A Hybrid Model of RoBERTa and Bidirectional GRU for Enhanced Sentiment Analysis

Nguyen Thi Mai Trang

Abstract— **Sentiment analysis, a natural language processing task, has gained significant attention because of its diverse applications in understanding user opinions and emotions in social media, customer feedback, and online reviews. Several studies have been conducted on this task using English datasets, yielding noteworthy outcomes. However, it is important to note that research on the same task for the Vietnamese language remains limited, and the available training data are currently not substantial. The proposed hybrid model is based on the power of two architectures: RoBERTa, a transformer-based model, and Bidirectional Gated Recurrent Units (Bi-GRU). By fusing the strengths of both models, the approach aims to enhance sentiment analysis performance and generalize better on diverse datasets. The research outlines the model's architecture, highlighting the seamless integration of RoBERTa and Bi-GRU components, and describes the fine-tuning process on a large corpus of Vietnamese texts for sentiment analysis. To evaluate the model's performance, comprehensive experiments are conducted on the benchmark datasets, comparing it with state-of-the-art approaches for sentiment analysis. The experimental results unequivocally demonstrate that the proposed model outperforms other methods across various metrics, including accuracy, precision, recall, and F1-score on both the IMDb and UIT-VSFC datasets.**

Keywords—**sentiment analysis, RoBERTa, Bidirectional GRU, natural language processing, Vietnamese text.**

I. INTRODUCTION

In recent years, sentiment analysis has developed significantly due to the widespread use of social media platforms, where people often express their opinions and emotions. Sentiment analysis is a task that determines the polarity of specific text content, such as positive, negative, or neutral sentiments. By predicting the public's opinions on a particular topic, sentiment analysis serves as an indicator of public preferences in many fields, such as business, education, social issues, and politics. For instance, businesses can leverage sentiment analysis to understand their customers' likes and dislikes towards their products or services and tailor their marketing strategies accordingly,

potentially leading to business growth.

Sentiment analysis poses unique challenges in the context of the Vietnamese language due to its complex linguistic characteristics and limited availability of labeled datasets. This research paper presents a novel hybrid deep learning model to address these challenges and improve sentiment analysis performance in Vietnamese texts.

II. RELATED WORK

Le et al. (2020) [1] presented the results of applying BERT, a transfer learning method, in the Vietnamese benchmark for one of the text classification problems, the Aspect Based Sentiment Analysis problem. The experiments were conducted on two data sets, named Hotel and Restaurant, in two tasks: Aspect Detection and Aspect Polarity. The obtained results have outperformed the previous research [2]–[4] in precision, recall, and F1 measures.

In the research [5], Tran et al. (2021) introduce a new approach to combine PhoBERT and SentiWordnet for Sentiment Analysis of Vietnamese reviews. The proposed sentiment analysis model uses PhoBERT for Vietnamese, which is a robust optimization for Vietnamese of the prominent BERT model, and SentiWordNet, a lexical resource explicitly devised for supporting sentiment classification applications. Experimental results on the dataset VLSP 2016 and AIVIVN 2019 have achieved good performance in comparison to other models for sentiment analysis system.

The study by Huang et al. (202) [6] proposes an effective sentiment analysis model in Vietnamese. This model combines semantic and sentiment features of words in the text while improving the problem of insufficient Vietnamese corpus through transfer learning. The model was pre-training with English corpus. Finally, the sentiment of the text was classified by stacked Bi-LSTM with attention mechanism, with the input of sentiment word vector. Experiments have shown that the model can effectively improve the performance of Vietnamese sentiment analysis under small language materials.

Dang et al. (2021) [7] proposed a hybrid deep-learning model for sentiment analysis. The model combines the advantages of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. The proposed model is evaluated on three datasets: IMDB, Yelp, and Amazon. The results show that the proposed model outperforms other state-of-the-art models.

Recently, Tan et al. (2023, 2022) [8], [9] have proposed hybrid deep learning models for sentiment analysis. Two models are a combination of the RoBERTa with the GRU and LSTM recurrent neural network, respectively. They evaluated their models on three benchmark datasets for sentiment analysis and showed that they outperformed all other comparison methods on these datasets. The use of the RoBERTa transformer model in these models is a promising development, as this model is effective for a variety of natural language processing tasks.

III. METHODOLOGY

A hybrid model will be proposed, using RoBERTa for tokenization and embedding, stacked Bi-GRU for contextual information, Dropout for preventing overfitting, and Dense layers for fully connected and classification. First, some pre-processing steps are performed to clean the text. The textual data are then sent into the proposed model for feature extraction and classification.

A. RoBERTa-BiGRU

The proposed model is a hybrid deep learning model that consists of a RoBERTa layer, a Bi-GRU layer, a fully connected layer (Dense), a Dropout layer, and a classification layer (Dense). The RoBERTa model was chosen for this structure because it is effective for sentiment analysis in previous research [8]–[11].

RoBERTa is a large language model that was introduced in 2019. It is a robustly optimized BERT pretraining approach that improves the performance of BERT in many NLP tasks, including sentiment analysis. The Bi-GRU layer uses gated recurrent units (GRUs), which are a type of recurrent neural network. GRUs are similar to Long Short-Term Memory units, but they have a simpler architecture and fewer parameters. GRUs are effective for a variety of tasks, including sentiment analysis [7], [12]–[14]. In the proposed architecture of the model, bidirectional GRUs are used because of the advantages of bidirectional recurrent neural networks (Bi-RNN). Bi-RNNs can process input sequences in both the forward and backward directions, which allows them to capture long-range dependencies in the data. This can be beneficial for tasks such as sentiment analysis, where it is essential to consider the context of a word or phrase to determine its sentiment [15], [16].

The input data is first tokenized and encoded by RoBERTa. It produces a sequence of word embeddings. The last hidden state of the RoBERTa encoder is dropped out and then passed to the Bi-GRU. The Bi-GRU layer processes the last hidden state in both the forward and backward directions. This allows the layer to capture long-range dependencies in the data. The output of the BiGRU layer is then passed to a Dense layer, a Dropout layer, and then a classification layer, which predicts the sentiment of the input text. The architecture of the proposed RoBERTa-BiGRU model is shown in Fig. 1.

1) RoBERTa

The first layer of the proposed model employs in Robustly Optimized BERT Pretraining Approach (RoBERTa). The RoBERTa is an improved version of the pre-trained BERT model. The models are based on the Transformer architecture [17]. The Transformer model was introduced in
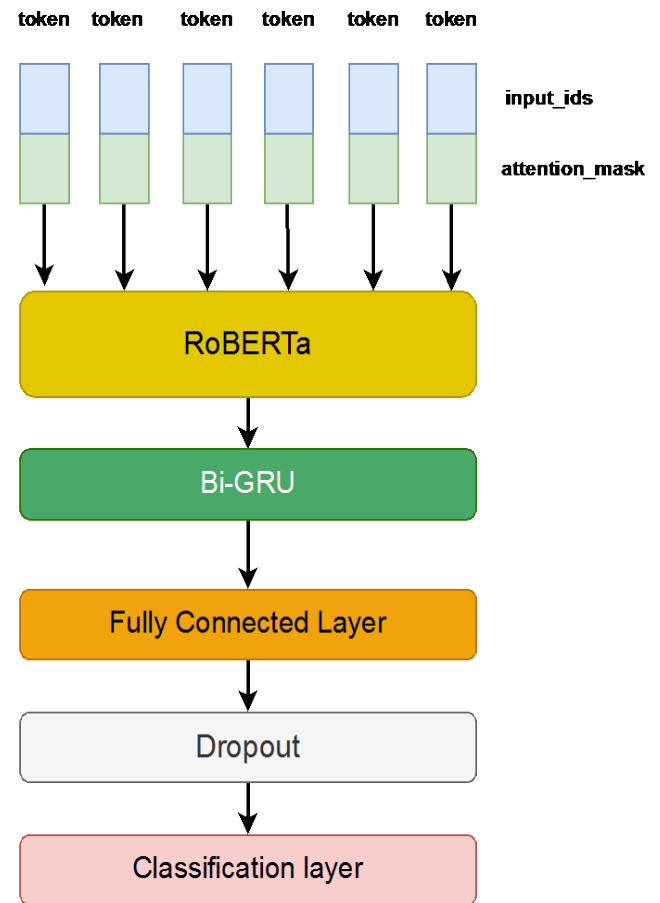

Fig. 1. The architecture of the proposed RoBERTa-BiGRU

2017 and was designed for sequence-to-sequence tasks with long-range dependencies. It uses self-attention mechanisms instead of recurrence neural networks to identify dependencies between inputs and outputs.

According to [11], some of the modifications of RoBERTa over BERT are:

- Removing BERT's next sentence pretraining objective

- Training with much larger mini-batches and learning rates, which improved the stability and convergence of the model.
- Training on a much larger dataset
- Using more dynamic masking of tokens, which reduced the exposure bias and memorization of the model.

These modifications help the RoBERTa to be more robust and perform better on a wide range of downstream tasks.

2) Bi-GRU

Bidirectional GRUs (Bi-GRU) are a neural network architecture used to process sequential data, such as text. They operate bidirectionally, capturing context from both the past and future. This feature makes them particularly effective for tasks like sequence labeling, sentiment analysis, machine translation, and speech recognition. Bi-GRU extend the traditional GRU architecture, which, along with LSTMs, addresses the vanishing gradient problem that arises when training standard recurrent neural networks on lengthy sequences.

In a standard GRU, input sequences are processed sequentially. At each time step, the GRU unit takes the

current input and the previous hidden state to generate a new hidden state and output. GRUs incorporate reset and update gates that regulate information flow, enabling the network to learn long-term dependencies and mitigate the vanishing gradient problem. The structural components of the GRU unit are illustrated in Fig. 2.
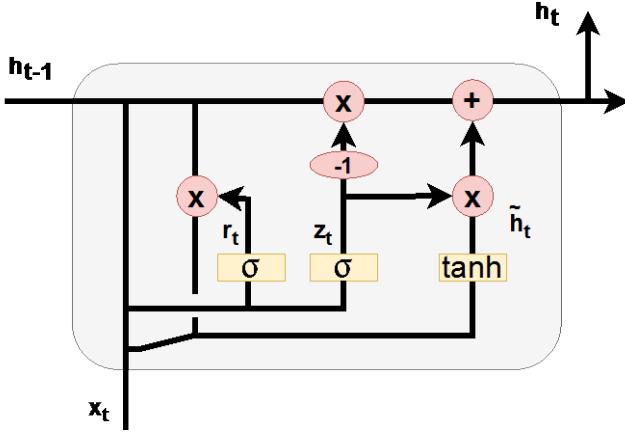

Fig. 2. The architecture of the GRU unit

The computations in the GRU unit are defined as:

$$r_t = \sigma(W_r \cdot [h_{t-1}; x_t] + b_r) \qquad (1)$$

$$z_t = \sigma(W_z \cdot [h_{t-1}; x_t] + b_z) \qquad (2)$$

$$\tilde{h}_t = tanh(W_h \cdot [(r_t \odot h_{t-1}); x_t] + b_h) \qquad (3)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \qquad (4)$$

where z_t and r_t denote the update gate and reset gate, h_t - the actual output state, $\tilde{h}_t$ - the candidate output state, and the symbol $\odot$ - element-wise multiplication. σ - a sigmoid function. $W_r, W_z, W_h \in \mathbb{R}^{d_h \times (d+d_h)}$, $b_r, b_z, b_h \in \mathbb{R}^{d_h}$ are the parameters of the update and reset gates and, d_h - the dimension of the hidden state.

In contrast, the Bi-GRU concatenate the hidden states of both forward and backward GRU units at each time step. This mechanism enables it to capture contextual information from both directions of the input sequence.

By comprehensively capturing the context at each step, Bi-GRU excels in identifying patterns and intricate dependencies within the data. Additionally, it reduces computational complexity compared to bidirectional LSTM.

3) Dropout layer

A dropout layer is a regularization technique commonly used in neural networks to prevent overfitting and enhance the model's generalization capabilities. Overfitting occurs when a model learns to perform well on the training data but struggles to generalize to new, unseen data. The dropout layer helps mitigate this issue by randomly dropping out (setting to zero) some of the neurons in a layer during training. The dropout rate is the probability that a neuron will be dropped out. The dropout rate can be adjusted depending on the network architecture and the amount of overfitting that is observed.

4) Dense layer

A fully connected layer or Dense layer is a layer in a neural network where each neuron is connected to every neuron in the previous layer. That means the output of each neuron in the previous layer is used as an input to each neuron in the fully connected layer.

The Dense layers are often used as the final layer in a neural network, where they are used to classify or predict the output. A last dense layer serves as the classification layer, where the softmax activation function is applied to compute the probability distributions of the sentiment classes.

B. Data

The proposed RoBERTa-BiGRU model was evaluated on two publicly available sentiment analysis datasets: the Internet Movie Database (IMDb) dataset [18] and the benchmark Vietnamese sentiment analysis dataset (UIT-VSFC) [19]. Fig.3 and 4. illustrate the datasets' distributions.

IMDb dataset consists of a collection of movie reviews from the IMDb, where each review is labeled with either a positive or negative sentiment based on the overall sentiment expressed in the text. The IMDb dataset typically consists of 50,000 labeled movie reviews, with approximately half being positive and the other half negative.

The dataset of 50,000 labeled samples was divided into training, validation, and test sets using a ratio of 60-20-20. Fig. 3. displays the data distribution applied to the training set, the validation set, and the test set. Table 1 shows the numbers of the samples in three sets.

UIT-VSFC dataset. The UIT-VSFC dataset contains 16,000 Vietnamese students' feedback, with three classes: positive, neutral, and negative. As the dataset is imbalanced, with a limited number of neutral samples, only positive and negative samples were used for experimentation. The sample distribution across the three datasets (training set, validation set, test set) is presented in Table 2 and Fig. 4.

In the research, the raw text of the IMDb dataset was pre-processed by removing stop words, punctuation, and digits and converting it to lowercase, followed by tokenization. However, when dealing with the UIT-VSFC dataset, it is not recommended to remove stop words due to the intricate semantics of the Vietnamese language. Furthermore, words with a length exceeding 7 letters were excluded from the text because this is the maximum length of a word in Vietnamese.

Table 1. The distributed samples in IMDb datasets

Dataset	Number of samples
Train set	30000
Test set	10000
Valid set	10000

Table 2. The distributed samples in UIT-VSFC datasets

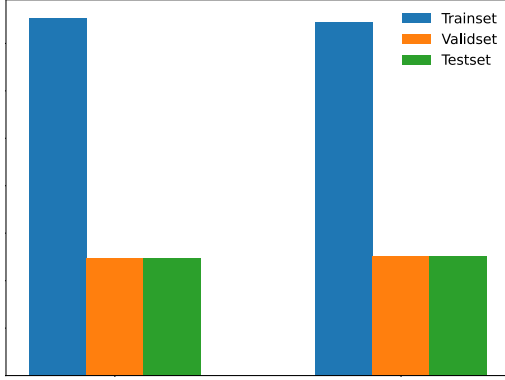| Dataset	Number of samples	Number of samples without neutral samples
Train set	11426	10968
Test set	3166	2999
Valid set	1583	1510

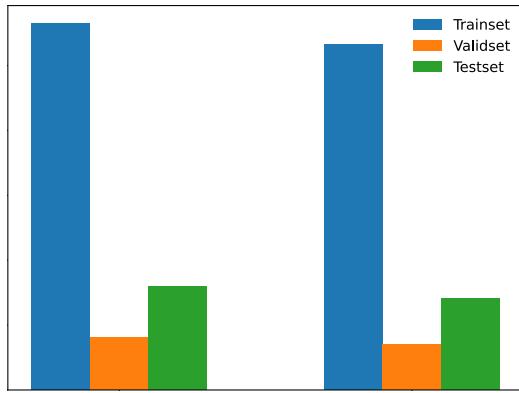Fig. 3. Data distribution in three sets of IMDb

Fig. 4. Data distribution in three sets of UIT-VSFC

C. Evaluation metrics

The proposed model was then compared with several baseline and hybrid models: BERT, BERT-GRU, BERT-BiGRU, RoBERTa, and RoBERTa-GRU. The performance of the models was evaluated using the metrics of accuracy, precision, recall, and F1-score.

Four functional accuracy metrics were considered, including false positive (FP), true positive (TP), false negative (FN), and true negative (TN).

The formulas to evaluate the four metrics are presented below.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (5)$$

$$Recall = \frac{TP}{TP + FN} \qquad (6)$$

$$Precision = \frac{TP}{TP + FP} \qquad (7)$$

$$F1 = \frac{2 * Precision * Recall}{(Precision + Recall)} \qquad (8)$$

IV. EXPERIMENTAL RESULT

For experimentation, the models were implemented using TensorFlow [19] and trained on Google Colab with TPUs. The Adam optimizer was used with a learning rate of $1e^{-5}$ and a batch size of 64. To prevent overfitting, a dropout rate of 0.25 was implemented in the Dropout layer. The models are trained with a limitation of 30 epochs. To prevent overfitting and conserve computational resources, the EarlyStopping technique was adopted. This involved utilizing the EarlyStopping callback from the tf.keras library with a "patience" of 5. The specific parameters of the models are shown in Table 3.

Table 3. The setting of the parameters

Parameter	Value
Optimizer	Adam
Dimension	256
Batch size	64
Dropout	0.25
Max epochs	30
Learning rate	1e-5
Activation function (fully connected layer)	ReLu
Activation function (classification layer)	softmax

The experimental results show that the proposed hybrid deep learning model achieves state-of-the-art performance on the IMDb and UIT-VSFC datasets. Specifically, the model achieves an accuracy of 95.35% on the IMDb dataset and an accuracy of 93.86% on the UIT-VSFC dataset. This is significantly better than the performance of the BERT-based models, which achieved accuracies of 93.03% and 92.26% on the IMDb and UIT-VSFC datasets, respectively.

The results of the experiments are shown in Tables 4, 5, and 6 and illustrated in Fig. 3 and Fig. 4.

The proposed RoBERTa-BiGRU model achieved the best results on the IMDb test dataset, with an accuracy of 95.35%, a precision of 95.40%, a recall of 95.35%, and an F1 score of 95.35%. Table 5 presents a comparison of accuracy, precision, recall, and F1-score among different methods in previous studies using the IMDb dataset. The proposed RoBERTa-BiGRU model surpasses all other models with an accuracy of 95.35%.

The RoBERTa-BiGRU also achieved the best results on the UIT-VFSC test dataset, with an accuracy of 93.86%, a precision of 93.88%, a recall of 93.86%, and an F1 score of 93.87% (Table 6).

To further explore the ability to discriminate between accurate and inaccurate classification, ROC curves were plotted for six models (Fig. 5 and Fig. 6) on the test set of IMDb and UIT-VSFC.

Table 4. The experimental result on IMDb dataset. ''Acc.'' is the abbreviation of accuracy; ''P.,'' precision; ''R.,'' recall; F1 represents F1-score.

Methods	Acc	P	R	F1
BERT	93.05	93.26	93.05	93.04
BERT-GRU	93.22	93.24	93.22	93.22
BERT-BiGRU	93.87	93.88	93.87	93.87
RoBERT	95.17	95.18	95.17	95.17
RoBERTa-GRU	94.87	94.87	94.87	94.87
RoBERTa-BiGRU (proposed)	**95.35**	**95.40**	**95.35**	**95.35**

Table 5. The comparison results on the IMDb.

Methods	Acc	P	R	F1
CNN-LSTM [20]	86.16	86.00	86.00	86.00
LSTM [21]	89.9	-	-	-
BiGRU-Attention + Hybrid CNN [22]	90.30	90.34	90.29	90.32
RoBERTa – LSTM [9]	92.96	93.00	93.00	93.00
RoBERTa-GRU [8]	94.63	95.00	95.00	95.00
RoBERTa-BiGRU (proposed)	**95.35**	**95.40**	**95.35**	**95.35**

Table 6. Comparative experimental results on the UIT-VSFC dataset

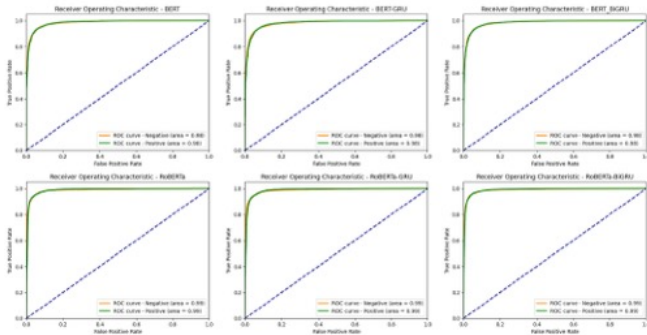| Methods	Acc	P	R	F1
BERT	92.60	92.62	92.60	92.59
BERT-GRU	93.10	93.11	93.10	93.09
BERT-BiGRU	93.10	93.10	93.10	93.10
RoBERT	93.43	93.46	93.43	93.43
RoBERTa-GRU	92.82	92.94	92.83	92.84
RoBERTa-BiGRU (proposed)	**93.86**	**93.88**	**93.86**	**93.87**

Fig. 5. Receiver Operating Characteristic (ROC) curves for six models on IMDb test set
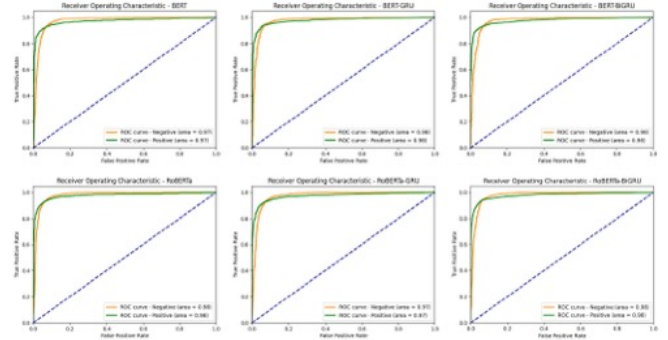

Fig. 6. Receiver Operating Characteristic (ROC) curves for six models on the UIT-VSFC test set

V. CONCLUSION

Sentiment analysis is an important task in natural language processing that allows us to understand the public sentiment expressed through various channels, such as applications, social media, and websites. The outcomes derived from sentiment analysis can subsequently be leveraged to inform decision-making processes concerning management and operations in a variety of fields.

In this study, a novel hybrid method was proposed to develop a deep learning model using two distinct datasets: IMDb (for English) and UIT-VSFC (for Vietnamese). The pre-processing phase involved applying techniques such as removing stop words, punctuation, and digits, and converting text to lowercase. Additionally, each sentence was tokenized to generate input ids and attention masks for text lines. Pre-trained BERT classifier models, namely BERT (base-uncased) and RoBERTa (roberta-base), were utilized in conjunction with Bi-GRU to enhance the performance of the models. The hybrid RoBERTa-BiGRU model has achieved significant performance improvements of 2.2% in accuracy for the IMDb dataset and 1.26% for the UIT-VFSC dataset compared to BERT alone.

The results of the experiments carried out on the IMDb and UIT-VSFC datasets indicate that the RoBERTa-BiGRU model outperforms all other comparison methods, with F1 scores of 95.35% and 93.87%, respectively.

The proposed model is effective for sentiment analysis, and its combination of RoBERTa and Bi-GRU is a promising approach for a variety of natural language processing tasks.

REFERENCES

[1] N. C. Le, N. The Lam, S. H. Nguyen, and D. Thanh Nguyen, "On Vietnamese Sentiment Analysis: A Transfer Learning Method," *Proceedings - 2020 RIVF International Conference on Computing and Communication Technologies, RIVF 2020*, Oct. 2020, doi: 10.1109/RIVF48685.2020.9140757.

[2] H. T. M. Nguyen *et al.*, "VLSP SHARED TASK: SENTIMENT ANALYSIS," *Journal of Computer Science and Cybernetics*, vol. 34, no. 4, pp. 295–310, Jan. 2019, doi: 10.15625/1813-9663/34/4/13160.

[3] H. Hamdan, P. Bellot, and F. Béchet, "Supervised Methods for Aspect-Based Sentiment Analysis," *8th*

International Workshop on Semantic Evaluation, SemEval 2014 - co-located with the 25th International Conference on Computational Linguistics, COLING 2014, Proceedings, pp. 596–600, 2014, doi: 10.3115/V1/S14-2104.

[4] D. Van Thin, V. D. Nguye, K. Van Nguyen, and N. L. T. Nguyen, "Deep Learning for Aspect Detection on Vietnamese Reviews," *NICS 2018 - Proceedings of 2018 5th NAFOSTED Conference on Information and Computer Science*, pp. 104–109, Jan. 2019, doi: 10.1109/NICS.2018.8606857.

[5] H. V. Tran, V. T. Bui, D. T. Do, and V. V. Nguyen, "Combining PhoBERT and SentiWordNet for Vietnamese Sentiment Analysis," *Proceedings - International Conference on Knowledge and Systems Engineering, KSE*, vol. 2021-November, 2021, doi: 10.1109/KSE53942.2021.9648599.

[6] Y. Huang, S. Liu, L. Qu, and Y. Li, "Effective Vietnamese Sentiment Analysis Model Using Sentiment Word Embedding and Transfer Learning," *Communications in Computer and Information Science*, vol. 1258 CCIS, pp. 36–46, 2020, doi: 10.1007/978-981-15-7984-4_3/COVER.

[7] C. N. Dang, M. N. Moreno-García, and F. De La Prieta, "Hybrid Deep Learning Models for Sentiment Analysis," *Complexity*, vol. 2021, 2021, doi: 10.1155/2021/9986920.

[8] K. L. Tan, C. P. Lee, and K. M. Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Applied Sciences 2023, Vol. 13, Page 3915*, vol. 13, no. 6, p. 3915, Mar. 2023, doi: 10.3390/APP13063915.

[9] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.

[10] W. Liao, B. Zeng, X. Yin, and P. Wei, "An improved aspect-category sentiment analysis model for text sentiment analysis based on RoBERTa," *Applied Intelligence*, vol. 51, no. 6, pp. 3522–3533, Jun. 2021, doi: 10.1007/S10489-020-01964-1/METRICS.

[11] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 2019, Accessed: Jul. 27, 2023. [Online]. Available: https://arxiv.org/abs/1907.11692v1

[12] R. Ni and H. Cao, "Sentiment Analysis based on GloVe and LSTM-GRU," *Chinese Control Conference, CCC*, vol. 2020-July, pp. 7492–7497, Jul. 2020, doi: 10.23919/CCC50068.2020.9188578.

[13] Y. Cheng *et al.*, "Sentiment Analysis Using Multi-Head Attention Capsules with Multi-Channel CNN and Bidirectional GRU," *IEEE Access*, vol. 9, pp. 60383–60395, 2021, doi: 10.1109/ACCESS.2021.3073988.

[14] Y. Pan and M. Liang, "Chinese Text Sentiment Analysis Based on BI-GRU and Self-attention," *Proceedings of 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference, ITNEC 2020*, pp. 1983–1988, Jun. 2020, doi: 10.1109/ITNEC48623.2020.9084784.

[15] A. Chaudhuri and S. K. Ghosh, "Sentiment analysis of customer reviews using robust hierarchical bidirectional recurrent neural network," *Advances in Intelligent Systems and Computing*, vol. 464, pp. 249–261, 2016, doi: 10.1007/978-3-319-33625-1_23/COVER.

[16] M. Jabreel, F. Hassan, and A. Moreno, "Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks," *Smart Innovation, Systems and Technologies*, vol. 85, pp. 39–55, 2018, doi: 10.1007/978-3-319-66790-4_3/COVER.

[17] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017.

[18] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning Word Vectors for Sentiment Analysis".

[19] K. Van 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," *Proceedings of 2018 10th International Conference on Knowledge and Systems Engineering, KSE 2018*, pp. 19–24, Dec. 2018, doi: 10.1109/KSE.2018.8573337.

[20] P. K. Jain, V. Saravanan, and R. Pamula, "A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, Sep. 2021, doi: 10.1145/3457206.

[21] S. M. Qaisar, "Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory," *2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020*, Oct. 2020, doi: 10.1109/ICCIS49240.2020.9257657.

[22] Q. Zhu, X. Jiang, and R. Ye, "Sentiment Analysis of Review Text Based on BiGRU-Attention and Hybrid CNN," *IEEE Access*, vol. 9, pp. 149077–149088, 2021, doi: 10.1109/ACCESS.2021.3118537.

Nguyen Thi Mai Trang, Ph.D., Lecturer of the Department of Computer Sience, Faculty of Information Technology, Posts and Telecommunications Institute of Technology (PTIT) of Vietnam.
Email: m.trang91@gmail.com.
ORCID : orcidID = https://orcid.org/0000-0002-3416-659X