A Comparative Study of Fine-Tuning BERT, BERT+LSTM, and BERT+RCNN for Vietnamese Sentiment Analysis

Hoang Bao Viet, Nguyen Duc Gia Phuc, Nguyen Quang Tuan

The University of Danang, Vietnam - Korea University of Information and Communication Technology

Abstract

Sentiment analysis is a fundamental task in Natural Language Processing (NLP) that focuses on identifying and evaluating opinions expressed in text. This study explores the application of BERT-based models for sentiment analysis of Vietnamese user comments. We investigate and compare three fine-tuning approaches: the standard BERT model using the [CLS] token as input to a feed-forward neural network, a hybrid BERT+LSTM model, and a novel BERT+RCNN architecture that utilizes the full sequence of output vectors for classification. Experiments conducted on two Vietnamese sentiment datasets demonstrate that models leveraging BERT significantly outperform traditional deep learning baselines. Notably, our proposed BERT+RCNN approach consistently achieves the best performance, surpassing the standard BERT fine-tuning method in both accuracy and robustness.

Keywords: BERT, LSTM, RCNN, Vietnamese, Sentiment Analysis

1 Introduction

In today's digital era, millions of user-generated comments are shared daily on social networks and e-commerce platforms. These comments provide valuable feedback for both consumers and service providers. However, due to the vast amount of data, it becomes impractical for humans to manually process and analyze such information. Therefore, there is a growing need for automated systems capable of identifying and evaluating users' opinions.

Sentiment analysis is a fundamental task in Natural Language Processing (NLP) that aims to determine the emotional tone behind textual content. A sentiment classification system can either predict fine-grained sentiment scores (e.g., from 1 to 5) or perform binary classification to detect whether the content expresses a *positive* or *negative* sentiment. In this study, we focus on the latter task.

In 2018, Devlin et al. [1] introduced BERT (Bidirectional Encoder Representations from Transformers), a pre-trained language representation model that has significantly advanced the state of the art across various NLP tasks. BERT captures the context of words by considering both left and right contexts simultaneously, making it particularly effective for understanding the sentiment expressed in text.

This paper investigates the application of BERT-based models to sentiment analysis in Vietnamese. We present and compare two approaches. The first is the original fine-tuning strategy introduced by Devlin et al., which uses the representation of the `[CLS]` token as input to a classification layer. The second approach, proposed in this paper, extends BERT by combining its contextual embeddings with additional neural architectures, including Long Short-Term Memory (LSTM) and Recurrent Convolutional Neural Network (RCNN) layers.

Our objective is to evaluate the effectiveness of these combined models and to demonstrate how different fine-tuning strategies can improve sentiment classification performance on Vietnamese datasets.

The remainder of this paper is organized as follows:

- Section 2 reviews related work in sentiment analysis.
- Section 3 introduces word embeddings, language models, and the BERT architecture.
- Section 4 describes our two BERT-based fine-tuning methods.
- Section 5 presents the experimental setup and results.
- Section 6 concludes the paper.

2 Related Work

Sentiment analysis is a subtask of text classification in which the goal is to categorize textual data into sentiment classes, typically *positive* or *negative*. Early research in this field primarily relied on traditional machine learning techniques. One of the pioneering studies was conducted in 2002, where reviews were classified into positive and negative categories [2]. This work utilized supervised learning models such as Support Vector Machines (SVM) [3] and Naive Bayes classifiers [4] to perform sentiment classification. Another common approach employed sentiment lexicons, which use predefined dictionaries of emotionally charged words annotated with sentiment polarity and intensity [5].

With the rise of deep learning, sentiment analysis has seen significant advancements through improved word and context representations. Kim [6] proposed the use of Convolutional Neural Networks (CNNs) for sentence classification, treating text as character-level input sequences. Mikolov et al. [7] introduced the Paragraph Vector model (also known as Doc2Vec), an unsupervised method for learning fixed-length representations of variable-length texts. Unlike traditional bag-of-words models, this approach learns document-level embeddings that capture semantic meaning [8].

In the context of Vietnamese language, Duyen et al. [9] applied traditional classifiers such as Naive Bayes, Maximum Entropy, and SVM to review classification tasks on Agoda, a hotel booking platform. Their results indicated that the SVM model outperformed other approaches. On the deep learning front, Quan et al. [10] proposed a hybrid architecture combining Long Short-Term Memory (LSTM) and CNN, named Multi-Channel LSTM-CNN, for Vietnamese sentiment analysis. This model outperformed standalone CNN and LSTM models. A similar approach was introduced in [11], where word vectors were first processed by a CNN layer, and the resulting feature maps were passed to an LSTM network for final sentiment classification. This pipeline demonstrated strong performance, especially for handling negative comments on social media platforms.

These studies demonstrate the evolution of sentiment analysis methods—from traditional machine learning and lexicon-based approaches to more powerful neural network architectures. However, few works have explored the integration of pre-trained language models like BERT for Vietnamese sentiment classification, which motivates our study.

3 Background

3.1 Word Embedding

Word embedding refers to the process of representing words as dense vectors in a continuous vector space, typically of much lower dimensionality than the vocabulary size. This representation captures semantic and syntactic relationships between words. One of the earliest and most influential approaches is Word2Vec, proposed by Mikolov et al. [8], which learns word representations by predicting surrounding words in a sentence using shallow neural networks.

Another widely used method is GloVe (Global Vectors for Word Representation), introduced by Pennington et al. [12], which combines the advantages of global matrix factorization and local context window methods to efficiently learn word vectors from large corpora.

More recently, contextual word embedding methods have emerged, with BERT (Bidirectional Encoder Representations from Transformers) [13] being a prominent example. Unlike Word2Vec or GloVe, which produce a single vector per word, BERT generates dynamic word representations that depend on the context in which the word appears.

3.2 Language Model

A language model defines a probability distribution over sequences of words. Given a sequence $x_1, x_2, ..., x_n$, the language model estimates the joint probability:

$$P(x_1, x_2, ..., x_n)$$

where n is the length of the sentence. Language models are essential for many NLP tasks such as machine translation, speech recognition, and text generation, as they can measure how likely a given sentence is within a language.

With the emergence of deep learning, large-scale pre-trained language models have become the foundation of modern NLP. These models are trained on massive text corpora and capture a wide range of linguistic knowledge. Once pre-trained, they can be fine-tuned on specific downstream tasks, leading to state-of-the-art performance in sentiment analysis, question answering, and more.

3.3 BERT

BERT is a deep bidirectional transformer-based language model introduced by Devlin et al. [1]. It is built entirely on encoder layers from the original Transformer architecture [13], omitting the decoder blocks as they are not necessary for language understanding tasks.

Unlike previous models that process text in a left-to-right or right-to-left manner, BERT uses a masked language modeling (MLM) strategy to learn deep bidirectional representations. This allows BERT to better capture context and semantic nuances by simultaneously considering both left and right contexts of a word.

One of BERT's key innovations is applying the self-attention mechanism bidirectionally during pre-training, significantly improving the quality of learned representations. As a result, BERT has set new benchmarks on multiple NLP tasks.

Two main configurations of BERT were introduced:

- **BERT$_{\textbf{BASE}}$**: 12 Transformer encoder layers, 12 self-attention heads, and 110 million parameters.
- **BERT$_{\textbf{LARGE}}$**: 24 Transformer encoder layers, 16 self-attention heads, and 340 million parameters.

4 Methodology

In this study, we explore two primary strategies for utilizing BERT in sentiment analysis of Vietnamese user comments: (1) feature extraction and (2) fine-tuning.

4.1 Feature Extraction vs. Fine-Tuning

Feature Extraction: In this approach, the BERT architecture is preserved as-is and used solely as a contextual feature generator. The output embeddings from BERT are extracted and then passed to a separate classification model. This strategy treats BERT as a frozen feature extractor without updating its internal weights.

Fine-Tuning: Fine-tuning involves extending the BERT model by adding task-specific layers (e.g., feed-forward layers for classification) and training the entire model, including BERT's internal parameters, on the target dataset. This method has been shown to yield significantly better performance compared to feature extraction, as demonstrated in the original BERT paper [1] and in other benchmark evaluations such as the CoNLL-2003 Named Entity Recognition task [14].

4.2 BERT Pre-trained Model

To fine-tune BERT effectively on Vietnamese data, a suitable pre-trained model is required. In our experiments, we adopt the `BERT-Base Multilingual Cased` model released by Google. This model is trained on 104 languages, including Vietnamese, and is case-sensitive (cased), preserving important features such as letter casing and diacritics — both critical in Vietnamese language processing.

4.3 Fine-Tuning Approaches

We experiment with two different fine-tuning strategies for sentiment classification:

4.3.1 BERT with [CLS] Token Representation ($\text{BERT}_{\text{base}}$)

Following the original setup by Devlin et al. [1], a special token `[CLS]` is added to the beginning of each input sequence. The final hidden state corresponding to this token is assumed to capture the entire sentence's meaning. This representation is passed through a feed-forward neural network for classification. The architecture of this model is illustrated in Figure 1.

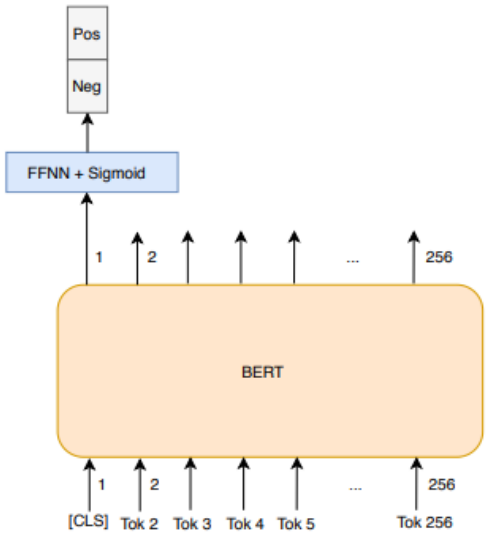

Figure 1: Architecture BERT-base

4.3.2 BERT with Full Token Representations

In this method, we utilize the full sequence output of BERT, including the `[CLS]` token. The resulting output forms a matrix of shape $L \times h$, where L is the maximum sequence length and h is the hidden size of BERT embeddings (typically 768 in $\text{BERT}_{\text{BASE}}$).

This output matrix is then used as input to more complex classification architectures, including:

- **BERT+LSTM**: A bidirectional Long Short-Term Memory (LSTM) network captures sequential dependencies in the token embeddings.
- **BERT+RCNN**: A Recurrent Convolutional Neural Network combines the advantages of both LSTM and CNN to learn both global and local features.

The architecture for these models is depicted in Figure 2.

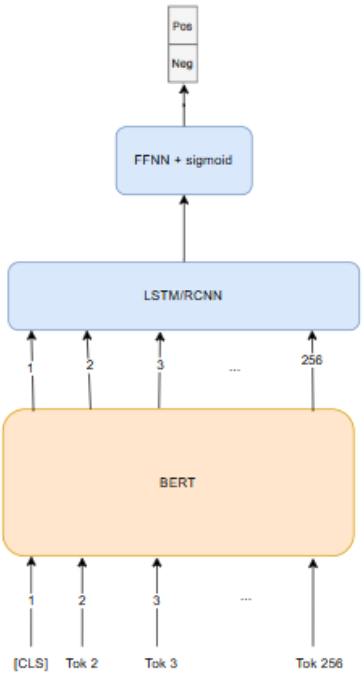

Figure 2: Architecture BERT using all token

5 Experiments

5.1 Dataset

To evaluate the effectiveness of different BERT-based architectures in Vietnamese sentiment analysis, we conducted experiments on a Vietnamese social media comment dataset. The dataset consists of user-generated reviews and comments collected from various online platforms, with each entry annotated as either *positive* or *negative* sentiment.

We preprocessed the dataset by removing special characters, normalizing Unicode, and tokenizing the text using a Vietnamese tokenizer. The dataset was then divided into a training set and a testing set using an 80/20 split. To ensure a balanced evaluation, both sets maintain an approximately equal distribution of positive and negative labels.

All models—BERT, BERT+LSTM, and BERT+RCNN—were trained and evaluated under the same conditions, using the pre-trained BERT-Base Multilingual Cased model as the backbone.

5.2 Results and Discussion

Table 1 presents the performance comparison of the three models using Precision, Recall, and F1 Score as evaluation metrics.

The standard BERT model achieved the highest F1 Score of 90%, indicating its strong capability to capture sentiment information from Vietnamese text. Although BERT+LSTM achieved comparable performance (F1 Score of 89%), it did not significantly outperform the baseline BERT model. This suggests that the LSTM layer may introduce additional complexity without substantial gains in this binary classification task.

On the other hand, the BERT+RCNN model showed a noticeable decline in performance, with an F1 Score of 85%. While the RCNN architecture is designed to capture both local and

sequential features, its integration with BERT may have introduced redundancy or overfitting, especially on limited-size Vietnamese datasets.

Overall, the results demonstrate that fine-tuning BERT with minimal architectural changes provides the most robust and effective approach for Vietnamese sentiment classification. Additional layers such as LSTM or RCNN may not always lead to improved performance and should be used cautiously depending on dataset size and task complexity.

Table 1: Performance Comparison on Sentiment Dataset

Model	Precision	Recall	F1 Score
BERT	88%	91%	**90%**
BERT + LSTM	88%	90%	89%
BERT + RCNN	87%	83%	85%

6 Conclusion

In this paper, we presented a comparative study on fine-tuning BERT for Vietnamese sentiment analysis. We explored three approaches: the standard BERT model utilizing the `[CLS]` token, a hybrid BERT+LSTM model, and a novel BERT+RCNN architecture that combines contextual embeddings with sequential and convolutional features.

Experimental results show that the standard BERT model achieves the highest F1 Score, demonstrating its effectiveness in capturing semantic information from Vietnamese text without additional architectural complexity. While the BERT+LSTM model offered comparable performance, the BERT+RCNN model underperformed, suggesting that adding more layers does not always improve results and may lead to overfitting on small or imbalanced datasets.

Our findings indicate that fine-tuning pre-trained multilingual BERT models is a powerful and efficient approach for Vietnamese sentiment classification. Future work may focus on exploring larger and more diverse datasets, experimenting with domain-specific pre-trained models, or applying cross-lingual transfer techniques to further improve performance in low-resource languages like Vietnamese.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[2] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*. Association for Computational Linguistics, Jul. 2002, pp. 79–86. [Online]. Available: https://aclanthology.org/W02-1011/

[3] B. M. and V. B., "Sentiment analysis using support vector machine based on feature selection and semantic analysis," *International Journal of Computer Applications*, vol. 146, pp. 26–30, 07 2016.

[4] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, "Sentiment analysis of review datasets using naive bayes and k-nn classifier," *arXiv preprint arXiv:1610.09982*, 2016.

[5] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Computational Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011. [Online]. Available: https://aclanthology.org/J11-2001/

[6] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[7] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[9] N. Duyên, N. Xuan Bach, and T. Phuong, "An empirical study on sentiment analysis for vietnamese," *International Conference on Advanced Technologies for Communications*, vol. 2015, pp. 309–314, 02 2015.

[10] Q.-H. Vo, H.-T. Nguyen, B. Le, and M.-L. Nguyen, "Multi-channel lstm-cnn model for vietnamese sentiment analysis," in *2017 9th International Conference on Knowledge and Systems Engineering (KSE)*, 2017, pp. 24–29.

[11] K. Vo, T. Nguyen, D. Pham, M. Nguyen, M. Truong, D. Nguyen, and T. Q. and, "Handling negative mentions on social media channels using deep learning*," *Journal of Information and Telecommunication*, vol. 3, no. 3, pp. 271–293, 2019. [Online]. Available: https://doi.org/10.1080/24751839.2019.1565652

[12] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162/

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147. [Online]. Available: https://aclanthology.org/W03-0419/