Comparing The Performance Of Models For Vietnamese Sentiment Analysis Task

Minh Tuan-Huu Nguyen [1,2,3]**, Thang Bao-Kien Nguyen**[1,2,4]**,**
Quan Van Nguyen[1,2,5] **, Hop Trong Do**[1,2,6]
[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
Email: {[3]21520348, [4]21521432, [5]21521333, [6]hoptd}@gm.uit.edu.vn

Abstract. Vietnamese sentiment analysis is an important and popular task in the field of natural language processing (NLP) and artificial intelligence (AI). In this project, we will select three machine learning and deep learning models, which are SVM, CNN and Transformer, to solve the problem of Vietnamese sentiment classification, and we will use the theory of statistical testing to compare their performances. We have trained these models on three different datasets: UIT-VSFC, UIT-VSMEC, and UIT-ViHSD. For each dataset, we have applied the models on both the original data and a new dataset generated from that original data using the word-segmentation tool py-vncorenlp. Furthermore, for each dataset, we have also used four different feature extraction methods: Count Vectorizer, Tf-Idf, and two PhoBERT architectures, to ensure replication principle in experimental design. After applying statistical analysis theories, we have concluded that the Transformer model shows the best accuracy among the three models. Source code is publicly available for model testing purposes.[1]

1 Introduction

Vietnamese sentiment analysis is an important and popular task in the field of natural language processing (NLP) and artificial intelligence (AI). Nowadays, with the growth of the economy, the need to understand the emotional trends of different types of people in order to maximize profits has become an important and urgent issue. As a result, many new technologies have been developed to support addressing these problems. These technologies help synthesize opinions from large data sources such as social media comments, product reviews, blog posts, and various other sources of information.

The application of sentiment analysis is diverse and has an impact on many different fields. For example, in the domain of business, we can analyze customer opinions to evaluate the quality of products or services. However, sentiment analysis in Vietnamese has posed certain challenges. Vietnamese, being a complex natural language, presents some difficulties. For instance, it contains numerous highly expressive words and relies heavily on context-specific emotional terms.

Currently, there are many machine learning and deep learning models available that can effectively solve the task of sentiment classification in Vietnamese with high

[1] https://github.com/minhquan6203/vietnamese-sentiment-analysis

accuracy. Therefore, selecting the best model for this task is crucial. This requires both practical and theoretical validation factors. Hence, we have decided to focus our group project on comparing the performances of these models. In this project, we will select several machine learning and deep learning models and compare their accuracy results. These models will be applied on different datasets using various data preprocessing methods to ensure replication principle in experimental design. Once we have obtained results regarding the accuracy of the models, we will apply statistical methods to verify the hypotheses we have set based on the experimental outcomes.

2 Related Works

The field of sentiment analysis has attracted significant attention from researchers, and numerous related studies have been conducted to evaluate and compare methods in this field. Below are some notable works in sentiment analysis.

An important research in this field is the study conducted by Pang and Lee (2008) [6], in which they compare and evaluate sentiment analysis methods on articles and product reviews. The methods include traditional machine learning models such as Naive Bayes and Support Vector Machines, along with dictionary-based and rule-based approaches. This work has provided solid foundations and standards for sentiment analysis in various application domains.

Another significant research in the field of sentiment analysis is the study by Socher et al. (2011) [7]. They proposed a model called Recursive Neural Tensor Network, which combines semantic and syntactic parsing to analyze sentence structure and classify sentiment. This work achieved impressive results and provided a new approach to sentiment analysis by integrating semantics and sentence parsing.

Furthermore, the work by Tang et al. (2015) [8] focused on comparing machine learning models and sentiment analysis methods based on social media. They compared the performances of traditional models such as Naive Bayes and SVM with social media-based models like Sentiment Propagation and Social LSTM on Twitter data. This study demonstrated the effectiveness of social media-based methods in sentiment analysis on data from social media platforms.

In conclusion, the field of sentiment analysis has witnessed numerous important researches aimed at accuracy evaluation and comparison.

3 Datasets

3.1 UIT-VSFC - Vietnamese Students' Feedback Corpus for Sentiment Analysis

Student feedback is not only an important resource for improving the quality of education but also plays a crucial role in the study of sentiment analysis in education. Creating a useful educational-related dataset not only helps to develop sentiment analysis in Vietnamese but also supports research about education, particularly in Vietnam.

One significant contribution in this regard is the collection of feedback from Vietnamese students (UIT-VSFC) [9] for reseaching in sentiment analysis and education. The end result was a collection of over 16,000 labeled sentences in terms of sentiment and topic. This dataset is publicly available for research purposes on the NLP@UIT group's website[2].

This dataset not only helps enhance the quality of education but also supports education research in Vietnam. Moreover, the detailed labeling instructions provide benefits for quality control and the development of effective text classification models.

3.2 UIT-VSMEC (version 1.0) - Vietnamese Social Media Emotion Corpus

Paul Ekman proposed six basic emotions of humans, including enjoyment, sadness, anger, surprise, fear, and disgust, which are expressed through facial expressions [1]. Nowadays, emotion recognition has become an important research field, not only based on facial expressions but also on various other types of data.

In recent years, emotion recognition in natural language has gained much popularity in AI community. This is due to its significant potential in various application domains such as marketing, security, psychology, human-computer interaction, artificial intelligence, and more.

Recognizing that immense potential, the UIT@NLP group has developed the UIT-VSMEC dataset [2], which is the first dataset for emotion recognition in Vietnamese social media text. The result is a collection of 6,927 labeled sentences expressing emotions. To ensure the best results with high consistency and accuracy, the group has also constructed a comprehensive and meticulous labeling process guideline for the dataset. The dataset is publicly available for research purposes.

3.3 UIT-ViHSD – Vietnamese Hate Speech Detection Dataset

Vietnam currently has approximately 70 million internet users, with a majority of them being familiar with Facebook and YouTube. Facebook accounts for around 70% of the total internet users in Vietnam, and on average, they spend 2.5 hours per day using this platform. However, on social media, the spread of hatred can easily occur.

According to Kang and Hall [3], inflammatory and hostile language does not always explicitly reveal an individual's hatred. Instead, it is often used as a mechanism to protect their identity and existence from others. However, hatred can lead to destruction, isolate individuals, and weaken society.

On social media platforms, hatred often manifests in the form of hostile comments, hateful content posts, or harassing messages, which quickly spread. The presence of inflammatory and hostile language makes the social media space toxic, threatening users, and causing distress within the community.

Automated detection and classification of hate speech are supervised tasks, and they are crucial in sentiment analysis on social media. Therefore, in 2021, the UIT@NLP group introduced "A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts" [4]. The dataset was manually labeled with over 30,000 sentences, categorized into three labels: "CLEAN" for sentences with positive sentiment, "OFFENSIVE" for sentences with offensive connotations, and "HATE" for sentences with

[2] https://nlp.uit.edu.vn/

hateful implications. This dataset is also publicly available for research purposes on the NLP@UIT group's website[3].

4 Method

4.1 Data Preprocessing

Once we have obtained the necessary datasets, we will proceed to clean them. The operations we will perform in this step include:

- Character removal: We will eliminate any excess characters and whitespace errors from the text, retaining only characters from the alphabet.
- Stopword removal: Stopwords are common and widely used words in natural language. They can be removed without significantly affecting the meaning of the text.
- Data normalization: We will convert all characters in the text to lowercase letters to avoid unnecessary duplication.
- Word segmentation: For each dataset, we will utilize the py-vncorenlp tool for Vietnamese language word-segmentation to create a new dataset consisting of segmented sentences. Word-segmentation is the process of dividing a sentence into individual words, which is an important step in natural language processing. Py-vncorenlp is an open-source toolkit based on the VnCoreNLP library developed by Thanh Vu et al. in 2018 [11] specifically for Vietnamese word-segmentation tasks.

4.2 Text Embedding

For the 6 datasets (including 3 original datasets and 3 preprocessed datasets), we will apply the following 3 text embedding methods to each dataset:

- CountVectorizer: CountVectorizer generates a count vector for each text line, where each element in the count vector corresponds to the number of occurrences of that word in the bag-of-words. This is the simplest embedding method for individual words.
- Tf-idf: This embedding method also transforms each line in the dataset into a vector. However, TfidfVectorizer calculates the importance of words in a text based on their frequency of occurrence across the entire dataset.
- PhoBERT [5]: This is an improvement of BERT performed by the VinAI research team, trained on a 20GB Vietnamese dataset to equip BERT with knowledge of the Vietnamese language. Utilizing the pre-trained model can extract features more effectively.

[3] https://nlp.uit.edu.vn/

4.3 Support Vector Machine

Support Vector Machine (SVM) is a powerful machine learning classification algorithm used in various fields. SVM is used to solve classification and regression problems, depending on its structure. SVM operates by finding a hyperplane in high-dimensional space to separate data points belonging to two different classes. It optimizes a loss function to find the hyperplane with the maximum margin between the two classes. For more details, refer to Figure 1.

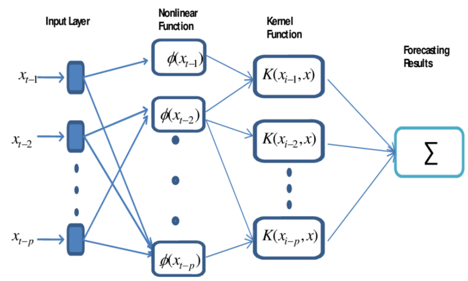

Fig. 1: Support Vector Machine

The process begins with sentences being passed through an embedding module to obtain feature vectors. These vectors are then fed into an SVM for classification. Subsequently, the logits along with the sentence labels are passed through the hinge loss function to calculate the loss.

4.4 Convolutional Neural Network

Convolutional Neural Network (CNN) is a widely used neural network architecture for text classification. It relies on two main components: convolutional layers and pooling layers. The convolutional and pooling layers are stacked together to form a convolutional neural network. Subsequently, fully connected layers and a softmax layer are commonly used for the classification output. For more details, refer to Figure 2.

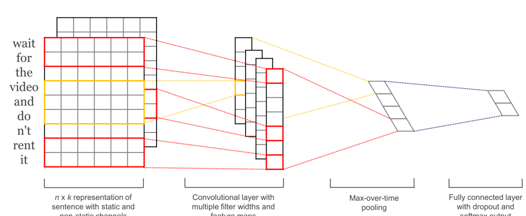

Fig. 2: Module CNN

CNN allows the model to automatically learn features from text data without requiring significant human intervention. In this sentiment analysis task, we have used a transformer to apply multi-head attention and self-attention mechanisms to the feature vectors. These vectors are then passed through the CNN module for classification.

4.5 Transformer

The Transformer architecture was initially introduced in the paper "Attention Is All You Need" by Vaswani et al. [10] in 2017. It replaces the traditional RNN network architecture with a self-attention mechanism to process context-related information within a sentence. See Figure 3 for details.

In this project, similar to SVM and CNN, the sentences are first passed through an embedding module to obtain feature vectors. These feature vectors then utilize the attention mechanism of the Transformer to represent information more effectively and weigh the importance of words, meaning that words with greater significance in the sentence are attended to more. After this step, the vectors are passed through a fully connected layer for classification into labels.

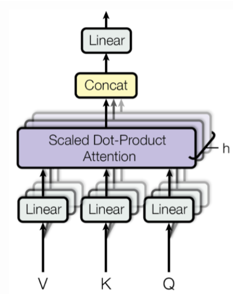

Fig. 3: Module Attention

4.6 Model Comparison

The SVM model has the simplest structure, focusing on finding the best hyperplane to classify the data.

The Transformer model has a significantly more complex structure, notable for its ability to process long sequence data and handle correlations between elements using the attention mechanism. The Transformer model is commonly used in tasks such as machine translation or natural language processing.

It is worth noting that the CNN model has a slight difference compared to the Transformer model. Instead of directly using the input feature vectors, the CNN model utilizes the feature vectors generated by the attention module for the classification process. This allows the CNN model to leverage the knowledge learned from the Transformer model and apply it to the data classification process.

In Figure 5 below, we have provided a detailed description of the models we applied, from the simplest SVM model to the more complex Transformer model and the model combining Transformer and CNN.

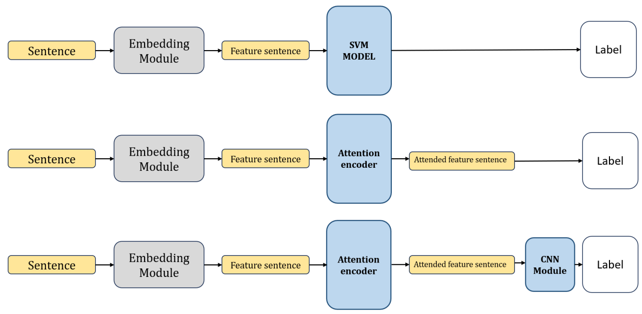

Fig. 4: Structure of the three models: SVM, Transformer and CNN

5 Result Analysis

We applied the three mentioned models to six datasets, including three original datasets and three datasets processed by the py-vncorenlp tokenization tool. For each dataset, we sequentially applied the four text embedding methods mentioned above. For each unit, we ran each model five times and recorded the results of each experiment. In the end, we obtained 120 experimental results for each model. The experiments we conducted ensured the three principles of Replication, Randomization, and Blocking. The results of the experiments have been publicly available at [4].

To compare the accuracy of the three aforementioned models, we formulated the following hypotheses:

- H0: There is no significant difference in term of accuracy among the three models.
- H1: There is a difference in accuracy among the models.

To test this hypothesis, we used the analysis of variance (ANOVA) method. The result in Fig 6 showed a p-value of 0.000238 - smaller than 0.05, indicating that the null hypothesis H0 is not true with a 95% confidence level. This suggests that the difference in accuracy among the three models mentioned above is generated by the difference in the quality of the models, rather than random factors.

Next, we used the Turkey HSD post-hoc analysis to compare the models specifically, more details in Fig 7. The conclusions we obtained are as follows:

[4] https://github.com/minhquan6203/vietnamese-sentiment-analysis/blob/main/data_new.csv

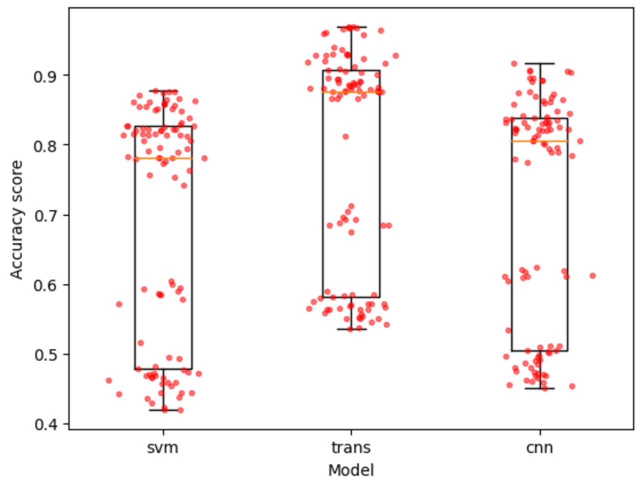

Fig. 5: Boxplot chart showing the accuracy of the three models

```
                      Df Sum Sq Mean Sq F value   Pr(>F)    
as.factor(data$model)   2  0.466 0.23297   8.584 0.000238 ***
Residuals             297  8.060 0.02714                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fig. 6: results of ANOVA showing the differnce of the three models

– The Transformer model has a significantly higher average accuracy than the CNN model by 0.07 with a p-value < 0.05, indicating a statistical significance.
– The Transformer model has a significantly higher average accuracy than the SVM model by 0.09 with a p-value < 0.05, indicating a statistical significance.
– The CNN model has a slightly higher average accuracy than the SVM model by 0.02 with a p-value > 0.05, indicating no statistical significance.

```
               diff         lwr        upr     p adj
svm-cnn   -0.020551 -0.07542951 0.03432751 0.6520248
trans-cnn  0.071409  0.01653049 0.12628751 0.0066980
trans-svm  0.091960  0.03708149 0.14683851 0.0002904
```

Fig. 7: results of Tukey HSD showing the difference of each model

From this, we can conclude that the Transformer model has the highest accuracy compared to the CNN and SVM model.

Furthermore, we applied various preprocessing and text-embedding methods with the hypothesis that the differences between these methods would impact the accuracy of the models used to solve this problem. We examined the impact of these factors on the accuracy of the models and obtained the following results:

– Using ANOVA and Turkey HSD analysis, we found that using different text-embedding methods leads to differences in experimental results. PhoBERT yields the highest accuracy for the models.
– Using a two-sample t-test, we found that the datasets preprocessed by the py-vncorenlp word segmentation tool result in higher accuracy for the models compared to the original datasets.

6 Error Analysis And Future Works

6.1 Error Analysis

Regarding the experimental results, we can only confirm that the Transformer model has the highest accuracy, but we have not yet identified the difference between two remaining models SVM and CNN. Another issue to consider is that there may be many other powerful models that can handle Vietnamese sentiment analysis task but we did not explore them in this project.

The scope of our project is still limited. We only applied the models to three datasets, and the number of experiments we conducted is still pretty small, so the results we obtained may not guarantee absolute objectivity.

6.2 Future Works

In the future, the scale of the experiments in this project can be expanded to provide a more objective evaluation of the models. Some possible ways to increase the scale of the experiments are: adding new datasets, incorporating additional preprocessing methods and different text-embedding techniques, and evaluating the accuracy of other machine learning and deep learning models.

7 Conclusion

In this project, we evaluated and validated the accuracy of different models for Vietnamese sentiment analysis task. The results showed that the Transformer algorithm achieved the highest effectiveness for this task. However, by comparing the accuracy of the models in various contexts, we also observed that the accuracy of a data processing method depends on the quality of the dataset, the preprocessing techniques, and the text-embedding methods used in the task.

The results of this project can serve as a practical basis for selecting a suitable model for Vietnamese sentiment analysis. Individuals and organizations can refer to the proposals of this project to gain a deeper understanding of the opinions and sentiments of Vietnamese users on online and social platforms.

References

[1] Ekman, P., et al.: Basic emotions. Handbook of cognition and emotion **98**(45-60), 16 (1999)

[2] Ho, V.A., Nguyen, D.H.C., Nguyen, D.H., Pham, L.T.V., Nguyen, D.V., Nguyen, K.V., Nguyen, N.L.T.: Emotion recognition for vietnamese social media text. In: Computational Linguistics: 16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019, Hanoi, Vietnam, October 11–13, 2019, Revised Selected Papers 16. pp. 319–333. Springer (2020)

[3] Hoare, D.J., Kowalkowski, V.L., Kang, S., Hall, D.A.: Systematic review and meta-analyses of randomized controlled trials examining tinnitus management. The Laryngoscope **121**(7), 1555–1564 (2011)

[4] Luu, S.T., Nguyen, K.V., Nguyen, N.L.T.: A large-scale dataset for hate speech detection on vietnamese social media texts. In: Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34. pp. 415–426. Springer (2021)

[5] Nguyen, D.Q., Nguyen, A.T.: Phobert: Pre-trained language models for vietnamese. arXiv preprint arXiv:2003.00744 (2020)

[6] Pang, B., Lee, L., et al.: Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval **2**(1–2), 1–135 (2008)

[7] Socher, R., Pennington, J., Huang, E.H., Ng, A.Y., Manning, C.D.: Semi-supervised recursive autoencoders for predicting sentiment distributions. In: Proceedings of the 2011 conference on empirical methods in natural language processing. pp. 151–161 (2011)

[8] Tang, D., Qin, B., Liu, T.: Deep learning for sentiment analysis: successful approaches and future challenges. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **5**(6), 292–303 (2015)

[9] Van Nguyen, K., Nguyen, V.D., Nguyen, P.X., Truong, T.T., Nguyen, N.L.T.: Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In: 2018 10th international conference on knowledge and systems engineering (KSE). pp. 19–24. IEEE (2018)

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

[11] Vu, T., Nguyen, D.Q., Nguyen, D.Q., Dras, M., Johnson, M.: VnCoreNLP: A Vietnamese natural language processing toolkit. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations. pp. 56–60. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-5012, https://aclanthology.org/N18-5012