Proceedings*, Chiang Mai, Thailand, 2022, Art. no. 45.
[20] M. Awatef, B. Hayet, and L. Zied, "Multimodal emotion recognition: Integrating speech and text for improved valence, arousal, and dominance prediction," *Annals of Telecommunications.,* vol. 80, no. 5, pp. 401-415, 2025.