***Original Research Article***

**A Hybrid PhoBERT-CNN-LSTM Model for Sentiment Analysis of Vietnamese Student Feedback**

# ABSTRACT

Student feedback plays a crucial role in improving educational quality and creating an effective learning environment. This study applies sentiment analysis to Vietnamese student feedback to extract and classify emotions into positive, negative, and neutral categories, providing valuable insights to support teaching improvements. We propose a method that utilizes the PhoBERT model for semantic feature extraction, followed by a CNN-LSTM architecture to capture both local features and sequential relationships in feedback data. Experimental results on the UIT-VSFC dataset demonstrate that the proposed PhoBERT-based CNN-LSTM model achieves an accuracy of 93.24% and an F1-score of 92.92%. This model demonstrates superior performance compared to several other advanced approaches. It surpasses ensemble model, which achieved an F1-score of 92.79%.These findings confirm the effectiveness of the model in extracting and classifying sentiments from student feedback while proposing a practical approach for analyzing Vietnamese educational data, contributing to teaching quality enhancement.

*Keywords: Sentiment analysis, Machine learning, Deep learning, PhoBERT, Vietnamese natural language processing.*

# 1. INTRODUCTION

In student evaluation of teaching (SET), which plays a growing role in learner-centered education, gathering and analyzing student feedback has become common both on online learning platforms and at higher education institutions. The primary objective is to enhance teaching quality and refine pedagogical approaches (Giang et al, 2020). In the past, closed-ended Likert-scale questions were popular because they were easy to collect and analyze. Open-ended questions now stand out for their ability to better reflect emotions and provide more detailed information (Stupans et al, 2016). At the same time, analyzing feedback from Vietnamese students is challenging, especially when using the UIT-VSFC dataset. These challenges include slang, spelling errors, emoticons, and data imbalance, with neutral labels accounting for only approximately 4% of the dataset (Nguyen et al, 2018). Similar challenges have been documented in other studies (Kastrati et al, 2021) (Chebolu et al, 2022), including small sample sizes, the lack of publicly available datasets, and data imbalance when performing sentiment classification on user comments or feedback. Overcoming these challenges calls for models with a strong grasp of language and context. Pre-trained language models stand out in this regard, as their large-scale training enables them to effectively understand text, context, and vocabulary. Transformer-based architectures such as BERT have achieved state-of-the-art performance in various sentiment classification tasks across different domains (Devlin et al, 2018). Research by Sharma and Si highlights BERT as the leading model, surpassing LSTM and CNN. Its advantage comes from its ability to understand context, learn from large datasets, handle unfamiliar words, and accurately recognize subtle emotional expressions (Kaur and Sharma, 2023) (Si, 2025).

This research proposes a novel model for sentiment classification of feedback from Vietnamese students by integrating a custom Transformer architecture. The model classifies sentiments into three categories: positive, negative, and neutral while incorporating a semantic attention mechanism. A key contribution of this study is the evaluation and comparison of the pre-trained PhoBERT model with previous approaches, highlighting its effectiveness in enhancing sentiment classification performance.

The rest of the paper is structured as follows. Section 2 details the datasets used in the research and describes the research methodology applied. Section 3 presents the experimental results and detailed analysis. Finally, we draw a conclusion in Section 4.

# 2. METHODOLOGY

The model development process follows a structured framework as illustrated in Figure 1. The first step involves preprocessing student feedback data, which includes handling missing values, removing punctuation, stop words, and numbers, as well as performing normalization and tokenization to prepare the data for training. After preprocessing, the sentiment analysis model is developed through sequential stages: model training, evaluation, and hyperparameter tuning to optimize performance. The model is then assessed using multiple metrics to ensure accuracy and reliability. Upon deployment, it can process user queries and classify sentiments (positive, negative, neutral), enabling educational institutions to extract valuable insights for enhancing teaching quality.

*A diagram of a data processing process

AI-generated content may be incorrect.*

**Figure 1: Sentiment Classification Processs**

## 2.1 Dataset

For this study, we utilized the UIT-VSFC dataset (Nguyen et al., 2018), which was published in 2018 and consists of student responses collected from end-of-term surveys at a Vietnamese university between 2013 and 2016. The dataset

contains over 16,000 Vietnamese sentences. Table 1 provides sample sentences from the UIT-VSFC dataset, each labeled with different sentiment categories: positive, neutral, and negative.

**Table 1.** Examples of Samples from the UIT-VSFC Dataset

|  |  |  |
| --- | --- | --- |
| **No.** | **Sentence** | **Sentiment** |
| 1 | Thầy rất tận tình và đi dạy rất đúng giờ.  (The teacher is very dedicated and always punctual.) | Positive |
| 2 | Tạo ra sự cạnh tranh trong mỗi buổi thực hành.  (Creates competition in each practice session.) | Neutral |
| 3 | Không có gì đặc biệt.  (Nothing special.) | Neutral |
| 4 | Thời lượng môn học quá dài, sinh viên cảm thấy mệt mỏi.  ( The course duration is too long, students feel tired.) | Negative |

The dataset exhibits a significant class imbalance, with over 90% of the data belonging to the positive and negative sentiment labels, while the neutral label comprises only approximately 4% of the total data. Table 2 illustrates this disproportionate distribution of the dataset, including 7,440 positive sentences (46%), 8,037 negative sentences (50%), and 698 neutral sentences (4%).

**Table 2.** Sentiment statistics in UIT-VSFC

|  |  |
| --- | --- |
| **Sentiment Type** | **Number of Sentences** |
| Positive | 7,440 |
| Negative | 8,037 |
| Neutral | 698 |
| Total | 16,175 |

Table 2 reveals a significant imbalance in the UIT-VSFC dataset, with only about 4% of the sentences belonging to the neutral class, while the positive and negative classes account for approximately 46% and 50%, respectively. This distribution partly reflects the reality that students tend to express their opinions only when they have particularly positive or negative experiences, whereas neutral sentiments are less commonly conveyed. However, this imbalance can impact the performance of sentiment analysis models, making them prone to bias and reducing their ability to accurately recognize neutral feedback. Therefore, understanding and being aware of the data distribution is essential to ensure that the analysis results are objective and comprehensive.

**Table 3.** Sentiment distribution statistics of UIT-VSFC Dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset** | **Total reviews** | **Positive** | **Neutral** | **Negative** |
| **Train** | 11,426 | 5,643 | 458 | 5,325 |
| **Validation** | 1,583 | 705 | 73 | 805 |
| **Test** | 3,166 | 1,589 | 167 | 1,410 |

To ensure rigorous training and objective evaluation, the UIT-VSFC dataset is proportionally split into Train, Validation, and Test sets, as shown in Table 3. The class imbalance, especially the small proportion of the neutral class, remains consistent across all subsets. This setup helps the model learn from a representative distribution while allowing fair evaluation. The sentiment classification task follows a structured pipeline with two stages: data cleaning (punctuation removal, Unicode normalization, abbreviation handling) and preprocessing (word segmentation, lowercase conversion, stopword removal). These steps enhance consistency and data quality, enabling the model to learn meaningful patterns effectively.

A white and yellow rectangular box with text

AI-generated content may be incorrect.

**Figure 2: Data Preprocessing Steps in Sentiment Classification**

## 2.2 Feature extraction

In our study utilizing machine learning techniques, we experimented with various text classification models, incorporating features such as TF-IDF, bi-grams, and Word2Vec. The TF-IDF (Term Frequency-Inverse Document Frequency) method quantifies the importance of individual words within a given response by balancing their frequency in a specific document against their rarity across the entire dataset (Sparck Jones, 1972). This approach effectively highlights key terms with strong discriminative power. N-grams, a probabilistic model, predict the likelihood of a word appearing based on the preceding N-1 words. By analyzing the frequency of consecutive word sequences within the corpus, N-grams capture linguistic patterns and sequential dependencies. In contrast, Word2Vec, a deep learning-based technique, generates word embeddings by mapping words into a multi-dimensional vector space. This method places words with similar contextual and semantic properties in close proximity within the vector space, allowing the model to identify semantic relationships and significantly improve performance in natural language processing tasks (Mikolov et al, 2013).

## 2.3 Deep learning method based on Transformer architecture

Our proposed model integrates Word2Vec for word representation, LSTM for capturing long-term contextual dependencies, CNN for extracting local features, and Multi-Head Attention to enhance focus capabilities, thereby optimizing performance in natural language processing tasks.).

**A diagram of a diagram

AI-generated content may be incorrect.**

**Figure 3:** **The CNN-LSTM architecture combined with Attention.**

The Multi-Head Attention mechanism, a pivotal component of our model, also serves as the foundation of the Transformer architecture, as introduced in the seminal paper "Attention Is All You Need" (Vaswani et al, 2017). As depicted in Figure 3, within the feature extraction phase, the input sequence, with a maximum length of 130 tokens, is preprocessed and passed through Word2Vec to derive sentence-level features, resulting in a 300-dimensional word vector. These vectors form an embedding matrix with dimensions of (130 × 300). The initial two layers of the encoder region consist of 1D convolutional layers. The Feature extraction section employs 1D convolutional layers with a kernel size of 3 to extract local features from the embedding matrix. The resulting local features are then normalized using LayerNormalization, subjected to nonlinear activation via ReLU, and mitigated for overfitting through a Dropout layer. To leverage sequential information from the data, the model employs two bidirectional LSTM (Bi-LSTM) layers with MaxPool1D and LayerNormalization, enabling the extraction of bidirectional features while preserving long-term dependencies. Following processing through the Bi-LSTM layers, the output consists of a sequence of context vectors corresponding to each time step. Rather than uniformly utilizing all these vectors, the Multi-Head Attention mechanism allows the model to selectively focus on and learn from the most salient information within the sequence.

The input to Multi-Head Attention consists of three matrices: Q (query), K (key), and V (value). These matrices are projected into lower-dimensional spaces through the learnable weight matrices , , corresponding to each attention head. For each attention head, the attention mechanism is computed as follows:

Where the attention function is defined as:

After that, the outputs from all heads are combined and projected through a weight matrix W0

The features obtained from the convolutional branch and the attention layer are processed through *GlobalMaxPooling1D*, then concatenated and normalized using *LayerNormalization*. Next, the data is passed through a series of Dense layers with 256, 128, and 64 units, respectively, using the ReLU activation function. Finally, the output layer consists of 3 units and applies the softmax function for classification. The model is trained with a learning rate of 1e-4, a batch size of 32, and the loss function used is cross-entropy.

The loss cross entropy is defined as:

Where:

* p(x) is the true probability of class x.
* q(x) is the probability predicted by the model for class x

## 2.4 Fine-Tuning the pretrained model for sentiment classification

The proposed model leverages the Transformer architecture, specifically fine-tuning PhoBERT for sentiment classification. It combines the encoder of PhoBERT, with hidden sizes of and 768, to handle deep semantic encoding. The CNN-LSTM component focuses on capturing sequential context and extracting local features. The CNN block comprises a 1D convolutional layer (256 channels, kernel size 3), followed by LayerNorm and GELU activation. The Gaussian Error Linear Unit (GELU) activation function is applied after the LayerNorm layer, introducing non-linearity, promoting sparsity, and ensuring smoothness, which enables the model to learn more complex representations and improves its generalization capabilities. The BiLSTM block, with 128 hidden units in each direction, processes bidirectional information in the data sequence. In the decoder block, a Fully Connected layer with 256 units is employed, and the output consists of 3 units, corresponding to the predicted labels.

A diagram of a layer

AI-generated content may be incorrect.

**Figure 4: The architecture of the PhoBERT + CNN-LSTM model**

During the fine-tuning process, the warm-up technique is used to minimize instability and convergence issues (Kalra and Barkeshli, 2024). The experiments on the PhoBERT models were conducted using the AdamW optimizer with warm-up for 10% of the training dataset. Following the warm-up phase, the model was trained with a linear scheduler, which gradually decreased the learning rate in a linear manner, allowing the model to converge better and avoid rapid learning rate decay. Both the PhoBERT + feed-forward model and the PhoBERT with CNN-LSTM model were configured with a learning rate of 2e-5, batch size of 64, with the loss function being cross-entropy.

# 3. RESULT AND DISCUSSION

**3.1 Evaluation metrics**

To evaluate the performance of the models, we use four popular metrics in the field of classification: Accuracy, Precision, Recall, and F1-score. These metrics were calculated using the formulas, where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

## 3.2 Result analysis

The experimental results indicate that Transformer-based models, particularly PhoBERT, outperform traditional machine learning methods and certain deep learning approaches. Among them, PhoBERT-v2 (Base) + CNN-LSTM achieved the highest performance with an F1-score of 92.92%. The LSTM-CNN model with an Attention mechanism achieved an F1-score of 90.63% on the original dataset, demonstrating its effectiveness in feature extraction for text classification. Among traditional machine learning models, Linear SVM achieved the highest accuracy of 88.91%, while Logistic Regression and XGBoost delivered a comparable performance with an accuracy of 88.31%. The experiments were conducted on the UIT-VSFC dataset, with comprehensive results presented in Table 4. PhoBERT-based models achieve superior performance over machine learning algorithms or deep learning models because they are the first large-scale monolingual language models specifically pre-trained for Vietnamese using a large (20GB). That is why the model can better capture context and emotional nuances.

**Table 4. Evaluation of Machine Learning and Deep Learning Models for Sentiment Analysis (%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **F1-Score** | **Recall** | **Precision** |
| **Logistic Regression** | 88.50 | 87.64 | 88.50 | 87.36 |
| **Linear SVM** | 88.91 | 87.64 | 88.91 | 88.12 |
| **XGBoost** | 88.31 | 87.38 | 88.31 | 87.24 |
| **CNN-LSTM + Attention** | 91.25 | 90.63 | 91.25 | 90.56 |
| **CNN-LSTM** | 89.29 | 89.09 | 89.29 | 88.93 |
| **PhoBert(Base) + FeedForward** | 92.99 | 92.41 | 92.99 | 92.43 |
| **PhoBert(Base) + CNN-LSTM** | 93.24 | 92.92 | 93.24 | 92.81 |

## 3.3 Comparison with other methods

Table 5 demonstrates that the proposed methods achieve performance that is either superior to or at least on par with previous state-of-the-art (SOTA) approaches. Deep learning models such as Bi-LSTM (Nguyen et al, 2018) and Bi-GRU + Attention (Trang and Hung, 2024) attained F1-scores ranging from 89% to 92%, indicating that RNN-based and Attention-enhance architectures provide certain benefits. Meanwhile, models utilizing pre-trained language models, such as XLM-R + VnEmoLex (Doan and Luu, 2022), have achieved higher performance, with an F1-score of approximately 93.97%, reflecting the growing trend of leveraging pre-trained Transformer models to enhance sentiment classification effectiveness.

**Table 5. Performance comparison with other Methods on UIT-VSFC (%)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Accuracy** | **F1-Score** | **Recall** | **Precision** |
| **BERT + CNN + BiLSTM + LSTM (Huynh et al, 2020)** | **-** | 92.79 | **-** | **-** |
| **Bi-GRU + Attention (Trang and Hung, 2024)** | 89.26 | 88.50 | 89.26 | 88.38 |
| **DNN-P (Vu Xuan et al, 2021)** | 88.56 | 89.90 | 94.18 | 86.00 |
| **XLM-R + VnEmoLex (Doan and Luu, 2022)** | **94.25** | **93.97** | **-** | **-** |
| **Bi-LSTM (Nguyen et al, 2018)** | **-** | 92.00 | 93.40 | 90.80 |
| **CNN-LSTM + Attention (Ours)** | 91.25 | 90.63 | 91.25 | 90.56 |
| **PhoBert(Base) + CNN-LSTM (Ours)** | 93.24 | 92.92 | 93.24 | 92.81 |

In this study, CNN-LSTM (Ours) and PhoBERT + CNN-LSTM (Ours), have shown significant improvements. The CNN-LSTM + Attention (Ours) model achieved an Accuracy of 91.25% and an F1-Score of 90.63%, which is comparable to or outperforms some previous studies. Notably, PhoBERT + CNN-LSTM (Ours) achieved the highest performance, with Accuracy, F1-Score, Recall, and Precision all around 93%. These results indicate that the combination of pre-trained language models (PhoBERT) with CNN/LSTM architectures enables the model to leverage deep contextual features and local feature extraction, thereby enhancing classification performance.Overall, the proposed methods not only keep up with but also have the potential to surpass previous state-of-the-art (SOTA) models. The combination of pre-trained language models with appropriate CNN/RNN architectures has demonstrated significant effectiveness in Vietnamese sentiment analysis. Future research directions may focus on experimenting with various datasets to assess the model's generalization ability, as well as optimizing deployment performance to enhance real-world applicability.

**Table 6. Sentiment Classification Results of CNN-LSTM + Attention and PhoBERT + CNN-LSTM**

| **Sentence** | | **Type** | **Label** | **CNN-LSTM + Attention** | **PhoBERT + CNN-LSTM** |
| --- | --- | --- | --- | --- | --- |
| 1 | Giáo viên không giảng dạy kiến thức , hướng dẫn thực hành trong quá trình học .  (Teachers do not provide instruction or practical guidance during the learning process.) | Raw | Negative | Negative | Negative |
| Giảng viên không giảng dạy kiến thức hướng dẫn thực hành trong quá trình học  (The lecturer do not provide instruction or practical guidance during the learning process.) | Standardized | **Negative** | **Negative** | **Negative** |
| 2 | Cụ thể doubledot về việc hướng dẫn bài tập về nội dung đó.  (Specifically, doubledot regarding guiding exercises on that content.) | Raw | Neutral | Negative | Positive |
| Cụ thể về việc hướng dẫn bài tập về nội dung.  (Specifically, regarding guiding exercises on the content.) | Standardized | **Neutral** | **Positive** | **Neutral** |
| 3 | Cô vui tính colonsmilesmile.  (She is cheerful colonsmilesmile.) | Raw | Positive | Positive | Positive |
| Giảng viên vui tính smile  (The lecturer is cheerful smile.) | Standardized | **Positive** | **Positive** | **Positive** |

Table 6 presents the sentiment classification results of our models: CNN-LSTM with Attention and PhoBERT with CNN-LSTM, applied to Vietnamese student feedback. The input data includes both raw and standardized sentences to evaluate the impact of preprocessing on model performance. The results show that the models generally align with the original sentiment labels in straightforward cases. For example, the sentence "Giáo viên không giảng dạy kiến thức, hướng dẫn thực hành trong quá trình học." (Teachers do not provide instruction or practical guidance during the learning process.) was correctly classified as negative by both models. However, discrepancies appear in cases where sentences exhibit neutral sentiment or require deeper contextual understanding. The disparity in performance can be attributed to the fundamental differences in the capabilities of the two models. Specifically, the CNN-LSTM + Attention model relies heavily on task-specific training data to learn patterns, whereas the PhoBERT + CNN-LSTM model leverages PhoBERT, a transformer model that has been pre-trained on a vast Vietnamese corpus. This pre-training enables PhoBERT to possess a deeper understanding of Vietnamese semantics and context, thereby enhancing its ability to capture the nuances of the language even before undergoing task-specific training. For instance, the sentence "Cụ thể về việc hướng dẫn bài tập về nội dung đó." (Specifically, regarding guiding exercises on that content.) was originally labeled as neutral.. The CNN-LSTM + Attention model misclassified it as negative, while the PhoBERT + CNN-LSTM model predicted a positive sentiment. Neutral statements, often lacking strong sentiment keywords, particularly challenge models to rely on this deeper contextual interpretation, where pre-trained models like PhoBERT may hold an advantage. On the other hand, in the case of positive sentiment, both models classified correctly. For example, consider the sentence "Cô vui tính colonsmilesmile (She is cheerful colonsmilesmile.)", and its standardized version "Giảng viên vui tính smile.". (The lecturer is cheerful smile.) were both labeled as positive, and both models predicted correctly. These results highlight variations in model performance when processing student feedback, especially for neutral statements. The input standardization can influence model predictions, sometimes leading to shifts in sentiment classification.

# 4. Conclusion

In this study, we analyzed student feedback from a university in Vietnam to classify sentiments into three categories: positive, negative, and neutral. Using machine learning and deep learning techniques on the UIT-VSFC dataset, we achieved promising results. Our PhoBERT + CNN-LSTM models performed best, reaching F1-Score 92.92%. These findings highlight the effectiveness of Transformer-based models, especially our customized versions of PhoBERT, in processing Vietnamese text. More importantly, they offer valuable insights for educators, helping institutions better understand student experiences and refine teaching methods to improve education quality. In the future, we aim to expand the dataset, optimize models, and explore additional factors affecting student emotions to further enhance learning outcomes.

**DISCLAIMER (ARTIFICIAL INTELLIGENCE)**Author(s) hereby declare that NO generative AI

technologies such as Large Language Models

(ChatGPT, COPILOT, etc.) and text-to-image

generators have been used during the writing or

editing of this manuscript.

**COMPETING INTERESTS**

Author has declared that no competing interests

**ACKNOWLEDGMENTS**

This research is funded by Ho Chi Minh City University of Education Foundation for Science and Technology under the student scientific research project for the academic year 2024–2025

# REFERENCE

Giang, N. T. P., Dien, T. T., & Khoa, T. T. M. (2020). Sentiment Analysis for University Students’ Feedback. In K. Arai, S. Kapoor, & R. Bhatia (Eds.), *Advances in Information and Communication* (Vol. 1130, pp. 55–66). Springer International Publishing. <https://doi.org/10.1007/978-3-030-39442-4_5>  
  
Stupans, I., McGuren, T., & Babey, A. M. (2016). Student Evaluation of Teaching: A Study Exploring Student Rating Instrument Free-form Text Comments. *Innovative Higher Education*, *41*(1), 33–42. <https://doi.org/10.1007/s10755-015-9328-5>

Nguyen, K. V., Nguyen, V. D., Nguyen, P. X. V., Truong, T. T. H., & Nguyen, N. L.-T. (2018). UIT-VSFC: Vietnamese Students’ Feedback Corpus for Sentiment Analysis. *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 19–24. <https://doi.org/10.1109/KSE.2018.8573337>

Kastrati, Z., Dalipi, F., Imran, A. S., Pireva Nuci, K., & Wani, M. A. (2021). Sentiment Analysis of Students’ Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences*, *11*(9), 3986. <https://doi.org/10.3390/app11093986>

Chebolu, S. U. S., Dernoncourt, F., Lipka, N., & Solorio, T. (2022). *Survey of Aspect-based Sentiment Analysis Datasets* (Version 5). arXiv. <https://doi.org/10.48550/ARXIV.2204.05232>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1810.04805>

Kaur, G., & Sharma, A. (2023). A deep learning-based model using hybrid feature extraction approach for consumer sentiment analysis. *Journal of Big Data*, *10*(1), 5. <https://doi.org/10.1186/s40537-022-00680-6>

Si, K. (2025). Research on Sarcastic Emotion Recognition Based on Multiple Feature Fusion. *ITM Web of Conferences*, *70*, 02008.<https://doi.org/10.1051/itmconf/20257002008>

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Springer International Publishing. <https://doi.org/10.1007/978-3-031-02145-9>

Al Katat, S., Zaki, C., Hazimeh, H., El Bitar, I., Angarita, R., & Trojman, L. (2024). Natural Language Processing for Arabic Sentiment Analysis: A Systematic Literature Review. *IEEE Transactions on Big Data*, *10*(5), 576–594. <https://doi.org/10.1109/TBDATA.2024.3366083>

Sáu, T. N. T., Sang, Đ. P., & Trang, P. T. T. (2021). PHÂN TÍCH Ý KIẾN THEO KHÍA CẠNH TRÊN BÌNH LUẬN PHẢN HỒI CỦA SINH VIÊN CHO TIẾNG VIỆT. *TNU Journal of Science and Technology*, *226*(18), 48–55. <https://doi.org/10.34238/tnu-jst.5101>

M.-N. Tran, P.-V. Nguyen, L. Nguyen, and D. Dinh, “ViGLUE: A Vietnamese General Language Understanding Benchmark and Analysis of Vietnamese Language Models,” in *Findings of the Association for Computational Linguistics: NAACL 2024*, Mexico City, Mexico: Association for Computational Linguistics, 2024, pp. 4174–4189. doi: [10.18653/v1/2024.findings-naacl.261](https://doi.org/10.18653/v1/2024.findings-naacl.261).  
  
Nguyen, M.-H., Nguyen, T. M., Van Thin, D., & Nguyen, N. L.-T. (2019). A corpus for aspect-based sentiment analysis in Vietnamese. *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, 1–5.<https://doi.org/10.1109/KSE.2019.8919448>  
  
Tran, Q.-L., Thanh, P., Le, D., & Do, T.-H. (n.d.). *Aspect-based Sentiment Analysis for Vietnamese Reviews about Beauty Product on E-commerce Websites*. <https://aclanthology.org/2022.paclic-1.84.pdf>

Ho, V. A., Nguyen, D. H.-C., Nguyen, D. H., Pham, L. T.-V., Nguyen, D.-V., Van Nguyen, K., & Nguyen, N. L.-T. (2019). *Emotion Recognition for Vietnamese Social Media Text*. <https://doi.org/10.48550/ARXIV.1911.09339>

Nguyen, D. Q., & Nguyen, A. T. (2020). *PhoBERT: Pre-trained language models for Vietnamese* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2003.00744>

Sparck Jones, K. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, *28*(1), 11–21. <https://doi.org/10.1108/eb026526>

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.1301.3781>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (Version 7). arXiv. <https://doi.org/10.48550/ARXIV.1706.03762>

Kalra, D. S., & Barkeshli, M. (2024). *Why Warmup the Learning Rate? Underlying Mechanisms and Improvements* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2406.09405>

Huynh, H. D., Do, H. T.-T., Van Nguyen, K., & Nguyen, N. L.-T. (2020). *A Simple and Efficient Ensemble Classifier Combining Multiple Neural Network Models on Social Media Datasets in Vietnamese* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2009.13060>

Trang, N. T. M., & Hung, N. N. (2024). BiGRU-Attention Sentiment Analysis for Enhancing Task-Oriented Chatbots. *2024 5th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI)*, 476–483. <https://doi.org/10.1109/ICMCSI61536.2024.00075>

Vu Xuan, Q. D., Laosen, K., & Laosen, N. (2021). An Evaluation of the UIT-VSFC Dataset Using Modern Machine Learning Techniques and Word Embeddings. *2021 25th International Computer Science and Engineering Conference (ICSEC)*, 394–399. <https://doi.org/10.1109/ICSEC53205.2021.9684597>

Doan, A. L., & Luu, S. T. (2022). Improving Sentiment Analysis By Emotion Lexicon Approach on Vietnamese Texts. *2022 International Conference on Asian Language Processing (IALP)*, 39–44. <https://doi.org/10.1109/IALP57159.2022.9961318>