AERA02 Aptitude Test: Sentiment Analysis on Booking Dataset

Huy M. Le
University of Information Technology
Vietnam National University, Ho Chi Minh, Vietnam
{20521394}@gm.uit.edu.vn

***Abstract*—In today's digital age, travelers rely heavily on online reviews to make informed decisions about their hotel stays. By analyzing the emotions and opinions expressed in reviews and social media, hotels can gain valuable insights into guest satisfaction; besides, customers are able to know which hotel they should pay for. This knowledge is crucial for improving guest experiences, addressing negative feedback promptly, and ultimately, attracting more bookings. However, current research in this field still faces some limitations, such as incomplete preprocessing techniques or not using the best-performing models. This paper focuses on building an intelligent system that can address these limitations. First, we propose an effective pre-processing technique to clean up the collected comments. Second, we apply the above given Booking dataset to current sota models as well as some models that have been used in previous papers. In addition, we also build Experimental results show that Transfer Learning models perform better than Machine Learning and Deep Learning models, with the PhoBERT model achieving the best result with approximately 97 F1-micro.**

***Index Terms*—Sentiment analysis, Transformer, Data Preprocessing.**

I. Introduction

THE rapid explosion of social media has dramatically changed our daily lives. In the past, it was very difficult for users to find suitable accommodation when traveling away from home or on business trips. The solution used by most people, if they did not have acquaintances in the desired location, was to book a hotel through a middleman or a travel agency. However, this still carried a very high risk in terms of satisfaction with the chosen location, the security of the surrounding area, and the convenience of getting around the place of stay.

Nowadays, we don't need to go anywhere far. We just need to have a smartphone with us, download the applications of the booking service providers, choose the hotel we find suitable, and book it. This greatly increases the convenience and autonomy of customers, and in addition, hotels also have an additional channel to spread information. However, to choose a good hotel, it is very important to read the reviews of customers who have used the service at that hotel. Thanks to these reviews, we can avoid unnecessary risks of dissatisfaction. In addition, hotels can also use the reviews to know what they need to improve to attract customers and increase profits.

According to statistics from Shiji[1] organization, from the second quarter of 2021 to the second quarter of 2023, there were more than 9 million comments on hotel booking services with full positive, neutral, and negative tones. With such demand, the Sentiment Analysis (SA) problem in this field is an essential issue. Tourism companies have made research moves related to analyzing the sentiment of comments on their platforms. But for content that lacks context, native cultural expertise, and slow updates for low-information language content, these systems still face significant challenges in processing data from special languages.

Therefore, this study will contribute to the SA problem in Vietnamese. We have used advanced techniques in the field of natural language processing to classify hotel review comments on the Booking.com platform, creating a tool that helps customers and sellers have a more general view of the product they will use. My main research is summarized as follows:

1) We perform rigorous and effective data preprocessing techniques to clean the comments collected from the Booking.com channel (details of the techniques are in Section III). These techniques improve data quality and significantly improve information extraction before training the model
2) We experiment with different methods on the constructed dataset: machine learning, deep learning, transfer learning, and a combination of transfer learning and deep learning. Details of the specific models for each method can be found in Section IV.

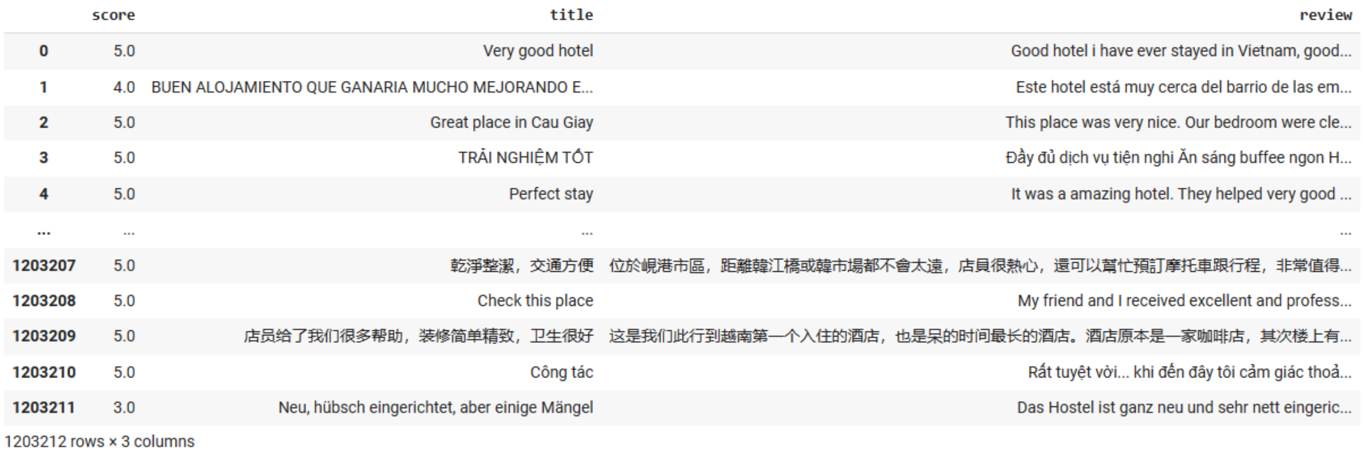
The findings of this study can be used to improve the customer experience in the hotel booking process. The sentiment analysis model can be used to help customers find hotels that meet their needs and preferences. It can also be used by hotels to improve their services and attract more customers.

II. Related Works

A. Introduction to Sentiment Analysis task

Sentiment analysis, also known as opinion mining, is a powerful technique in Natural Language Processing (NLP) that extracts and categorizes opinions, feelings, and attitudes expressed in text data, then identifies whether the sentiment is positive, negative, or neutral. This task recently growing

[1] https://reviewpro.shijigroup.com/



	score	title	review
0	5.0	Very good hotel	Good hotel i have ever stayed in Vietnam, good...
1	4.0	BUEN ALOJAMIENTO QUE GANARIA MUCHO MEJORANDO E...	Este hotel está muy cerca del barrio de las em...
2	5.0	Great place in Cau Giay	This place was very nice. Our bedroom were cle...
3	5.0	TRẢI NGHIỆM TỐT	Đầy đủ dịch vụ tiện nghi Ăn sáng buffee ngon H...
4	5.0	Perfect stay	It was a amazing hotel. They helped very good ...
...	...	...	...
1203207	5.0	乾淨整潔，交通方便	位於峴港市區，距離韓江橋或韓市場都不會太遠，店員很熱心，還可以幫忙預訂摩托車跟行程，非常值得...
1203208	5.0	Check this place	My friend and I received excellent and profess...
1203209	5.0	店员给了我们很多帮助，装修简单精致，卫生很好	这是我们此行到越南第一个入住的酒店，也是呆的时间最长的酒店。酒店原本是一家咖啡店，其次楼上有...
1203210	5.0	Công tác	Rất tuyệt vời... khi đến đây tôi cảm giác thoả...
1203211	3.0	Neu, hübsch eingerichtet, aber einige Mängel	Das Hostel ist ganz neu und sehr nett eingeric...

1203212 rows × 3 columns

Fig. 1: *Raw dataset*

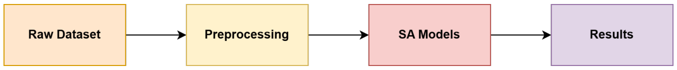


Fig. 2: *General process to solve the SA task*

not just among researchers but also among businesses, governments, and organizations (Sánchez-Rada and Iglesias [1]). Particularly, the task is described below:

Input: Vietnamese comment on booking platforms.

Output: Predict one of three labels: positive, neutral, negative.

The simple process of this task is described in figure 2

B. Existing SA models

One of the most simple approaches is lexicon-based. Lexicons are the collection of tokens where each token is assigned with a predefined score which indicates the neutral, positive and negative nature of the text ([2]). A score is assigned to tokens based on polarity such as +1, 0, -1 for positive, neutral, negative or the score may be assigned based on the intensity of polarity and its values range from [+1, -1] where +1 represents highly positive, and -1 represents highly negative. In Lexicon Based Approach, for a given review or text, the aggregation of scores of each token is performed, i.e., positive, negative, neutral scores are summed separately. In the final stage, overall polarity is assigned to the text based on the highest value of individual scores. Thus, the document is first divided into tokens of single words, where the polarity of each token is calculated and aggregated in the end.

Another research uses Machine Learning approach, in the project of Hajek[3] and Bordes[4], they proposed a NB model along with a SVM model. They used a movie review dataset for training and testing the models. Two thousand reviews were trained after preprocessing and vectorization of the training dataset. Count Vectorizer and TF-IDF were used before training the model. Furthermore, there are some publications that use SVM (Dave et al. [5]) and Logistic Regression (Hamdan et al. [6])

In recent years, with the advent of more powerful and complex models trained on massive datasets, the Neural Network Approach has become a widely used approach. For example, Kitaev and Klein[7] The paper inherits the Transformer architecture of Vaswani [8] to design an Encoder-Decoder based model for the SA task.

III. Dataset

A. General Information

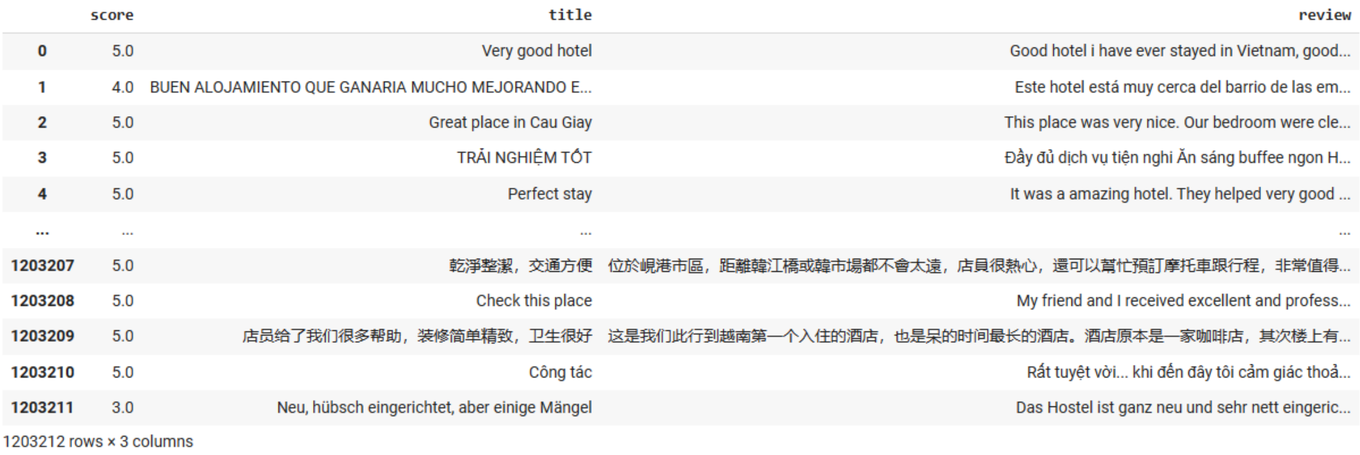
This dataset (show in figure 3) offers information on traveler sentiment towards hotels, containing over 1.2 million customer reviews scraped from Booking.com. Each review is broken down into three key components: a concise 'title' summarizing the guest's experience, a detailed 'review' written in various languages, and a final 'score' reflecting the guest's overall satisfaction on a 5-point scale. By focusing solely on the Vietnamese reviews (and not concerned with other languages) within this dataset, we aim to build a model that deciphers travelers' satisfaction levels based on the content of their reviews. This will allow us to gain valuable insights into guest preferences and optimize the hospitality experience in Vietnam.

B. Dataset Discussion

This dataset offers a rich resource for understanding traveler sentiment, but it also presents several hurdles that need to be cleared before building a robust model. One major challenge is the presence of missing data I. While the dataset boasts over 1.2 million comments, a significant portion lacks crucial information. For instance, not all reviews have a corresponding score (over 160.000 reviews lack a score), title (over 160.000 lack a title), or detailed text (nearly 400.000 reviews are missing the actual review content).

TABLE I: Number of null and non-null values each column in dataset

Column	Number of null values	Number of non-null data
score	163856	1039356
title	163871	1039341
review	389364	813848



	score	title	review
0	5.0	Very good hotel	Good hotel i have ever stayed in Vietnam, good...
1	4.0	BUEN ALOJAMIENTO QUE GANARIA MUCHO MEJORANDO E...	Este hotel está muy cerca del barrio de las em...
2	5.0	Great place in Cau Giay	This place was very nice. Our bedroom were cle...
3	5.0	TRẢI NGHIỆM TỐT	Đầy đủ dịch vụ tiện nghi Ăn sáng buffee ngon H...
4	5.0	Perfect stay	It was a amazing hotel. They helped very good ...
...	...	...	...
1203207	5.0	乾淨整潔，交通方便	位於峴港市區，距離韓江橋或韓市場都不會太遠，店員很熱心，還可以幫忙預訂摩托車跟行程，非常值得...
1203208	5.0	Check this place	My friend and I received excellent and profess...
1203209	5.0	店员给了我们很多帮助，装修简单精致，卫生很好	这是我们此行到越南第一个入住的酒店，也是呆的时间最长的酒店。酒店原本是一家咖啡店，其次楼上有...
1203210	5.0	Công tác	Rất tuyệt vời... khi đến đây tôi cảm giác thoả...
1203211	3.0	Neu, hübsch eingerichtet, aber einige Mängel	Das Hostel ist ganz neu und sehr nett eingeric...

1203212 rows × 3 columns

Fig. 3: *Raw dataset*

Another challenge lies in the data's scope. The dataset appears to contain entries for locations in 'score' column like Busan, Taipei, and Hanoi. This extraneous data needs to be filtered out to ensure the model focuses solely on numerical scores. Additionally, while the focus is on Vietnamese reviews, the titles and reviews themselves are documented in over 40 different languages. This multilingual aspect necessitates a pre-processing step to identify and isolate the Vietnamese reviews for analysis.

Moreover, the amount of 5.0 score through this dataset is quite big (about 700.000 reviews have 5.0 score). This unbalanced data can introduce bias into the model, potentially skewing its understanding of guest experiences.

Finally, the quality of the review text itself presents cleaning challenges. The reviews may contain non-standard characters like icons (e.g., smiley faces, stars) that don't contribute to sentiment analysis. Similarly, extraneous spaces and duplicate characters can introduce noise into the data. Techniques like text normalization and tokenization will be crucial to clean the review text and ensure the model can accurately interpret the sentiment expressed by Vietnamese travelers.

C. Data Preprocessing

Therefore, we proceed to build a data preprocessing pipeline to improve the quality of the datasets, in order to extract valuable features before using them to train classification models. Figure 4 provides an overview of the data preprocessing pipeline consisting of two phases.

1) Phase 1: :

Combine review and title: We realize that both review and title have meaning for the SA problem, in addition, there are some comments that only have reviews and no title and vice versa. So we decided to combine these two fields into a single field called 'content'

Language Detection: Because we only focus on Vietnamese data, however, the data set contains more than 40 different languages, forcing us to use language detection first to filter out other language reviews.

Lowercase Conversion: All characters in the comments in the dataset are converted to lowercase. We do this to avoid Python interpreting two words that are spelled the same but capitalized differently as two different words.

Removal of Extra Whitespace: Social media users often do not know or realize that they type multiple spaces in their comments. Therefore, we decided to remove these extra spaces.

Removal of Links: We believe that website links in comments do not affect the sentiment of the comment. Therefore, we decided to remove all of them.

Unicode Normalization: We also found that many Vietnamese words in the dataset have the same meaning but are detected by Python as different due to different Unicode. The reason is that there are many Unicode Transformation Formats (UTFs) such as UTF-8, UTF-16, UTF-32 that are widely used, but our choice is to normalize to UTF-8.

Removal of Redundant Characters: We remove redundant characters that users intentionally create. Take the example of removing redundant characters: the word ”vuiiii”, which means happy, and the word ”siuuuu” is a Vietnamese teen code that is very in English, after removing redundant characters, they become ”vui” and ”siu”. However, the duplicate characters in the word ”call” will not be removed because it is a meaningful word in English.

After completing the steps in Phase 1, the output of Phase 1 will be fed into Phase 2.

2) Phase 2: :

Word Tokenization: The input sentence is split into meaningful words or phrases. To do this, we used the word segmenter of VnCoreNLP [9] for the PhoBERT model as well as for other models. Because the comments in the dataset are raw text, word segmentation is necessary to prepare the data for training the models. Moreover, PhoBERT uses the VnCoreNLP RDRSegmenter word segmenter to preprocess the training data (including word segmentation and Vietnamese sentences), so it is easier to use the same word segmentation.

De-Teencode: On social media, people often use abbreviations of words to type faster. Some people use these abbreviations to avoid being detected by the system when swearing. Moreover, these abbreviations also have their own names in Vietnamese, called teen code. Therefore, to help

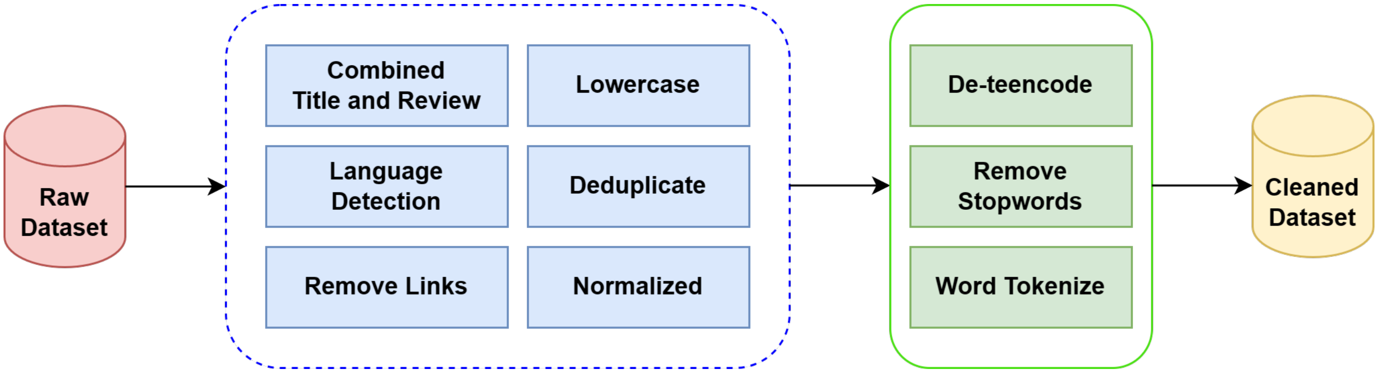


Fig. 4: *Our Preprocessing Process*

TABLE II: Some teencode and their meaning

Teencode	Vietnamese meaning	English meaning
j z	gi vay	what
hay lm	hay lam	good
so ciu	de thuong lam	so cute
siu cap	sieu cap	very

our models better understand the input sentences, we need to map those teen codes to their original words. Moreover, the process of mapping teen codes, we call De–teencode, and Table II below shows some examples of them.

Removal of stopwords: We also remove stopwords from the comments because they are meaningless. In our experiments, we use the Vietnamese stopwords dictionary to remove stopwords from the sentence.

In Phase 2, the data is segmented, de-teencoded, and stopwords are removed. The steps are performed in that order because the output of word segmentation is a list of words, phrases, and individual characters separated by spaces. These characters are then checked to see if they are teen codes and will be de-teencoded in the next step. Therefore, it is a logical decision to perform the de-teencoding step after the word segmentation step. Finally, after the de-teencoding step, we remove all stopwords, and the reason we remove stopwords after the de-teencoding step is that these teen codes can also be stopwords.

D. Completed Vietnamese Booking dataset

After going through 2 preprocessing phases, we have a complete dataset (in figure 5) consisting of only Vietnamese reviews. The dataset consists of 48491 rows with 10 different columns. However, after preprocessing, we only need to pay attention to two columns: 'label' containing the POSITIVE/NEUTRAL/NEGATIVE labels of the review (encoded to 0,1,2) and the 'normalize' column containing The review content has been normalized. Our goal is to predict the label of that review based on the content of these reviews.

IV. Model

A. Machine Learning Approach(ML)

SVM: SVM [10] (Support Vector Machine) is a supervised learning method that can be used for both classification and regression. SVM was first introduced in 1992 by Boser, Guyon, and Vapnik in COLT-92. SVM is a classification and regression model that uses machine learning theory to maximize prediction accuracy and avoid overfitting data. SVM works by finding a hyperplane in a multidimensional space, where the data points are divided into two different classes. SVM became popular when using pixel maps as input.

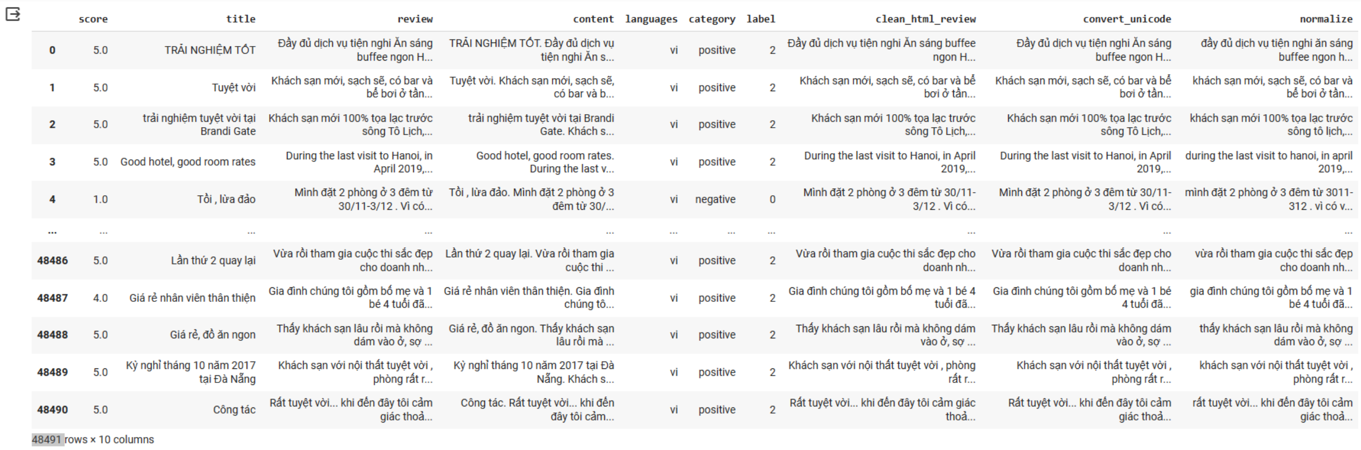
Random Forest: Random forest [11] (RF) is a machine learning method developed by Leo Breiman and Adele Cutler in 2001. While Decision Tree is a well-known model for both classification and supervised learning tasks, it has limitations despite its high accuracy. Random Forest addresses this by combining multiple Decision Trees to improve accuracy and reduce overfitting. Random Forests is an ensemble of trees, where each tree is built on a random subset of the data. After building a collection of trees, Random Forests makes predictions by averaging the predictions of individual trees. Random forests typically provide accurate and stable predictions.

B. Deep Learning Approach(DL)

Text CNN: CNN (Convolutional Neural Network) is a type of artificial neural network that was first proposed by Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner in 1998 [12]. CNNs use layers of convolutional filters that are applied to local features. Originally developed for computer vision, CNN models have since been shown to be effective for natural language processing (NLP) tasks as well. In 2014, the Text CNN model (Convolutional Neural Networks for Sentence Classification) was proposed by Yoon Kim in the paper "Convolutional Neural Networks for Sentence Classification." [13] This model has become an important tool in the field of NLP and has been widely applied to a variety of tasks, including text classification, sentiment analysis, question answering, and many other applications involving language processing.

C. Transfer Learning Approach(TL)

PhoBERT: PhoBERT [14] is known to be the first language model dedicated to Vietnamese, built in 2020. The model has a similar structure and approach to the RoBERTa [15] model and was pre-trained on 20GB of Vietnamese data, about 1GB



	score	title	review	content	languages	category	label	clean_html_review	convert_unicode	normalize
0	5.0	TRẢI NGHIỆM TỐT	Đầy đủ dịch vụ tiện nghi Ăn sáng buffee ngon H...	TRẢI NGHIỆM TỐT. Đầy đủ dịch vụ tiện nghi Ăn s...	vi	positive	2	Đầy đủ dịch vụ tiện nghi Ăn sáng buffee ngon H...	Đầy đủ dịch vụ tiện nghi Ăn sáng buffee ngon H...	đầy đủ dịch vụ tiện nghi ăn sáng buffee ngon h...
1	5.0	Tuyệt vời	Khách sạn mới, sạch sẽ, có bar và bể bơi ở tần...	Tuyệt vời. Khách sạn mới, sạch sẽ, có bar và b...	vi	positive	2	Khách sạn mới, sạch sẽ, có bar và bể bơi ở tần...	Khách sạn mới, sạch sẽ, có bar và bể bơi ở tần...	khách sạn mới, sạch sẽ, có bar và bể bơi ở tần...
2	5.0	trải nghiệm tuyệt vời tại Brandi Gate	Khách sạn mới 100% tọa lạc trước sông Tô Lịch,...	trải nghiệm tuyệt vời tại Brandi Gate. Khách s...	vi	positive	2	Khách sạn mới 100% tọa lạc trước sông Tô Lịch,...	Khách sạn mới 100% tọa lạc trước sông Tô Lịch,...	khách sạn mới 100% tọa lạc trước sông tô lịch,...
3	5.0	Good hotel, good room rates	During the last visit to Hanoi, in April 2019,...	Good hotel, good room rates. During the last v...	vi	positive	2	During the last visit to Hanoi, in April 2019,...	During the last visit to Hanoi, in April 2019,...	during the last visit to hanoi, in april 2019,...
4	1.0	Tồi , lừa đảo	Mình đặt 2 phòng ở 3 đêm từ 30/11-3/12 . Vì có...	Tồi , lừa đảo. Mình đặt 2 phòng ở 3 đêm từ 30/...	vi	negative	0	Mình đặt 2 phòng ở 3 đêm từ 30/11-3/12 . Vì có...	Mình đặt 2 phòng ở 3 đêm từ 30/11-3/12 . Vì có...	mình đặt 2 phòng ở 3 đêm từ 3011-312 . vì có v...
...	...	...	...	...	...	...	...	...	...	...
48486	5.0	Lần thứ 2 quay lại	Vừa rồi tham gia cuộc thi sắc đẹp cho doanh nh...	Lần thứ 2 quay lại. Vừa rồi tham gia cuộc thi ...	vi	positive	2	Vừa rồi tham gia cuộc thi sắc đẹp cho doanh nh...	Vừa rồi tham gia cuộc thi sắc đẹp cho doanh nh...	vừa rồi tham gia cuộc thi sắc đẹp cho doanh nh...
48487	4.0	Giá rẻ nhân viên thân thiện	Gia đình chúng tôi gồm bố mẹ và 1 bé 4 tuổi đã...	Giá rẻ nhân viên thân thiện. Gia đình chúng tô...	vi	positive	2	Gia đình chúng tôi gồm bố mẹ và 1 bé 4 tuổi đã...	Gia đình chúng tôi gồm bố mẹ và 1 bé 4 tuổi đã...	gia đình chúng tôi gồm bố mẹ và 1 bé 4 tuổi đã...
48488	5.0	Giá rẻ, đồ ăn ngon	Thấy khách sạn lâu rồi mà không dám vào ở, sợ ...	Giá rẻ, đồ ăn ngon. Thấy khách sạn lâu rồi mà ...	vi	positive	2	Thấy khách sạn lâu rồi mà không dám vào ở, sợ ...	Thấy khách sạn lâu rồi mà không dám vào ở, sợ ...	thấy khách sạn lâu rồi mà không dám vào ở, sợ ...
48489	5.0	Kỳ nghỉ tháng 10 năm 2017 tại Đà Nẵng	Khách sạn với nội thất tuyệt vời , phòng rất r...	Kỳ nghỉ tháng 10 năm 2017 tại Đà Nẵng. Khách s...	vi	positive	2	Khách sạn với nội thất tuyệt vời , phòng rất r...	Khách sạn với nội thất tuyệt vời , phòng rất r...	khách sạn với nội thất tuyệt vời , phòng rất r...
48490	5.0	Công tác	Rất tuyệt vời... khi đến đây tôi cảm giác thoả...	Công tác. Rất tuyệt vời... khi đến đây tôi cảm...	vi	positive	2	Rất tuyệt vời... khi đến đây tôi cảm giác thoả...	Rất tuyệt vời... khi đến đây tôi cảm giác thoả...	rất tuyệt vời... khi đến đây tôi cảm giác thoả...

48491 rows × 10 columns

Fig. 5: *Vietnames booking dataset*

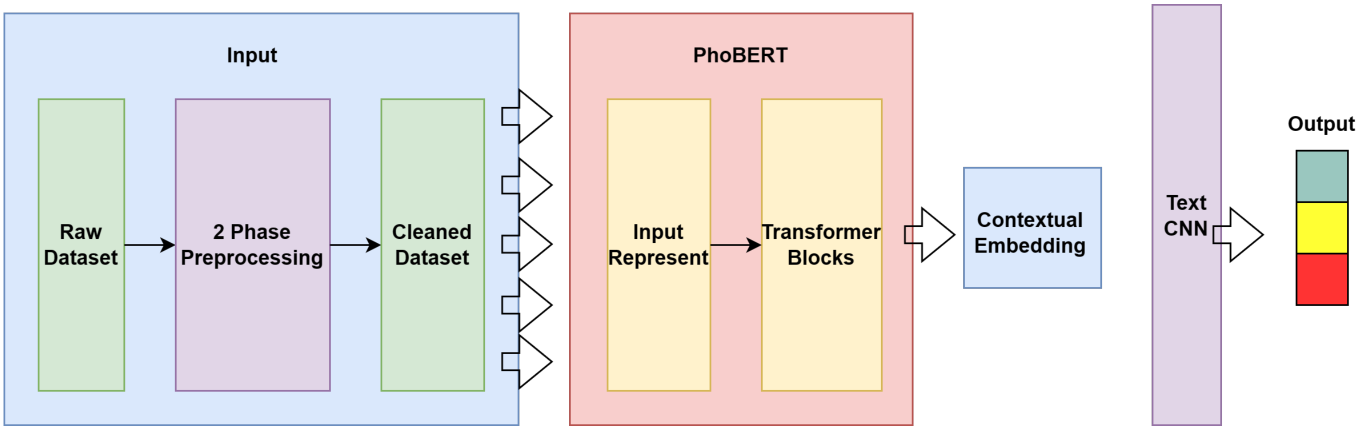


Fig. 6: *General Architecture of PhoBERT-CNN.*

of Vietnamese Wikipedia and the remaining 19GB from the Vietnamese news corpus. Because of its similar structure to BERT [16], PhoBERT also has 2 versions, base and large, with the corresponding number of transformers blocks being 12 and 24, respectively. In addition, with the characteristic of using RDRSegmenter of VNCoreNLP to separate words for the input data before going through the BPE encode layer, the input text tokens are different from other multilingual training models. Currently, PhoBERT is known to be the most modern language model for Vietnamese people to handle natural language processing tasks.

D. *Combine*

PhoBERT + Text CNN: Many BERT-CNN [17] hybrid models have been widely used recently for classifying short texts collected from social networks, specifically for the task of classifying offensive and hateful comments, achieving many promising results. In our study, a PhoBERT-CNN hybrid model was used to evaluate the effectiveness for the task of classifying Vietnamese comments. Thanks to the resonance mechanism of the two models that helps to reduce the error between the predicted and actual labels, PhoBERT-CNN outperforms other models in the Vietnamese comment classification task, especially on our dataset. Figure 6 shows an overview of our approach in this task.

We use PhoBERT to perform word embedding to extract information from data. The PhoBERT model is chosen because it outperforms previous monolingual and multilingual pretrained models, achieving SOTA on Vietnamese natural language processing tasks, including hate speech classification. The architecture of the PhoBERT model is a multi-layer architecture consisting of multiple Bidirectional Transformer encoders. PhoBERT takes as input a sentence consisting of a sequence of words with context. The input representation of the model is constructed by concatenating the tokens together with the segment vector and the corresponding position of the word in the sentence. Figure 7 illustrates the above ideas more intuitively.

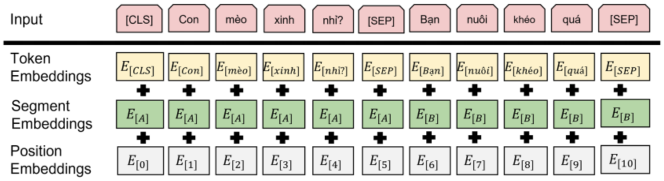


Fig. 7: *Visualize of input of PhoBERT*

The Fully-connected layer at the end of the PhoBERT model is replaced by a CNN architecture. [13] Because CNN is currently the most effective model for short text classification tasks [18], CNN is used instead of other typical deep neural networks such as LSTM, Bi-LSTM, or GRU.

Convolution and Pooling techniques of the CNN network help to extract the main concepts and keywords of the text as features, thereby significantly improving the performance of the classification model. However, CNN networks have a relatively large limitation that they are not suitable for processing text at the string level[13], [18]. To overcome this limitation, a large-scale pretrained language model for Vietnamese such as PhoBERT is a reasonable combination, because PhoBERT has the task of extracting sentence features to input into the Text-CNN model. Then, the word embedding with context of the sentences from PhoBERT is fed into Text-CNN to extract the feature maps. Finally, the predicted label is determined through a Softmax layer.

V. Experiments

A. Setup Experiments

We use the dataset that has gone through 2 Preprocessing Phases as mentioned in the previous section. However, there are some differences between the models at the word embedding step. Specifically, with the machine learning approach, we use the tf-idf vectorize and count vectorize techniques for this step. On the other hand, with the Text CNN model, we use the famous pre-trained word embedding for Vietnamese, which is Phow2v [19]. As for the transfer learning models, they have the ability to self-vectorize sentences to turn them into features.

ML approach: For SVM, Random Forest, and XGBoost models, we used the tf-idf vectorize and count vectorize techniques to normalize the data to fit the model.

With SVM, the gamma parameter determines the influence of one data point on other data points. Here, it is set to 0.0001. The model is trained with a relatively simple classification boundary. The C parameter is set to 1000 to evaluate the impact of misclassified data points on the model's learning process. The kernel used is ”rbf” to classify data more accurately by transforming the data into a higher space.

For Random Forest, we set the number of Decision Trees (n-estimators) to 108 to build the Random Forest model and the maximum depth of the Decision Tree (max-depth) to 40.

DL Approach: For the deep learning approach, we chose Text-CNN with two pre-trained Word Embedding sets: phow2v(syllables), and phow2v(words) with an Embedding size of 100. We experimented with 40 epochs, a batch-size of 256, a sentence length of 100, and a dropout rate of 0.5. Our model uses a 2D Convolution Layer with 32 filters and sizes of 2, 3, and 5, respectively. During training, the Adam optimizer with a learning rate of 1e-4 was applied and the results are shown in Table III.

TL Approach: We conducted transfer learning experiments using the transformer library. We tokenized words using VNCoreNLP, and the parameters used for both PhoBERT base were the Adam optimizer, a learning rate of 0.001, a batch-size of 64, and trained for 10 epochs due to resource constraints. For PhoBERT large, the Adam optimizer was used with a learning rate of 1e-3, a batch-size of 64, and trained for 20 epochs.

Combined: For PhoBERT + CNN, we trained the model for 20 epochs, with the Adam optimizer, a learning rate of 1e-5, epsilon of 1e-8, and a batch-size of 64.

B. Experiments Results

Table III shows the results of experimenting with the models on the dataset. The results are calculated based on micro-F1 score and macro-F1 score.

The table also shows that the RF + TF-IDF model achieved the lowest result with 94.72 on F1-micro and 51.65 on F1-macro, the RF + CountVectorize had a similar result with F1-micro of 94.87 but a higer F1-macro of 52.70, SVM + TF-IDF was the most effective model in this group with 96.49 F1-micro and 70.38 F1-macro.

The DL approach performed better than the ML approach, as all two models in this group performed better than the ML group (except for SVM + TF-IDF). Similarly, the TL approach also performed the best, with PhoBERT achieving the highest results than the other three groups with 96.91 in F1-micro and 77.54 in F1-macro. When combining Text CNN with features extracted from PhoBERT, the result was a disappointing result with 95.52 on F1-micro and 65.47 on F1-macro. This model became the best model in the experiment by topping both metrics.

The fact that pre-trained models performed better than non-pre-trained models suggests that training data is an important issue to consider in this problem in general and in the field of natural language processing in particular.

VI. Conclusion

We have built a dataset for hate speech detection in Vietnamese in the gaming field. Furthermore, we have tested this dataset on models using various approaches. The results show that the models perform quite well on the dataset. However, due to the dataset's size being insufficiently large and some shortcomings in preprocessing and data augmentation, there are still inherent errors. In the future, the team will improve the experimental method by enhancing the preprocessing process and adding more data.

References

[1] J. F. Sánchez-Rada **and** C. A. Iglesias, “Social context in sentiment analysis: Formal definition, overview of current trends and framework for comparison,” *Inf. Fusion*, 2019.

[2] S. Kiritchenko, X. Zhu **and** S. M. Mohammad, “Sentiment Analysis of Short Informal Texts,” *J. Artif. Intell. Res.*, 2014.

[3] P. Hájek, A. Barushka **and** M. Munk, “Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining,” *Neural Comput. Appl.*, 2020.

TABLE III: Result of all models in our experiments

Approach	**Model**	**Embedding**	**F1-micro**	**F1-macro**
Machine Learning	SVM	TF-IDF	96.49	70.38
	SVM	CountVectorize	95.52	68.67
	RF	TF-IDF	94.72	51.65
	RF	CountVectorize	94.87	52.70
Deep Learning	TextCNN	Phow2v (syllables)	96.03	69.32
	TextCNN	Phow2v (words)	96.14	70.07
Transfer Learning	PhoBERT	PhoBERT	**96.91**	**77.54**
Combined	PhoBERT + TextCNN	PhoBERT	95.52	65.47

[4] A. Bordes, X. Glorot, J. Weston **and** Y. Bengio, "A semantic matching energy function for learning with multi-relational data - Application to word-sense disambiguation," *Mach. Learn.*, 2014.

[5] K. Dave, S. Lawrence **and** D. M. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," **in***WWW2003*.

[6] H. Hamdan, P. Bellot **and** F. Béchet, "Lsislif: CRF and Logistic Regression for Opinion Target Extraction and Sentiment Polarity Analysis," **in***SemEval@NAACL-HLT 2015* D. M. Cer, D. Jurgens, P. Nakov **and** T. Zesch, **editors**, 2015.

[7] N. Kitaev **and** D. Klein, "Constituency Parsing with a Self-Attentive Encoder," **in***ACL 2018*.

[8] A. Vaswani, N. Shazeer, N. Parmar **andothers**, "Attention is All you Need," **in***NIPS 2017*.

[9] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras **and** M. Johnson, "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* 2018.

[10] B. E. Boser, I. Guyon **and** V. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," **in***COLT 1992*.

[11] L. Breiman, "Random Forests," *Mach. Learn.*, 2001.

[12] Y. LeCun, L. Bottou, Y. Bengio **and** P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, 1998.

[13] Y. Kim, "Convolutional Neural Networks for Sentence Classification," **in***EMNLP 2014*.

[14] D. Q. Nguyen **and** A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," **in***EMNLP2020* 2020.

[15] Y. Liu, M. Ott, N. Goyal **andothers**, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *CoRR*, 2019.

[16] J. Devlin, M. Chang, K. Lee **and** K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," **in***NAACL HLT 2019*.

[17] A. Safaya, M. Abdullatif **and** D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," **in***SemEval@COLING 2020* 2020.

[18] C. He, S. Chen, S. Huang, J. Zhang **and** X. Song, "Using Convolutional Neural Network with BERT for Intent Determination," **in***IALP 2019*.

[19] A. T. Nguyen, M. H. Dao **and** D. Q. Nguyen, "A Pilot Study of Text-to-SQL Semantic Parsing for Vietnamese," **volume** EMNLP 2020.