CTU Journal of Innovation and Sustainable Development
ISSN 2588-1418 | e-ISSN 2815-6412

DOI:10.22144/ctujoisd.2024.322

Fine-tuned PhoBERT for sentiment analysis of Vietnamese phone reviews

Ngo Minh Tan[1,2], Ngo Ba Hung[1*], and Stuchilin Vladimir Valerievich[2]

[1]*College of Information and Communication Technology, Can Tho University, Viet Nam*

[2]*Institute of Computer Science, NUST MISIS, Russia*

**Corresponding author (nbhung@ctu.edu.vn)*

Article info.

Received 15 Jun 2024
Revised 30 Aug 2024
Accepted 3 Oct 2024

Keywords

Fine-tuned PhoBERT, natural language processing, sentiment analysis, text classification, Vietnamese language

ABSTRACT

This paper presents an exploration of sentiment analysis applied to Vietnamese phone reviews, leveraging the PhoBERT model. While significant advancements have been made in sentiment analysis for English and other widely spoken languages, Vietnamese remains relatively under investigated. Our study addresses this gap by constructing a comprehensive dataset that integrates data from the UIT-ViSFD dataset and data collected through web scraping. We experimented with various models including naive Bayes, Support Vector Machine, and PhoBERT, utilizing multiple data preprocessing techniques. PhoBERT, a state-of-the-art pre-trained language model specifically designed for Vietnamese, demonstrated superior performance. The final PhoBERT model with optimized preprocessing achieved an accuracy of 92.74%, highlighting its efficacy in accurately identifying sentiments.

1. INTRODUCTION

In recent years, the rapid growth of smartphone usage in Vietnam has been accompanied by an increasing volume of online reviews and feedback from users. There were over 57 million smartphone users in Vietnam in 2023 (Start.io, 2024), constituting a significant portion of the population. The number of smartphone users is projected to rise steadily from 2024 to 2029, with an anticipated growth of 12.7 million users, marking a 15.04 percent increase. By 2029, the smartphone user base is expected to reach a peak of 97.19 million users, following nineteen consecutive years of growth since 2017 (Statista, 2024). This trend highlights the consistent increase in the number of smartphone users over recent years.

Sentiment analysis, a subset of natural language processing (NLP), involves extracting subjective information from text to determine the sentiment expressed. It has become an essential tool for businesses to analyse customer feedback, improve products, and enhance user satisfaction (Medhat et al., 2014). Despite its widespread application in English and other major languages, sentiment analysis in Vietnamese remains underdeveloped due to the complexities of the language and the scarcity of annotated datasets (Hoang et al., 2007; 1StopAsia, 2024). Therefore, our research focuses on advancing sentiment analysis techniques to enhance the processing of product reviews, with a particular emphasis on phone reviews.

In this study, we present a novel framework for sentiment analysis of Vietnamese phone reviews, which integrates PhoBERT (Nguyen et al., 2020) with various preprocessing techniques. We evaluate the performance of a sentiment analysis model utilizing the PhoBERT pre-trained model tailored for Vietnamese text and explore multiple preprocessing strategies to determine the optimal configuration. To demonstrate the effectiveness of our approach, we employ Support Vector Machines (SVM) and naive Bayes (NB) classifiers. These

models were chosen due to their proven efficacy in text classification tasks, providing a baseline for assessing the performance gains achieved through the use of a transformer-based model like PhoBERT.

The remainder of this paper is organized as follows: In Section 2, we review related work on sentiment analysis within the Vietnamese context. Section 3 outlines our methodology, focusing on the integration of PhoBERT with various preprocessing techniques for sentiment analysis. Section 4 describes the experimental setup, presents the results, and provides a detailed analysis of our classification approach. Finally, Section 5 concludes the paper by summarizing key findings and potential directions for future research.

2. RELATED WORKS

Sentiment analysis for the Vietnamese language has seen various advancements over the past decade. Early efforts, such as those by Kieu and Pham (2010), focused on document-level sentiment analysis using rule-based systems for Vietnamese. This pioneering work addressed sentiment at the sentence level, using a corpus of computer product reviews. More recent studies have incorporated advanced machine learning techniques to enhance accuracy and efficiency. For instance, Nguyen et al. (2018) compared traditional machine learning classifiers with deep learning models on a corpus of Vietnamese students' feedback, finding that the Bi-Directional Long Short-Term Memory (BiLSTM) outperformed other algorithms.

Aspect-based sentiment analysis (ABSA) has also been an area of growing interest. Nguyen et al. (2019) introduced a new annotated corpus for ABSA, specifically targeting restaurant reviews in Vietnamese, achieving notable F1-scores for aspect detection and polarity detection. This work highlights the potential of supervised learning methods to handle the complexities of sentiment analysis in Vietnamese. Similarly, Le et al. (2022) applied the pre-trained PhoBERT model to the Vietnamese Smartphone Feedback Dataset (UIT-ViSFD), demonstrating its superiority over other transformer-based models in terms of macro-F1 scores for both aspect and sentiment detection tasks.

In the specific context of phone reviews, sentiment analysis has been explored using various methodologies. Shaheen et al. (2019) conducted a comprehensive analysis of mobile phone reviews fetched from Amazon.com, employing multiple sentiment classification algorithms. Their study provided a comparative analysis of the performance of different classifiers, including Random Forest and Long Short-Term Memory (LSTM), highlighting the effectiveness of these models in predicting customer ratings based on user reviews. Moreover, Yiran and Srivastava (2019) utilized Latent Dirichlet Allocation (LDA) for aspect-based sentiment analysis on a large dataset of phone reviews, showcasing the utility of topic modelling in extracting meaningful information from unstructured data.

Specifically focusing on Vietnamese phone reviews, the UIT-ViSFD dataset has been instrumental. Phan et al. (2023) developed a social listening system using aspect-based sentiment analysis based on this dataset. Their approach, based on a Bi-LSTM architecture with fastText embeddings, achieved notable F1-scores for aspect and sentiment tasks, underscoring the dataset's value in facilitating advanced NLP research. Our study builds upon these foundations by fine-tuning the PhoBERT model on a newly created dataset that combines UIT-ViSFD data with additional user reviews collected from online sources. This integration aims to improve the model's performance in sentiment analysis, addressing the unique challenges posed by the unstructured and diverse nature of online reviews.

3. MATERIALS AND METHOD

3.1. Overview

In this study, we propose a comprehensive framework that integrates extensive preprocessing techniques with PhoBERT for sentiment analysis on phone reviews. This section details the preprocessing techniques we experimented with and identifies the final framework that yielded the best results. The process, from input sentence to output label, encompasses three main stages: preprocessing, learning, and evaluation. Figure 1 illustrates the implementation of our method. We provide an in-depth description of the framework through these stages, as outlined in Sections 3-2, 3-3, and 3-4 respectively.

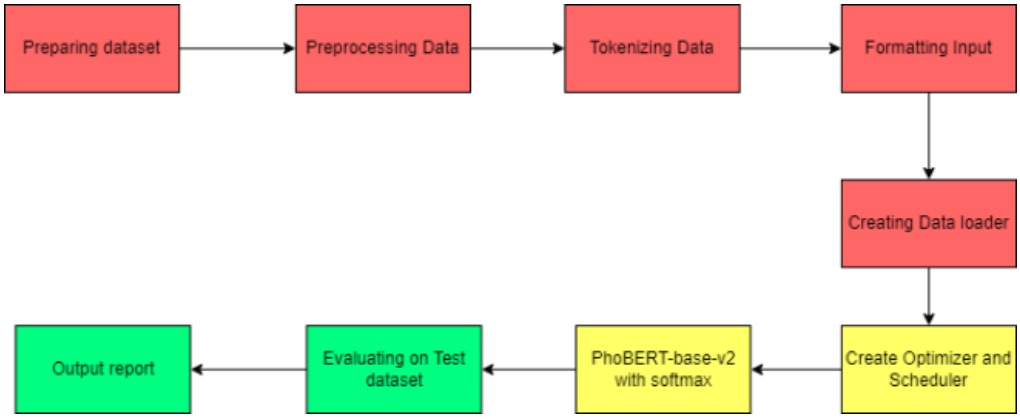

Figure 1. The entire process of our method includes preprocessing steps highlighted in red, the learning stage marked in yellow, and the evaluation process depicted in green

3.2. Preprocessing stage

3.2.1. Preparing dataset

Our sentiment analysis system produces binary outputs, classifying sentiments as either positive or negative. The dataset we used consists of 44,072 entries, combining the publicly available UIT-ViSFD dataset and an additional 32,950 reviews obtained through web scraping from phone retailer websites. The UIT-ViSFD dataset contains 11,122 entries with four features: comment (content), n_star (user's star rating), date_time (posting date and time), and label (sentiment). However, we dropped the date_time feature, as it did not provide relevant information for sentiment analysis, and the label feature, since it followed a different labeling system.

For the web-scraped dataset, we collected two features: review (content) and star (user's star rating). The web-scraped reviews were sourced from popular Vietnamese phone retailer websites, focusing specifically on product reviews for smartphones. This selection ensures that both datasets are contextually consistent and relevant to the target domain of phone-related sentiment analysis. Moreover, the process of web scraping followed a systematic approach, collecting data from a set of uniform sources, thus providing additional data that complements the UIT-ViSFD dataset.

To create a unified dataset, we merged the comment and review columns and combined the n_star and star columns. Finally, we added a label column to reflect the sentiment, where reviews with 1 to 3 stars were labeled as negative (0) and those with 4 or 5 stars as positive (1). After merging both datasets and standardizing the feature set, our final dataset consists of three key features: comment/review content, n_star/star rating, and sentiment label. The final distribution of sentiments in the dataset is 71.2% positive and 28.8% negative.

3.2.2. Data preprocessing

Our preprocessing steps are crucial for preparing the dataset for sentiment analysis. Figure 2 illustrates the workflow. Below is a brief overview of each preprocessing step:

– **Lower Case and Removal of Null Values**: All text is converted to lower case to ensure uniformity. Zero values and irrelevant data points are removed to clean the dataset.

– **Text Standardization**: This step is critically important and the most challenging because online reviews often lack consistent formatting and contain numerous slangs and abbreviations that vary based on reviewer preferences. We standardized the text by correcting spelling errors, normalizing text formats, and converting variations of the same word to a common format. A specialized dictionary for emojis and slangs/abbreviations, specifically tailored for phone reviews, is used in this step. Example: “đt”, “dt” and “dthoai” are standardized to "điện thoại" <telephone>.

– **Converting Emoticons and Punctuation Marks**: Emoticons and punctuation marks are converted to meaningful words using the same specialized dictionary.

– **Text Segmentation**: The text is segmented into sentences or tokens to facilitate analysis. This step is essential for breaking down the text into manageable pieces. Due to the distinctions in Vietnamese syllables, PhoBERT uses the Vietnamese Word Segmenter from VNCoreNLP

(Sennrich et al., 2016) prior to applying Byte-Pair Encoding (BPE) methods (Vu et al., 2018). Example: “điện thoại này tuyệt vời” <this phone is great> is tokenized into [“điện_thoại”, "này", "tuyệt_vời"].

– **Deleting Short Entries**: Removing very short entries that may not provide enough context or information for sentiment analysis.

By incorporating these preprocessing techniques, we enhance the quality and consistency of the dataset, making it more suitable for training the PhoBERT model. After completing the preprocessing steps, the dataset now consists of 37575 entries. To ensure consistency, experiments with the unprocessed data will randomly sample 37575 entries to match the size of the preprocessed dataset.

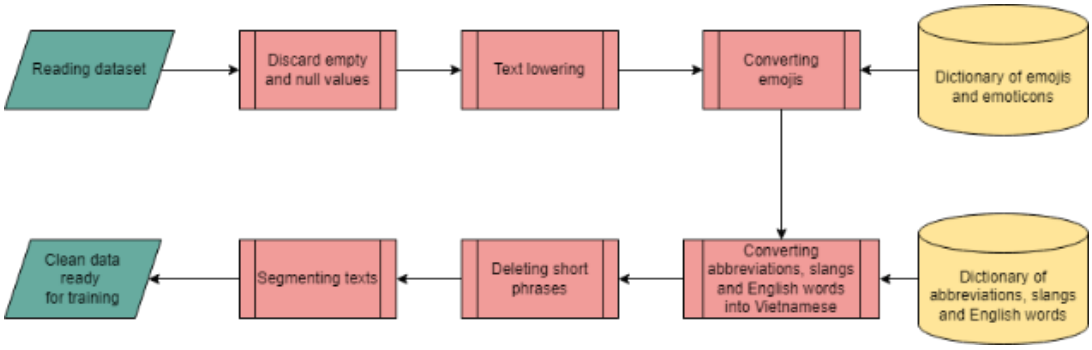

Figure 2. The workflow for the preprocessing steps is illustrated, with steps highlighted in red, dictionaries in yellow, and the input and output data in green

3.2.3. Tokenizing and formatting data

In this step, we first load the dataset, consisting of review texts and their corresponding labels. The data is split into training and test sets, ensuring a robust evaluation of the model. For tokenization, the PhoBERT-base-v2 tokenizer is employed to process the text data. This tokenizer converts the text into input IDs and attention masks suitable for PhoBERT. The texts are truncated or padded to a maximum length of 256 tokens to maintain consistency.

The tokenized data is then converted into PyTorch tensors, which are essential for efficient processing during model training. Training and test data are prepared using DataLoader objects, with the training set shuffled and sampled randomly to enhance model learning, while the test set is sequentially sampled for accurate evaluation. This setup ensures that the data is properly formatted and ready for fine-tuning the PhoBERT model.

3.3. Learning stage

Originally, PhoBERT is available in two versions: PhoBERT-base and PhoBERT-large. These models comprise approximately 135 million and 370 million parameters, respectively. Recently, an updated version, PhoBERT-base-v2, was introduced. This new version incorporates additional training data and improvements over the original PhoBERT-base. Specifically, PhoBERT-base-v2 includes the same 20GB of Wikipedia and News texts as the original PhoBERT-base but adds 120GB of texts from OSCAR-2301 dataset. Given these enhancements, we employ PhoBERT-base-v2 in our study, as its additional training data offers a more comprehensive and nuanced understanding of Vietnamese compared to previous versions. Additionally, its novelty and limited prior experimentation make it a compelling choice for our research.

We employ the AdamW optimizer with weight decay and a learning rate scheduler. This combination allows for efficient training with gradual learning rate adjustments, promoting stable convergence. The cross-entropy loss function is used for its suitability in binary classification tasks, and the softmax activation function in the final layer converts model outputs into class probabilities, ensuring accurate sentiment predictions.

3.4. Learning stage

We assess the model’s performance using accuracy and F1-score metrics, which are critical for understanding both the precision and recall of our classifier. These metrics are calculated after each epoch and averaged across all folds to provide a comprehensive evaluation.

4. EXPERIMENTAL RESULTS

4.1. Configurations

For all data preprocessing, training, and evaluation stages, we use Google Colaboratory, a free tool provided by Google. It offers a robust configuration, including:

- Nvidia Tesla K80 GPU
- 12GB of RAM
- Intel(R) Xeon(R) CPU @ 2.00GHz.

4.2. Experimental details

In our experiments, we evaluated the performance of three models: PhoBERT-base-v2, naive Bayes, and SVM. For PhoBERT, we used the pre-trained vinai/phobert-base-v2 model, implemented as a RobertaForSequenceClassification classifier for binary sentiment classification. We fine-tuned the model on our dataset for 5 epochs, using the AdamW optimizer with a learning rate of 2e-5 and an epsilon of 1e-8. A linear learning rate scheduler with warm-up steps was applied during training. The model was trained and evaluated on a GPU using a standard procedure of gradient clipping and step-based optimizer updates.

For the naive Bayes and SVM models, we used implementations from the scikit-learn library and applied the TfidfVectorizer for feature extraction, limiting the vocabulary to the 3000 most relevant terms. We used a linear kernel in the SVM with a regularization parameter C=1.0, chosen for its efficiency in text classification tasks. For naive Bayes, we employed the Multinomial naive Bayes classifier, which is well-suited for handling text data transformed by the TF-IDF technique.

We ran all three models on both unprocessed and preprocessed datasets to compare their effectiveness. The preprocessing steps are according to section 4. By evaluating the models on both unprocessed and preprocessed data, we aimed to determine the impact of preprocessing on model performance.

4.3. Results with unprocessed data

The results highlight how PhoBERT has an inherent capability to handle unprocessed Vietnamese text better than traditional machine learning models. Despite the effectiveness of these models, the presence of noise, slang, and inconsistent formatting in unprocessed phone reviews limited their overall performance. The F1-score and accuracy of models are shown in Table 1.

Table 1. Experimental results on unprocessed dataset

Model	Accuracy	F1-score
Naive Bayes	83.67%	85.66%
SVM	85.96%	85.95%
PhoBERT	89.33%	89.33%

4.4. Results with preprocessed data

Upon applying extensive preprocessing techniques to clean and standardize the dataset, the performance of all models improved significantly. The preprocessing steps, including text standardization, emoji and slang conversion, and text segmentation, were crucial in refining the dataset. Text standardization addressed inconsistencies in review formats, emoji and slang conversion normalized colloquial language, and text segmentation improved the granularity of textual analysis. By refining the dataset and reducing noise, these steps provided clearer and more consistent input for the models.

These steps addressed the variations in online reviews, leading to clearer and more consistent input for the models. The notable increase in accuracy underscores the importance of preprocessing in enhancing model performance, particularly for specialized language tasks. The F1-score and accuracy of models are shown in Table 2.

Table 2. Experimental results on preprocessed dataset

Model	Accuracy	F1-score
Naive Bayes	85.25%	85.23%
SVM	87.97%	87.96%
PhoBERT	92.74%	92.72%

5. CONCLUSION

In this study, we developed a robust framework for sentiment analysis of Vietnamese phone reviews using the PhoBERT-base-v2 model. The integration of comprehensive preprocessing techniques significantly improved dataset quality and model performance, as evidenced by substantial gains in accuracy and F1 scores across all models. This work demonstrates the critical role of preprocessing in refining data for complex language tasks. Future research could focus on expanding the framework's application to other Vietnamese NLP tasks, such as sentiment analysis in different domains (e.g., social media, product reviews) and opinion mining. Additionally, exploring adaptations of this framework for multilingual sentiment analysis could further enhance its applicability and impact in diverse linguistic contexts.

REFERENCES

1StopAsia. (2024). *Difficulties with developing NLP for Vietnamese*. 1StopAsia. https://www.1stopasia.com/blog/challenges-developing-nlp-for-vietnamese/

Hoang, V. C. D., Dinh, D., Nguyen, N. L., & Ngo, H. Q. (2007). A comparative study on Vietnamese text classification methods. In *Proceedings of the 2007 IEEE International Conference on Research, Innovation and Vision for the Future* (pp. 267–273). IEEE. https://doi.org/10.1109/RIVF.2007.369167

Kieu, B. T., & Pham, S. B. (2010). Sentiment analysis for Vietnamese. In *Proceedings of the 2010 Second International Conference on Knowledge and Systems Engineering* (pp. 7–9). IEEE. https://doi.org/10.1109/KSE.2010.33

Le, B. H., Nguyen, H. M., Nguyen, N. K. P., & Nguyen, B. T. (2022). A new approach for Vietnamese aspect-based sentiment analysis. In *Proceedings of the 2022 14th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 19–21). IEEE. https://doi.org/10.1109/KSE56063.2022.9953759

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Nguyen, D. Q., & Nguyen, T. A. (2020). PhoBERT: Pre-trained language models for Vietnamese. In T. Cohn, Y. He, & Y. Liu (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1037–1042). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.92

Nguyen, M. H., Nguyen, T. M., & Nguyen, D. V. (2019). A corpus for aspect-based sentiment analysis in Vietnamese. In *Proceedings of the 2019 11th International Conference on Knowledge and Systems Engineering (KSE)* (pp. 24–26). IEEE. https://doi.org/10.1109/KSE.2019.8919448

Nguyen, P. X. V., Hong, T. T. T., Nguyen, K. V., & Nguyen, N. L. T. (2018). Deep learning versus traditional classifiers on Vietnamese students feedback corpus. In *Proceedings of the 5th NAFOSTED Conference on Information and Computer Science (NICS)* (pp. 23–24). IEEE. https://doi.org/10.1109/NICS.2018.8606837

Phan, L. L., Pham, P. H., Nguyen, K. T. T., Huynh, S. K., Nguyen, T. T., Nguyen, L. T., & Huynh, T. V. (2023). SA2SL: From aspect-based sentiment analysis to social listening system for business intelligence. In H. Qiu, C. Zhang, Z. Fei, M. Qiu, & S. Y. Kung (Eds.), *Knowledge Science, Engineering and Management. KSEM 2021. Lecture Notes in Computer Science* (Vol. 12816, pp. 662–677). Springer. https://doi.org/10.1007/978-3-030-82147-0_53

Sennrich, R., Haddow, B., & Birch, A. (2016). Neural machine translation of rare words with subword units. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1715–1725). Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1162

Shaheen, M., Awan, S. M., Hussain, N., & Gondal, Z. A. (2019). Sentiment analysis on mobile phone reviews using supervised learning techniques. *International Journal of Modern Education and Computer Science*, *7*, 32–43. https://doi.org/10.5815/ijmecs.2019.07.04

Start.io. (2024). *Smartphone users in Vietnam*. Start.io. https://www.start.io/audience/smartphone-users-in-vietnam

Statista. (2024). *Number of mobile internet users in Vietnam from 2010 to 2029*. Statista. https://www.statista.com/forecasts/1147340/mobile-internet-users-in-vietnam

Vu, T., Nguyen, D. Q., Nguyen, D. Q., Dras, M., & Johnson, M. (2018). VnCoreNLP: A Vietnamese natural language processing toolkit. In Y. Liu, T. Paek, & M. Patwardhan (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 56–60). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-5012

Yiran, Y., & Srivastava, S. (2019). Aspect-based sentiment analysis on mobile phone reviews with LDA. In *Proceedings of the 4th International Conference on Machine Learning Technologies (ICMLT)*. ACM. https://doi.org/10.1145/3340997.3341012