Vietnamese social media text classification

1st Vo Quang Nhat Hoang*, 2nd Do Thanh Dat*, 3rd Tran Cong Hien*,
4th Phan Vo My Huyen* 5th Doan Minh Tuan*
*Faculty of Information and Computer Engineering,
University of Information Technology, VNU-HCM, Ho Chi Minh City, Viet Nam
{22520482, 22520206, 22520425, 225205912, 22521600}@gm.uit.edu.vn

***Abstract*—Sentiment analysis plays an important role in various fields such as e-commerce, social media, and education. In this paper, we propose a method that combines machine learning, deep learning, Transformer-based models, and large language models (LLMs) to classify emotions in Vietnamese texts collected from social media. In addition, we integrate the VnEmolex emotion lexicon to enhance the accuracy of deep learning models. Experiments are conducted on three large datasets: UIT-VSMEC, UIT-VSFC, and UIT-ViCTSD. The results show that incorporating the emotion lexicon and applying Vietnamese-specific language preprocessing techniques significantly improves model accuracy, especially when using Transformer-based models such as PhoBERT and Few-shot Prompting with LLMs.The experimental results demonstrate that the combination of modern models and knowledge from the emotion lexicon significantly enhances emotion classification performance in real-world scenarios.**

***Index Terms*—sentiment analysis, emotion lexicons, text classification, machine learning, deep learning, transformers models, LLM**

I. Introduction

The topic of Sentiment Analysis (SA) has garnered significant interest and research in academia, particularly in the development of predictive models. SA has numerous applications in daily life as it serves as a tool for monitoring opinions from user-generated data and supporting decision-making [1]. Sentiment analysis is applied across various fields, including e-commerce, social media, blogs, discussion forums, and education.

To address these challenges, we'll use three datasets—UIT-VSMEC [2], UIT-VSFC [3], and UIT-ViCTSD [4] — to investigate the effectiveness of the proposed methods mentioned in the abstract. All three are large-scale datasets manually annotated by humans through a rigorous labeling process in specific domains like social media (UIT-VSMEC), students and education (UIT-VSFC), and hate speech detection (UIT-ViCTSD). Additionally, VnEmolex [5] is a sentiment dictionary comprising eight different emotion categories and containing 12,795 emotion-laden words.

Our paper is structured as follows. Section II presents an overview of current research on Vietnamese sentiment analysis. Section III briefly introduces the datasets used in our experiments. Section IV describes the proposed methods involving machine learning, deep learning, and large language models, and the integration of sentiment lexicon features with deep learning models. Section V presents the experimental results and error analysis of the proposed methods. Finally, Section VI concludes the study and suggests future research directions.

II. Related Works

Sentiment analysis is a crucial task in the field of natural language processing (NLP), with many real-world applications such as monitoring public opinion, product evaluation, education, and detecting negative speech. In Vietnamese language research, this task has also attracted considerable attention in recent years.

Traditional approaches are typically categorized into three main directions: (1) lexicon-based methods, (2) classical machine learning methods, and (3) deep learning or pretrained language models.

Among these, lexicon-based methods aim to assign "sentiment scores" to individual words without the need for training. A prominent example is the VnEmolex lexicon , which contains over 12,000 Vietnamese words associated with 8 basic emotions, divided into positive and negative polarities. Although simple, this approach faces challenges in handling complex contexts and ambiguous expressions commonly found in online texts.

Machine learning methods such as Naive Bayes, SVM, and Logistic Regression use statistical features (e.g., TF-IDF, BoW) combined with labeled data to train classifiers. In previous studies, the UIT-VSMEC dataset was used for sentiment classification on social media, while UIT-VSFC was employed to assess emotions in student feedback within educational settings. More recently, the UIT-ViCTSD dataset was developed to detect toxic and non-constructive comments on social platforms, requiring binary classification between constructive and harmful language.

Deep learning methods and modern language models (such as BERT, PhoBERT, GPT) are gradually replacing traditional approaches due to their superior ability to learn contextual semantic representations. The PhoBERT model, trained on 20GB of Vietnamese text, has demonstrated outstanding performance in sentiment analysis tasks by effectively capturing Vietnamese-specific linguistic structures.

Some recent studies have proposed integrating emotion lexicons into deep learning models to enhance semantic information. For instance, in the authors showed that incorporating VnEmolex into models such as Text-CNN, BiGRU, or BERT significantly improves sentiment classification accuracy. Moreover, large language models (LLMs) such as GPT have opened up opportunities for zero-shot and few-shot prompting, which are particularly useful in low-resource scenarios with limited labeled data.

In this study, we adopt the combined approach of deep learning and knowledge from emotion lexicons, and further extend it with Transformer-based models and LLMs to handle a wide range of emotional scenarios in Vietnamese.

III. VIETNAMESE SENTIMENT ANALYSIS DATASETS

UIT-VSMEC was created for the purpose of detecting emotions in Vietnamese social media texts. This dataset includes seven different emotion levels, as described in Table I. We use this dataset as a benchmark to evaluate the effectiveness of our proposed method. Besides UIT-VSMEC, we also analyze the results on two other benchmark datasets, UIT-VSFC and UIT-ViCTSD, to emphasize the effectiveness of our method. UIT-VSFC was constructed to analyze student feedback on educational activities. This dataset includes two tasks: a sentiment analysis task to detect user emotions from texts related to educational activities, and a topic classification task to identify categories within teaching and learning activities such as lecturers, facilities, and curricula. In this paper, we utilize the sentiment analysis task for our experiments. The labels for UIT-VSFC are presented in Table I. Finally, UIT-ViCTSD is a dataset built for the task of detecting hate speech in Vietnamese. This dataset also features three labels, as shown in Table I. All three datasets were manually annotated by humans following a detailed and rigorous annotation process.

The UIT-VSMEC dataset contains a high proportion of emojis and emotive characters, accounting for approximately 32.57%, clearly reflecting the informal and emotionally rich language of social media. As this dataset was developed for the purpose of emotion classification, it includes many short sentences, numerous emojis, as well as abbreviations and spelling errors, which pose challenges in preprocessing and annotation. In contrast, the UIT-VSFC dataset is more formal, with fewer emojis, though teencode and abbreviations still appear. Lastly, the UIT-ViCTSD dataset focuses on detecting hostile, aggressive, or non-constructive language; its texts are typically short, containing slang, strong emotional expressions, and vulgar words.

IV. METHODOLOGY

The task of sentiment analysis is categorized as the text classification task. Figure 2 illustrates briefly our methodology,

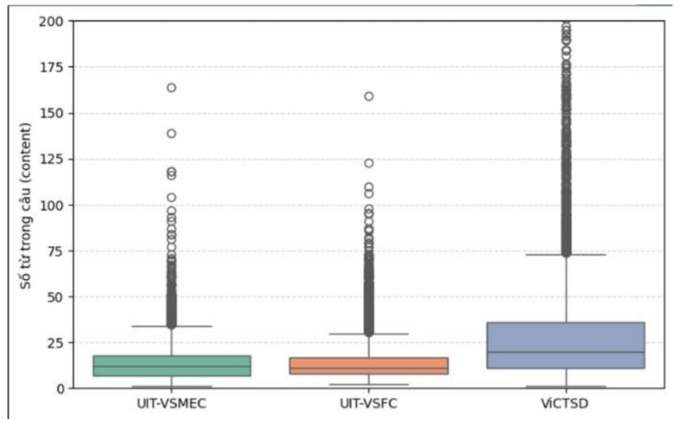

Figure 1: Distribution of the length of the comments in the three Vietnamese Sentiment Analysis datasets.

Table I: Overview Statistics of the Three Vietnamese Sentiment Analysis Datasets.

Dataset	Size	Labels
UIT-VSMEC	6,927	FEAR SURPRISE ANGER ENJOYMENT SADNESS DISGUST OTHER
UIT-VSFC	16,175	POSITIVE NEGATIVE NEUTRAL
UIT-ViCTSD	33,400	CLEAN OFFENSIVE HATE

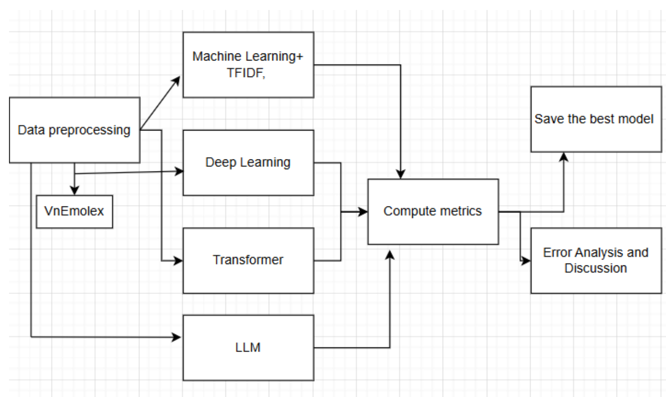

Figure 2: Experimental procedure.

including pre-processing techniques, combining the VnEmolex to the feature vectors, and fitting them to the classification models.

A. Data pre-processing

Vietnamese text data collected from social media or user feedback often contains many non-standard characteristics such as emojis, abbreviations (teencode), misspellings, and the use of words that do not follow standard orthographic rules. These factors directly impact the effectiveness of sentiment analysis models if not properly handled.

In this paper, we apply a five-step preprocessing pipeline as follows:

1) **Normalize data to string format**
All input data is converted into a plain character string format to ensure processing consistency. This is especially useful when the original data is stored in various forms, such as lists of words, word-label pairs, or JSON format.
2) **Normalize Emojis and Emoticons**
Statistics show that over 30% of the text in the UIT-VSMEC dataset contains emojis or emoticons, which serve as clear indicators of emotion. Therefore, we normalize character-based symbols (e.g., :), =))) into their corresponding emojis (🙂, 😂) to increase the accuracy of sentiment analysis.
3) **Normalize Abbreviations and Teencode**
Social media texts frequently feature abbreviations (e.g., "mik" for "mình," "vk" for "vợ"), which degrade the model's vocabulary recognition capabilities. We use a list of common abbreviations and convert them to their standard forms through a lookup table of teencode mappings compiled from the community.
4) **Normalize Spelling Errors**
Common typing errors like "qúa" instead of "quá" (too much/over), or "đẹppp" instead of "đẹp" (beautiful) are handled using a list of Vietnamese spelling errors containing over 700 patterns developed from **Vietnamese Natural Language Processing Community**. Incorrect words are replaced with their corresponding standard forms to ensure consistency.

Table II: Some example sentences after pre-processing

Sentence	After pre-processing
vk tui h mới đi lm về á :)))	vợ tôi vừa mới đi làm về à 🙂
có j ko hiểu thì ib mik nhé	có gì không hiểu thì inbox mình nhé
cực kỳ qúa đáng lun á	cực kỳ quá đáng luôn à

B. *Emotion Lexicons*

To enhance the emotional semantic information for our deep learning models, we integrate the Vietnamese sentiment lexicon – VnEmolex – into the training process. Supplementing the models with knowledge from the lexicon helps them better exploit emotional signals that are often ambiguous in social media texts.The VnEmoLex is used to represent raw texts from three datasets: UIT-VSMEC, UIT-VSFC, and ViHSD to vectors as illustrated in Table III.

1) **VnEmolex – Vietnamese Sentiment Lexicon**
VnEmolex is a lexicon developed by combining EmoLex (English) and VietWordNet, comprising 13,000 Vietnamese words annotated with emotional labels. Each word in VnEmolex is assigned one or more labels belonging to 8 basic emotion categories and 2 opposite state
   - **Emotion categories**: Enjoyment, Trust, Anticipation, Surprise, Sadness, Fear, Anger, Disgust
   - **Opposite state:** Positive, Negative

   Additionally, VnEmolex clearly categorizes words as either positive or negative sentiment, facilitating their quantitative integration into deep learning models.
2) **Technique for Integrating VnEmolex into Text**
We use a method of counting the number of emotional words that appear in each input text. This forms a feature vector whose dimension equals the number of emotions in the lexicon (e.g., a vector of 8 values representing the count of words carrying emotions like "Enjoyment," "Fear," etc.).
Example: For the sentence "Tao ghét tụi mày, tao muốn đập hết" (I hate you guys, I want to smash everything) → the VnEmolex vector might be: [Enjoyment: 0, Sadness: 0, Anger: 2, Fear: 0, Disgust: 1, ...]
Subsequently, this emotion vector is concatenated with the feature vector from a deep learning model, such as Text-CNN, Bi-GRU, or Transformer models, to serve as input for the classification layer.
3) **Adjusting Labels for Each Dataset**
Since each dataset has different label definitions, we remap the emotion categories as follows:
   - **UIT-VSMEC:** We use all 7 labels that directly match VnEmolex.
   - **UIT-VSFC:** We group labels into two polarities: Positive and Negative. Emotions not belonging to these two groups are combined into Neutral.
   - **UIT-ViCTSD:** As this is a binary classification task, we map all negative, abusive, or hateful texts to the "Toxic" label, while texts containing constructive content are mapped to the "Constructive" label.
4) **The goal**
Integrating VnEmolex isn't just to add emotional information to the model. It also helps deep learning models better understand the subtle emotional nuances that are often confused when relying solely on contextual representations, thereby increasing classification accuracy.

C. *Classification models*

In this paper, we use four kinds of methodologies to evaluate three different benchmark datasets, including UIT-VSMEC, UIT-VSFC, and UIT-ViCTSD. The methodologies consist of two machine learning models (SVM and Logistic Regression), two deep learning models (Text-CNN and BiGRU), a Transformer-based model(PhoBert), and a Large Language Model (Gemini-2.0-flash), which are described as follows.

1) **Machine Learning**
In this paper, we initially evaluated SVM and Logistic Regression algorithms for emotion reecognition - UIT-

Table III: Describe some example of the sentences in the UIT-VSMEC dataset that were obtained from combining it with VnEmoLex [1].

Sentence	Disgust	Fear	Enjoyment	Sadness	Suprise	Anger
cho **đáng đời con quỷ** . về nhà **lôi** con nhà mày ra mà **đánh** . ***(English:** go home and fight with your son because you deserve the devil.)*	2	1	0	0	0	2
chả **mong** gì nhiều chỉ **mong** về **già được như** hai ông bà! ***(English:** expecting nothing more than to be like my grandparents!)*	0	0	2	0	1	1
làm công nhân đã bị **vắt kiệt sức khỏe** cho đến khi **bị thải.** ***(English:** draining workers' health till they are cut loose.)*	1	1	0	2	1	2

Table IV: EVALUATION RESULT 5 BASE MODEL IN UIT-VSMEC CORPUS(%)

Method	Accuracy	Macro F1-score
SVM	**53.82**	**51.03**
Logistic Regression	**53.39**	**46.79**
Decision Tree	39.11	37
Naive Bayes	46.32	27.44
Random Forest	48.92	43.68

VSMEC dataset. We also tested three more machine learning algorithms including Decision Tree, kNN, and Naive Bayes by training them on this dataset. Consequently, Logistic Regression achieved the second-best result after SVM, which is displayed in Table IV. This is the main reason why we chose SVM and Logistic Regression for subsequent experiments on the remaining two datasets.

Logistic Regression: Logistic Regression is a model used to estimate the probability of an event occurring (e.g., the probability of belonging to a specific class) and overcomes the shortcomings of linear regression in modeling probabilities, particularly regarding predictions outside the [0,1] range.

Support Vector Machine (SVM): SVM is a robust classification algorithm that focuses on finding the most optimal decision boundary, one that maximizes the distance to the closest data points, thereby improving its generalization ability on new data.

2) **Deep Learning**

Convolutional Neural Network (Text-CNN): A type of deep learning model that utilizes layers with convolving filters applied to local features, while also emphasizing their successful adaptation from computer vision to various NLP tasks, particularly when leveraged with pre-trained word embeddings.

BiGRU: GRU is a more advanced kind of Bi-LSTM, often referred to as update gate and reset gate, which was utilized to alleviate the gradient loss issue that traditional RNNs encountered. The two vectors essentially determine what data should be sent to the output. The unique feature is that it can be taught to retain old data without erasing output prediction-related data.

3) **Transformer**

PhoBERT: A pre-trained model based on the RoBERTa architecture, optimized for Vietnamese, and trained on a 20GB text dataset. PhoBERT utilizes the VnCoreNLP word segmenter to better handle compound words and punctuation compared to multilingual models like mBERT.

4) **Large Language Model**

Gemini-2.0-flash: Delivers next-gen features and improved capabilities, including superior speed, built-in tool use, multimodal generation, and a 1M token context window.

V. EXPERIMENTS AND RESULTS

A. Model settings

1) **Machine Learning**

The two best models will be trained on TF-IDF and BoW (Bag of Words) features. To further enhance their performance, we'll employ Grid Search for hyperparameter tuning

2) **Deep Learning**

Text-CNN: We set up four conv2D layers with 32 filters at sizes 1, 2, 3, 5 and used softmax for activation. In addition, we set batch size equal to 32, max sequence length is 100, and dropout is 0.2 for this model. The model is compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric. For datasets with many labels, particularly like UIT-VSMEC, we employ Focal Loss to address class imbalance and improve the model's focus on hard-to-classify examples.

BiGRU: We set up the bidirectional layer followed by a max-pooling 1D, a dense layer has 50 in size for both activation and the softmax activation. We set batch size equal to 32, max sequence length is 80, and dropout is 0.2. The model is compiled using the Adam optimizer, categorical cross-entropy loss, and accuracy as the evaluation metric.

Both models will be integrated with VnEmolex to evaluate the effectiveness of enhancing emotional knowledge.

3) **Transformer**

PhoBERT: We set the max sequence length is 128, batch size equals 16, the learning rate is at 2e-5, accumulation steps are 5, and the number of train epochs is 10.

4) **Large Language Model**

To leverage the reasoning power of Large Language Models (LLMs) like Gemini, we employ two main strategies:

- **Zero-shot Prompting:** In this strategy, the model predicts sentiment based on the task description provided in the prompt, without requiring any fine-tuning or re-training.
- **Few-shot Prompting:** This strategy enhances accuracy by including a few illustrative examples (samples) directly within the prompt, helping the model better understand the desired format and response style for the specific task.

B. Experimental Results

When combining the two features TFIDF and BoW, the machine learning models showed significant improvements across all three datasets in table V. Meanwhile, integrating VnEmoLex helped enhance the accuracy of deep learning models on all three datasets, with a particularly notable improvement on the UIT-ViCTSD dataset in table VI. PhoBERT delivered outstanding results compared to basic deep learning models in table VII, especially on the UIT-VSMEC dataset, where it showed a remarkable improvement of approximately 2.07% in accuracy over the best results from previous research [2]. In addition, the few-shot prompting technique outperformed the zero-shot approach across all datasets in table VIII. Notably, for the VSFC dataset, the model achieved results comparable to or even better than PhoBERT.

Table V: Comparison of Machine Learning on 3 datasets(%)

Dataset	Algorithm	Accuracy	Macro F1-Score
UIT-VSMEC	SVM + TFIDF	55.56	52.36
	SVM + BoW	54.4	52.08
	Logistic Regression + TFIDF	**55.99**	**53.85**
	Logistic Regression + Bow	53.68	50.32
UIT-VSFC	SVM + TFIDF	**88.76**	68.87
	SVM + BoW	88.03	**72.69**
	Logistic Regression + TFIDF	88.38	70.70
	Logistic Regression + Bow	88.28	71.57
UIT-ViCTSD	SVM + TFIDF	89.20	66.36
	SVM + BoW	85.90	64.97
	Logistic Regression + TFIDF	**89.40**	**65.93**
	Logistic Regression + Bow	87	65.70

Table VI: Comparison of Deep Learning on 3 datasets(%)

Dataset	Algorithm	Accuracy	Macro F1-Score
UIT-VSMEC	Text-CNN	55.56	53.53
	Text-CNN + VnEmolex	**58**	**557**
	BiGRU	52.96	51.23
	BiGRU + VnEmolex	55.70	52.45
UIT-VSFC	Text-CNN	88.60	75.36
	Text-CNN + VnEmolex	90.02	76.58
	BiGRU	90.21	76.17
	BiGRU + VnEmolex	**90.24**	**77.21**
UIT-ViCTSD	Text-CNN	88.10	64.90
	Text-CNN + VnEmolex	89.40	**69.46**
	BiGRU	87.50	66.61
	BiGRU + VnEmolex	**88.90**	65.60

Table VII: Comparison of PhoBert on 3 datasets(%)

Dataset	Accuracy	Macro F1-Score
UIT-VSMEC	61.81	58.90
UIT-VSFC	93.40	82.62
UIT-ViCTSD	90.70	73.33

Table VIII: Comparison of LLM on 3 datasets(%)

Dataset	Algorithm	Accuracy	Macro F1-Score
UIT-VSMEC	Zero-shot	58.40	54.87
	Few-shot	**59.20**	**55.83**
UIT-VSFC	Zero-shot	83.33	70.22
	Few-shot	**93**	**82.01**
UIT-ViCTSD	Zero-shot	82.67	65.38
	Few-shot	**83.47**	**68.09**

C. Error analysis and Limitations

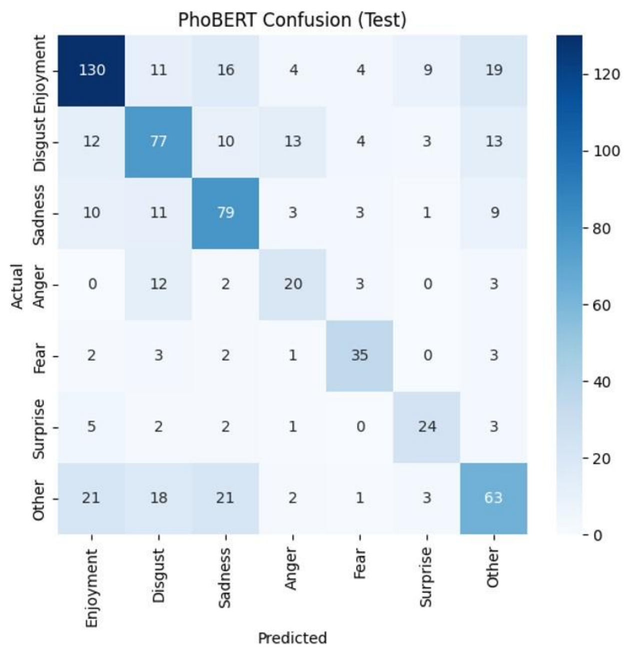

Figure 3: Confusion matrix of PhoBert on the VSMEC dataset

1) **Error analysis**

While models like PhoBERT and LLMs achieve high performance on most datasets, some classification errors still exist due to the unique characteristics of the Vietnamese language and the informal expressions used by social media users. We use confusion matrix and specific example analysis to clarify these cases.

In the UIT-VSMEC dataset, labels like "Anger" and "Disgust," or "Sadness" and "Fear," are often confused by models due to the similarity in how these negative emotions are expressed. Additionally, social media language frequently employs emojis or sarcasm, making it challenging for models to determine true sentiment. Meanwhile, some sentences contain multiple emotional nuances, making single-label assignment difficult, or are too short/fragmented for the model to accurately identify

Table IX: Some Error Cases

Comments	Emotion	Predictions	Explanations
chia sẻ cho ai thích thể hiện nè (**English:** share with those who enjoy showing)	Other	Enjoyment	The use of the word “thích” (like) confuses the meaning of the sentence and gives the impression of enjoyment.
tao đâu biết nó kinh khủng thế này :((((**English:** I had no idea it was so terrible :(( )	Fear	Sadness	The “:((“ emotion has a stronger influence than the word “kinh khủng” (terrible).
đây gọi là nghiệp vụ (**English:** this is known as karma)	Enjoyment	Other	The user sometimes has difficulty identifying the enjoyable nuance of the sentence.
người ta có bạn bè nhìn vui ghê (**English:** they look so happily having friends)	Sadness	Enjoyment	Our model cannot understand the context of this comment.
tao không sợ đi làm mệt mỏi tao chỉ sợ không có niềm vui với nó (**English:** I’m worried of not having fun with it, not of going to work tired)	Fear	Sadness	Both fear and sadness are present in this sentence, but since both “sợ” (fear) and “không sợ” (not afraid) are present, our model chooses sadness.
tao đéo ngờ tới trường hợp này :)) (**English:** I f*cking cannot know it)	Surprise	Enjoyment	The term “đéo ngờ tới” (did not expect) has not appeared in the train data so it cannot be recognized to express surprise.

the correct emotion. Table IX contains a few illustrations of prediction errors along with their corresponding explanations.

We used the confusion matrix to evaluate how well the models performed by better visualizing the prediction and actuality of the labels. The confusion matrix of the PhoBert - the best classification model on the test set of the UIT-VSMEC is shown in Figure 3. As can be seen, the best classification model on the Enjoyment label has more than 72.22% correct prediction points, followed by the Fear label’s accuracy of 70% and the Surprise label’s accuracy of more than 60%. Label Anger has the lowest accuracy at 45.45%.

2) **Limitations**

When using the Gemini 2.0 Flash model for text classification, we encountered limitations regarding the number of tokens allowed per API query. As the volume of text increased or prompts became longer (especially during the few-shot phase), the model sometimes failed to respond in time or was interrupted. Although I implemented a waiting and retry mechanism, some queries still didn’t return results after the maximum allowed time (70 seconds). This led to missing prediction values for some samples, affecting the completeness of the results and making it difficult to assess the model’s actual performance.

Models still struggle with sentences containing sarcasm, mixed emotions, or ambiguous emojis.

VnEmolex does not yet fully cover slang, new words, or modern variations found on social media, leading to the omission of subtle emotions.

VI. CONCLUSION AND FUTURE WORKS

A. Conclusion

In this study, we proposed and experimented with a system for Vietnamese emotion classification from social media text, combining various techniques ranging from traditional machine learning to deep learning models, Transformer models (PhoBERT), and Large Language Models (LLMs). Concurrently, the integration of the Vietnamese sentiment lexicon VnEmolex into deep learning models clearly demonstrated significant effectiveness in improving classification accuracy.

Experiments were conducted on three manually annotated datasets: UIT-VSMEC, UIT-VSFC, and UIT-ViCTSD, representing three distinct semantic domains: social media emotions, educational feedback, and toxic speech. The results show:

- PhoBERT yielded the highest effectiveness among traditionally trained models, particularly with superior accuracy on VSFC and ViCTSD.
- LLMs (Gemini-2.0-flash) with the few-shot prompting method achieved results comparable to or even better than Transformer models without retraining, demonstrating strong potential in Vietnamese NLP tasks.
- VnEmolex played a crucial role when integrated into Text-CNN, Bi-GRU, or PhoBERT models, helping to add an extra dimension of emotional information to the input data.

B. Future works

- Experiment with and fine-tune new LLMs like GPT and LLaMA on Vietnamese data to improve their effectiveness in specific tasks.
- Expand and update VnEmolex, including new words, "teencode" (Vietnamese slang/textspeak), and popular emojis from current social media platforms.
- Combine semantic and syntactic information by using data from dependency parsing or POS tagging to enrich the structural context for deep learning models.

With this future works, we believe that the Vietnamese emotion analysis system can be further enhanced in terms of both accuracy and practical applicability in digital platforms and human-computer interaction systems.

REFERENCES

[1] A. L. Doan and S. T. Luu, “Improving sentiment analysis by emotion lexicon approach on Vietnamese texts,” in *Proc. 2022 Int. Conf. Asian Lang. Process. (IALP)*, 2022.

[2] V. A. Ho, D. H.-C. Nguyen, D. H. Nguyen, L. Thi-Van Pham, D.-V. Nguyen, K. Van Nguyen, and N. L.-T. Nguyen, “Emotion recognition for vietnamese social media text,” in Computational Linguistics - 16th International Conference of the Pacific Association for Computational Linguistics, 2019.

[3] K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L.-T. Nguyen, “Uit-vsfc: Vietnamese students’ feedback corpus for sentiment analysis,” in 2018 10th International Conference on Knowledge and Systems Engineering (KSE), 2018.

[4] Nguyen, Luan Thanh and Van Nguyen, Kiet and Nguyen, Ngan Luu-Thuy, "Constructive and toxic speech detection for open-domain social media comments in Vietnamese"", in *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices*, Cham: Springer International Publishing, 2021

[5] KTLab, “Vnemolex: A vietnamese emotion lexicon for sentiment intensity analysis,” in OpenAIRE, 2017.

[6] D. Q. Nguyen and A. T. Nguyen, “PhoBERT: Pre-trained language models for Vietnamese,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.