International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 13, Issue 12, December 2023)

Manuscript Received: 18 Oct 2023, Received in Revised form: 26 November 2023, Accepted: 02 December 2023 DOI: 10.46338/ijetae1223_01

XBert - A Model for Hate Speech Detection in Vietnamese Text

Duy Nguyen Minh Le[1], Huy Gia Le[2], Hai Thanh Hoang[3], Vu Anh Hoang[4]

[1,2,3,4]*Information Technology Department, Artificial Intelligence Program, FPT University, Ho Chi Minh City, Viet Nam*

***Abstract*— In the digital age, social media's pervasive influence has inadvertently escalated the prevalence of hate speech and offensive comments, with alarming implications for mental health. There is increasing evidence indicating a clear correlation between two factors. exposure to such toxic online content and the onset of depression among users, particularly affecting vulnerable groups like content creators and channel owners. Addressing this critical issue, our research introduces XBert, a model for detecting hostile and provocative language in Vietnamese. We propose an approach related to data preprocessing, improved tokenization, and model fine-tuning. We have modified the architecture of the Roberta model, used the EDA technique, and added a dropout parameter to the tokenizer. Our model achieved an accuracy of 99.75% and an F1-Macro score of 98.05%. This is a promising result for a model detecting provocative and hostile language in Vietnamese.**

***Keywords*—Hate speech detection, NLP, Roberta, Tokenizer, Transformer**

I. INTRODUCTION

In today's era where science, technology, and engineering are rapidly evolving, everyone has been able to easily own an electronic device to access social media sites. And the most popular social networks nowadays, such as Facebook and Instagram, serve as a venue for people to chat, discuss, and exchange ideas on various topics. However, with the excessive popularity and easy accessibility, numerous intractable issues have emerged. And one of the most pressing issues is the increasing prevalence of hate speech, which is becoming difficult to control. This is a problem that needs to be thoroughly addressed, or else it will lead to unpredictable consequences.

Hate speech can be easily found in posts, user comments, and even private messages. According to Stacy Jo Dixon[1], in the first quarter of 2023, the social media platform Instagram detected and removed 5.1 million hate speech contents, an increase of 4.7 million contents compared to the previous quarter. The peak of hate speech content detection was in the second quarter of 2021 with 9.8 million contents. As for the social media platform Facebook, according to the Statista Research Department[2], this platform detected and removed 18 million hate speech contents, an increase of 10.7 million contents compared to the first quarter of 2023.

The peak was between April and June 2021, during which Facebook removed over 31 million hate speech contents. And according to Microsoft[3], in 2023, on online platforms, approximately 35% of the content contains hate speech.

Recent studies have established a notable connection between exposure to online hate speech and the onset of mental health issues like depression. A study on adolescents highlighted this link, showing that those victimized by online hate speech were more likely to exhibit depressive symptoms. The study, involving 1,632 adolescents, also indicated that resilience could somewhat mitigate these effects[4]. Similarly, research on the psychological effects of hateful speech in online college communities found that exposure to such speech increased stress expression, negatively impacting mental health. This study, which analyzed 6 million Reddit comments[5], noted that individuals with lower psychological endurance were more vulnerable to emotional outbursts and more prone to neuroticism.

Additionally, general research on social media's impact on mental health corroborates these findings, linking heavy social media use to an increased risk for depression, anxiety, and loneliness, with about 10% of teens reporting being bullied on social media[6]. These studies collectively underscore the detrimental impact of online hate speech and offensive comments on mental health, particularly among younger and more vulnerable populations. For the Vietnamese language, based on a study by VPIS[7], in a survey of about 1,000 people, 78% of the respondents admitted that they had been victims or were aware that hate speech on social media is becoming increasingly prevalent. Any spelling, grammar, and punctuation errors have been corrected.

Such speech is not merely an expression of negative sentiments. It has the potential to incite violence, amplify discrimination against individuals or groups based on factors like skin color, religion, gender, or ethnicity, and can also spread misleading information that tarnished the reputation of individuals or organizations. The repercussions of unchecked hate speech can be severe, leading to unforeseen consequences, deepening internal divisions, and fostering an unhealthy environment. In this context, this article introduces an enhanced model designed to detect hate speech in the Vietnamese language. The essential sections of the article can be described as follows:

The article employs the BPE-Dropout algorithm[8] to process Vietnamese words. It stochastically corrupts the segmentation procedure of BPE, leading to the generation of multiple segmentations within the same fixed BPE framework. We introduced a dropout variable to BPE, randomly eliminating some of the merge operations that BPE typically performs.

We have altered the existing architecture of the Roberta model[9]. Through modifications and enhancements, we have restructured a new configuration tailored for hate speech detection. This improved model demonstrates commendable accuracy in the Vietnamese language.

II. Preprocessing Data

Firstly, we will handle the placement of diacritics in Vietnamese because, currently, there are two ways to place diacritics, the old style and the new style. Therefore, we will synchronize and change any words using the old style to the new style. The rules for the new formatting style include:

For syllables whose main vowel is a single vowel, the placement of diacritics will be at the position of the letter representing the main vowel.

For syllables whose main vowel is a double vowel, for vowels written as: iê, yê, uô, ươ; the ending vowels are written with: p, t, c, ch, m, n, ng, nh, o, u, i, the diacritic will be placed on the second letter of the two letters of the main vowel. Whereas for vowels written as: ia, ya, ua, ưa, the diacritic will be placed on the first letter of the two letters of the main vowel.

Next, we will clean up the input text by replacing emojis with spaces, removing Unicode characters, deleting special characters, deleting URLs, removing extra spaces, and converting all letters to lowercase in the text.

Next, we will process stop words, which are common words that do not carry significant meaning and are often removed to reduce the dimensionality of the text data. We will use a list of Vietnamese stop words to remove these words from the input text[10].

Finally, for special characters that are not letters and are repeated many times in a sentence, we will remove them and only keep one special character.

III. Increase Text Data

Since the dataset that we collected is quite limited, we used the EDA (Easy Data Augmentation)[11] method to improve our dataset. This method includes four techniques: word replacement, word insertion, word position change, and word deletion. With the use of the EDA method, the performance of our model has significantly improved compared to before using it.

IV. Tokenizer

A. Byte Pair Encoding (BPE)

Byte Pair Encoding (BPE) is a data compression technique, first introduced in 1994, that enhances the efficiency of NLP models. The algorithm works by breaking words into subwords.

We employ a technique introduced in the paper 'Neural Machine Translation of Rare Words with Subword Units' [12] in 2016, which is capable of breaking words into subword units. Based on this technique, we can represent most subwords and will handle words that have never appeared.

The BPE technique statistically analyzes the frequency of subwords and merges them if they have the highest occurrence. This process continues until there are no more subwords to merge.

B. Subword regularization

Subwordregularization[13] is introduced as a technique that combines multiple segmentation approaches. This method entails training a unigram model with a dedicated algorithm to predict the probability of each subword with the utmost precision. Additionally, it involves employing the EM algorithm to optimize the vocabulary and the Viterbi algorithm to generate segmentation samples.

Subword regularization has been demonstrated to yield significant improvements compared to using a single subword sequence. However, this subword regularization technique is quite intricate and cannot be applied to regular BPE.

C. BPE-Dropout

BPE-Dropout[8] is a simple yet highly effective subword regularization technique. It is built upon BPE and can be applied to traditional BPE. The key operation of the BPE-Dropout technique is that it randomly perturbs the segmentation process of the BPE algorithm, thereby leading to the generation of multiple segmentations within the same fixed BPE framework.

D. Our Approach

We have implemented the dropout technique in the BPE process, specifically our dropout parameter is used to control the vocabulary compression frequency of subword units during BPE. The reason we apply this technique is because the dropout parameter helps control the compression level and ensures that the BPE process does not lose too much information compared to the original data. The efficiency after applying the dropout parameter has improved by an additional 1-2% accuracy compared to the initial state.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 13, Issue 12, December 2023)

V. Model Architecture: Enhanced XBert For Vietnamese Hate Speech Detection

For our research, our model is based on the "Roberta For Sequence Classification" class from the Hugging Face Transformer library[14], a specialized version of the RoBERTa-based model[9] for Vietnamese hate speech detection[15]. This section breaks down the model's component:

A. Roberta Embeddings

The Foundation of our model is the "Roberta Embedding" layer. This layer is responsible for converting input tokens into dense vectors that can be processed by the subsequent layers:

Word Embeddings: An embedding matrix of size 64001x768, responsible for converting token IDs to vectors, capturing the semantic essence of each token[16].

Position Embeddings: An embedding matrix of size 256x768. This provides the model with information about the relative or absolute position of tokens within a sequence[17].

Token Type Embedding: An embedding layer of size 1x768. While RoBERTa doesn't utilize token type embeddings in the same manner as BERT, this layer can be adapted for specific tasks[18].

Layer Normalization and Dropout: These are standard regularization techniques to prevent overfitting, respectively[19].

B. Roberta Encoder

This is the heart of the RoBERTa model, where the actual processing of the embeddings takes place. It consists of 12 RobertaLayer modules, each containing:

Self Attention Mechanism: This mechanism allows the model to weigh the importance of different tokens in a sequence relative to a given token. It comprises query, key, and value linear transformations, followed by a dropout for regularization[17].

Intermediate Layer : A feed-forward neural network with a GELU activation function. It temporarily expands the dimensionality to 3072, allowing the model to capture more complex relationships before compressing it back[20].

Output Layer: A subsequent feed-forward network that reduces the dimensionality back to 768. This is followed by layer normalization and dropout for stabilization and regularization[21].

C. XBert Layer

Our novel contribution, the XBert layer, is introduced to further refine the encoder's output for the specific task of Vietnamese hate speech detection:

- A linear layer with input and output features of size 768.
- Layer normalization and dropout layers for regularization[21].

D. Classifier

The final classification head that predicts the hate speech labels. It consists of:

- A dense layer with input features of size 768 and input features of size 768.
- A dropout layer for regularization
- An output projection layer that maps the 768-dimensional vector to 3 classes, corresponding to the hate speech labels[22]

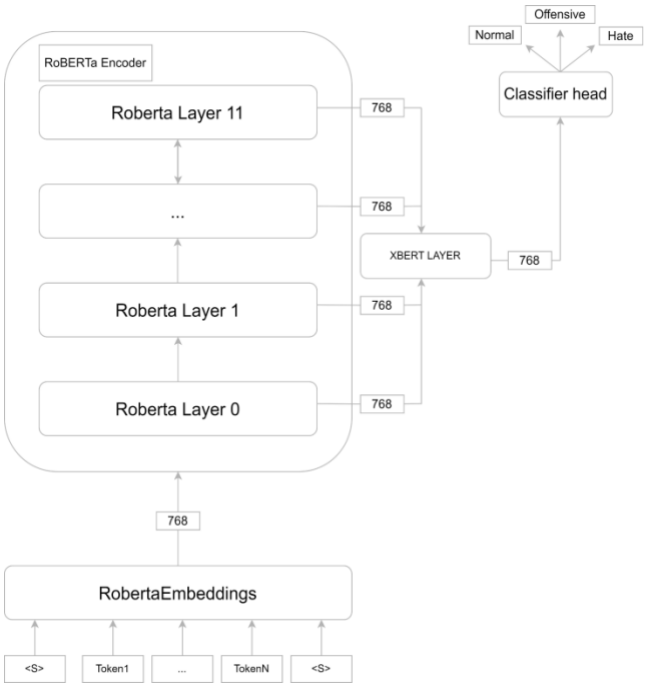

Fig .1.ModelXbert with add 1 layer XBert before layer classifier head[23]

The model is designed to classify each input sequence into one of three categories: CLEAN, OFFENSIVE, or HATE. The architect leverages the power of the RoBERTamodel[9], known for its superior performance in various NLP tasks, and customizes it for the specific task of Vietnamese hate speech detection.

VI. Dataset Description

We use two datasets:

The first dataset is introduced in the paper "HSD Shared Task in VLSP Campaign 2019: Hate Speech Detection for Social Good."[24].

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 13, Issue 12, December 2023)

This dataset has been curated to facilitate research in the domain of hate speech detection, specifically for the Vietnamese language. The dataset includes 25,431 samples. It was crawled from Facebook, a social media platform widely used by many people in Vietnam. This has three classes: CLEAN (18,614 samples), OFFENSIVE (1,022 samples), HATE (709 samples). The dataset was crawled from comments and posts on social networks. So, the dataset has a lot of noise, such as emoji, teencode(e.g., ``k", ``ko", ``khum" which means ``không"; etc.) , special characters, Vietnamese stopwords (e.g., ``và", ``là", ``đi", ``khi", etc.), etc.

TABLE I
EXTRACTION DATASET HSD-VLSP

ID	Text	Label
1	Đmmấythằngcườitrôngbựavl	HATE
2	t đây, cómôn đ cầnônvẫn qua môn, nể t vl	OFFENSIVE
3	MìnhcầnbáncâyIphone X bản 64gb đểlên XS Max	CLEAN

The second dataset is introduced in the paper "A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts[25]." Collect data from user comments on Vietnamese Facebook's comments and Youtube videos that are highly interactive. This dataset contains over 30,000 comments, and each comment is annotated with three labels: CLEAN(0), OFFENSIVE(1), and HATE(2).

VII. EXPERIMENTS AND RESULT

A. Evaluation metrics

Macro F1-score serves as a widely used assessment measure in classification tasks. F1-score is the harmonic mean of Precision and Recall.

F1-score: performance evaluation for classification.

$$F_1 = \frac{2}{Recall^{-1} + Precision^{-1}}$$

where *Precision* is the number of the correct positive results divided by the number of positive results, and *Recall* is the number of correct positive results divided by the number of all samples that should have been identified as positive.

F1-macro: computed as mean of F1 scores for each class.

$$F_{1-Macro} = \frac{F_{1-HATE} + F_{1-OFFENSIVE} + F_{1-CLEAN}}{3}$$

B. Training Procedure

To ensure the model's generalization to mitigate overfitting, we employed K-fold cross-validation[26] with K = 10. This means the VLSP and ViHSD datasets were divided into 10 subsets. In each iteration, 9 subsets were used for training, and the remaining subset was used for validation. These datasets are particularly relevant to the task at hand, given their comprehensive collection of labelled instances pertinent to Vietnamese hate text. The hyperparameters for the model were judiciously selected after a series of preliminary experiments. Specifically, we utilized a maximum sequence length of 60 tokens, a batch size of 128, and a learning rate of 2e-5. These hyperparameters were held constant across all folds to ensure the consistency and reliability of our results.

C. Baseline Models

For the purpose of comparative analysis, we selected two existing state-of-the-art models as baselines: PhoBERT[27] and PhoBERT-CNN[28].PhoBERT serves as a robust benchmark, given its prior success in various Vietnamese natural language processing tasks. PhoBERT-CNN, on the other hand, represents the most recent advancement in the specific domain of Vietnamese hate text detection, as evidenced by its performance metrics reported in the latest literature.

D. Results

Our XBERT model manifested exceptional performance, surpassing the baseline models across both evaluation metrics—accuracy and macro F1-score. These results were consistent across all ten folds of the cross-validation, thereby affirming the robustness of our model.

TABLE II
RESULT

Model	VLSP		ViHSD	
	Accuracy	Macro-F1	Accuracy	Macro-F1
PhoBert	94.1	66.03	86.61	53.0
PhoBert-CNN	98.26	90.89	87.17	64.43
XBert	**99.75**	**98.05**	**96.55**	**91.67**

E. Discussion

In the discussion, we integrate a comparison of XBERT with existing research, mainly focusing on its

advancements over models like RoBERTa in the context of Vietnamese hate text detection. We highlight that while RoBERTa and similar models have established a foundation for language-specific hate speech detection, XBERT excels in handling imbalanced datasets, a challenge often inadequately addressed by previous models. This comparative analysis showcases XBERT's technical superiority and situates it within the existing landscape of hate speech detection research. Additionally, the discussion acknowledges the limitations in the current scope of XBERT's application and suggests expanding future research to diversify datasets and explore practical integration into content moderation systems.This approach addresses the feedback by illustrating how XBERT builds upon and surpasses existing methodologies in the field.

VIII. CONCLUSION

In summary, our empirical evaluation substantiates the efficacy of XBERT in the task of Vietnamese hate text detection. The model not only achieves state-of-the-art performance but also establishes new benchmarks for future research in this domain.

REFERENCES

[1] S. J. Dixon, "Instagram: hate speech-containing content removal as of Q1 2023," Statista, 2023

[2] Statista Research Department, "Facebook: hate speech content removal as of Q2 2023," Statista, 2023

[3] Microsoft, "Digital Civility Challenge," Microsoft, 2023

[4] S. Wachs, M. Gámez-Guadix, M.F. Wright, "Online Hate Speech Victimization and Depressive Symptoms Among Adolescents: The Protective Role of Resilience," Cyberpsychology, Behavior, and Social Networking, vol. 1, pp 416-423, 2022

[5] K. Saha, E. Chandrasekharan, M. De Choudhury, "Prevalence and Psychological Effects of Hateful Speech in Online College Communities," in Proceedings of the ACM Web Science Conference, pp. 255-264, 2019

[6] L. Robinson, M. Smith, "Social Media and Mental Health," HelpGuide.org, 2023

[7] VPIS, "Symposium: Hate Speech and Solutions Towards a Safe and Sustainable Social Network Environment," Vietnam Program for Internet and Society, 2017.

[8] I. Provilkov, D. Emelianenko, and E. Voita, "BPE-Dropout: Simple and Effective Subword Regularization," Association for Computational Linguistics, pp. 1882–1892, 2020

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, " RoBERTa: A Robustly Optimized BERT Pretraining Approach," Computer Science - Computation and Language, 2019

[10] An Long Doan, Son T. Luu. "Improving Sentiment Analysis By Emotion Lexicon Approach on Vietnamese Texts," International Conference on Asian Language Processing, 2022

[11] J. Wei and K. Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks," Association for Computational Linguistics, pp 6382–6388, 2019.

[12] R. Sennrich, B. Haddow, A. Birch, "Neural Machine Translation of Rare Words with Subword Units," Association for Computational Linguistics, vol. 1, pp 1715–1725, 2016

[13] T. Kudo, "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates," Association for Computational Linguistics, vol. 1, pp 66–75, 2018

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, et al., "Transformers: State-of-the-Art Natural Language Processing," Association for Computational Linguistics, pp 38–45, 2020

[15] S. T. Luu, H. P. Nguyen, K. V. Nguyen, N. L.-T. Nguyen, "Comparison Between Traditional Machine Learning Models And Neural Network Models For Vietnamese Hate Speech Detection," RIVF International Conference on Computing and Communication Technologies, 2020

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," Neural Information Processing Systems, 2013

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention Is All You Need," Neural Information Processing Systems, 2017

[18] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp 4171–4186, 2019

[19] J. L. Ba, J. R. Kiros, G. E. Hinton, "Layer Normalization," Department of Computer Science, University of Toronto, 2016

[20] D. Hendrycks, K. Gimpel, "Gaussian Error Linear Units (GELUs)," International Conference on Learning Representations (ICLR), 2023

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," Department of Computer Science, University of Toronto, pp 1929-1958, 2014

[22] I. Goodfellow, Y. Bengio, A. Courville, "Deep Learning," MIT Press, 2016

[23] Quang Huu Pham, Viet Anh Nguyen, Linh Bao Doan, Ngoc N. Tran, "From Universal Language Model to Downstream Task: Improving RoBERTa-Based Vietnamese Hate Speech Detection" Computation and Language, 2020

[24] X.-S. Vu, T. Vu, M.-V. Tran, T. Le-Cong, and H. T. M. Nguyen, "HSD Shared Task in VLSP Campaign 2019: Hate Speech Detection for Social Good," 2020

[25] S. T. Luu, K. V. Nguyen, and N. L.-T. Nguyen, "A Large-scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts," Lecture Notes in Computer Science, vol. 12798, 2021

[26] Jung, Y., Hu, J. (2015). "A K-fold averaging cross-validation procedure". Journal of Nonparametric Statistics.

[27] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," Findings of the Association for Computational Linguistics: EMNLP 2020, pp 1037-1042, 2020

[28] K. Q. Tran, A. T. Nguyen, P. G. Hoang, C. D. Luu, T.-H. Do, and K. V. Nguyen, "Vietnamese Hate and Offensive Detection using PhoBERT-CNN and Social Media Streaming Data," Neural Computing and Applications, pp 573-594, 2022