Enhancing Performance in Targets for Hate Speech Detection Task on Vietnamese Social Media

Ma Kim Phat and **Nguyen Tien Nam** and **Tran Tri Duc** and **Luu Thanh Son**
Faculty of Information Science and Engineering
University of Information Technology, Ho Chi Minh City, Vietnam
Vietnam National University, Ho Chi Minh City, Vietnam
{22521071, 22520920, 22520276}@gm.uit.edu.vn, sonlt@uit.edu.vn

Abstract

Classifying toxic comments and identifying their targeted subjects are becoming increasingly important. The ViTHSD dataset—a dataset for detecting hate speech on Vietnamese social media platforms—enables us to determine both the level of hostility in the comments and the subjects they target. However, the performance of models on this dataset has not been satisfactory. Therefore, we propose normalization and data augmentation methods. Subsequently, the dataset will undergo these techniques and be trained using a combination of Bi-GRU-LSTM-CNN with a pre-trained language model to leverage the power of BERTology. The best results achieved after applying these techniques show an improvement of approximately 1.7% in the F1-macro score.

1 Introduction

With the rapid expansion of social media platforms, the prevalence of toxic content and harmful speech has become a significant concern. Classifying such content and identifying the specific targets of hate speech are essential tasks for addressing online toxicity. The ViTHSD dataset, introduced to detect hate speech in Vietnamese social media, not only enables the identification of offensive comments but also determines the specific targets and the level of hostility associated with them. This focus makes it a valuable resource for nuanced analyses of online hostility.

Current hate speech detection models trained on the ViTHSD dataset encounter opportunities for improvement. Challenges such as the imbalance in target label distribution and variations in linguistic expressions within the dataset highlight areas where performance can be enhanced. By addressing these aspects, the robustness and effectiveness of hate speech detection systems can be further developed.

This study aims to enhance the performance of models applied to the ViTHSD dataset through data normalization based on ViLexNorm dataset and augmentation techniques. By refining the dataset and leveraging advanced deep learning architectures such as Bi-GRU-LSTM-CNN with a pre-trained lan-guage model to leverage the power of BERToloogy, we strive to achieve significant improvements in identifying hate speech and its associated targets. The proposed methods contribute to both the accuracy and interpretability of hate speech detection, paving the way for more effective solutions to combat online toxicity.

2 Related Work

Son et al. (2024) [1] introduced a groundbreaking dataset called ViTHSD, specifically designed for offensive language detection. Unlike traditional datasets in this field, ViTHSD allows researchers to not only detect offensive comments but also identify the specific targets those comments are directed toward. This distinction makes the dataset particularly valuable for more nuanced analyses of online hostility. However, despite its innovative design, the dataset has certain limitations. Notably, it suffers from an imbalance in the distribution of target labels, which could affect the performance of machine learning models trained on it. Additionally, the comments in the dataset have not been thoroughly normalized, which may introduce inconsistencies and hinder the effectiveness of applied models.

Son et al. (2020) [2] proposed a data augmentation technique aimed at improving the quality of Vietnamese social media comments for training machine learning models. Their method was based on the data augmentation approach developed by Wei et al. (2019) [3], which had achieved significant success with English-language data. By tailoring this method to Vietnamese, Son et al. demon-

strated its potential to enhance model performance significantly. However, a critical limitation of their approach is its inability to handle multi-label data effectively, which poses a challenge for applications where comments may simultaneously target multiple entities or categories.

In another significant contribution, Thanh and Phong et al. (2024) [4] introduced the ViLexNorm dataset, which was explicitly created to normalize Vietnamese social media comments. This dataset was compiled from a diverse range of popular social media platforms in Vietnam to ensure comprehensive coverage and high accuracy. The normalization process aimed to standardize comments, addressing issues such as irregular formatting and non-standard language use. Experiments conducted using this dataset demonstrated substantial improvements across various models, with most achieving better performance compared to baseline results. Among these, the $\text{BART}_{\text{phosyllable}}$ model achieved the best performance, showcasing the effectiveness of the normalization process in enhancing model accuracy and robustness.

3 Dataset

The ViTHSD dataset was specifically developed to address the challenges of hate speech detection in Vietnamese social media. It contains 10,000 annotated comments, each meticulously labeled with a target and its corresponding level of hostility. The targets include five categories: individuals, groups, religion/creed, race/ethnicity, and politics. Each target is assigned one of three levels: Clean, Offensive, or Hateful, depending on the nature of the comment. Below is a detailed explanation of these categories with examples.

- **Individuals**: Hate speech targeting individuals often focuses on personal characteristics, such as disabilities, sexual orientation or orphanage. This is a widespread form of attack on Vietnamese social media, frequently conveyed using personal pronouns or names.

- **Groups**: Hate speech directed at groups often involves discrimination or disdain toward collective identities, such as organizations or communities. This type of content frequently includes sexism, misogyny, or hostility toward NGOs.

- **Religion/Creed**: Hate speech in this category targets individuals based on their religious beliefs or spiritual affiliations, often promoting intolerance and division.

- **Race/Ethnicity**: Hate speech related to race or ethnicity marginalizes individuals or groups based on physical traits or cultural backgrounds. Such remarks reinforce harmful stereotypes and biases. They often use racist words to insult others.

- **Politics**: Political hate speech attacks governmental institutions, political leaders, or ideologies, often with divisive and hostile rhetoric.

The ViTHSD dataset is divided into training, development, and testing subsets in an approximate ratio of 7:1:2. Table 1 presents the general statistics for these subsets, with the vocabulary size determined at the token level and the average length calculated by dividing the total length of comments by the number of comments in each subset. The training set features the largest vocabulary size, while the development set exhibits the highest average word length, suggesting that comments in this subset tend to be longer. Interestingly, despite the longer comments in the development set, its vocabulary size is smaller compared to both the training and testing sets.

	Train	**Dev**	**Test**
Num.comments	7000	1201	1800
Avg.word length	57.33	58.25	55.54
Vocab.size	12701	4547	5684

Table 1: Overview of the ViTHSD dataset

Table 2 illustrates the distribution of targets across the comments in the training subset of the ViTHSD dataset. Most of the hate speech comments target individuals and groups, making these the most prominent categories in the training set. In contrast, religion/creed has the smallest number of comments, indicating that hate speech directed at religious beliefs is relatively rare. This distribution highlights that hate speech in Vietnamese social media is predominantly aimed at individuals or groups rather than at religion, race/ethnicity, or politics.

	Num.cmts	Avg.wl	Vocab.s
Total	7,000	57.33	12,701
Individuals	5,480	3.97	5,806
Groups	2,976	3.97	4,528
Religion/creed	26	4.22	287
Race/ethnicity	502	4.06	2,263
Politics	363	4.09	1,921

Table 2: Distribution of different targets of the ViTHSD training set

4 Our Proposed Method

4.1 Data Augmentation

In this study, we apply the EDA technique introduced by Wei and Zou (2019) and Son et al. (2020) [2]. This technique takes a comment as input and performs one of the following operations to generate a new comment. Notably, some data may have more than one target. When the target which was needed to augment was applied these techniques below, we have decided that only the data having only that target was taken:

- **Synonym Replacement (SR):** This operation generates a new sentence by randomly selecting n words from the input comment (excluding stopwords) and replacing them with their synonyms. In our experiments, we use the Vietnamese WordNet by Nguyen et al. (2016) to construct the synonym and stopword dictionary.
- **Random Insertion (RI):** This operation generates a new comment by selecting a random word from the input sentence that is not a stopword, retrieving its synonym, and inserting that synonym at a random position in the sentence. The synonym is obtained from the Vietnamese WordNet.
- **Random Swap (RS):** This operation generates a new comment by randomly selecting two words from the input sentence and swapping their positions
- **Random Deletion (RD):** This operation generates a new comment by randomly deleting p words from the input sentence, where p is a predefined probability set by the user.

According to Wei and Zou (2019), the parameter n is defined as the number of words modified for the operations: synonym replacement, random insertion, random swap, and random deletion. The parameter n is calculated using the formula $n = a \times l$, where a is the percentage of words to be changed in the sentence, and l is the length of the sentence. Specifically, for the Random Deletion operation, the random deletion rate p is equal to a. The parameter a is defined by the user. [1] [2]

4.2 Normalization technique

The ViLexNorm dataset was constructed from over 10K sentence pairs manually labeled by humans, sourced from popular social media platforms in Vietnam. After applying it to datasets such as UIT-VSMEC, ViHSD, and ViSPAM, improvements in F1-macro scores ranging from 0.5% to approximately 2.7% were observed. Therefore, we decided to reconstruct the dictionary with the goal of normalizing the ViTHSD dataset, which was collected from social media platforms where comments often exhibit high complexity, irregularities, and lack of standardization.

After successfully constructing the dictionary, we decided to apply it to the dataset we are striving to improve, aiming to enhance its overall quality and usability. Furthermore, to maintain consistency and uniformity across the dataset, we also implemented a process to remove stopwords, ensuring that the data aligns with the preprocessing standards established for our analysis.

5 Experimental and Result

5.1 The augmentation and normalization data result

5.1.1 Augmentation data

Using the method outlined in section 4.1 , we now describe the distribution of the targets in the dataset before and after augmentation.

[1] https://github.com/zeloru/vietnamese-wordnet
[2] https://github.com/MaPhat/DS310-Enhancing-Peformance-ViTHSD-task

Table 3: The number of levels by each target in original and augmented dataset

Target	CLEAN	OFFENSIVE	HATE
Original Dataset			
Individual	2,480	1,169	1,831
Groups	1,406	639	932
Religion/Creed	8	8	8
Race/Ethnicity	120	163	219
Politics	37	81	245
Target	**CLEAN**	**OFFENSIVE**	**HATE**
Augmented Dataset			
Individual	2,480	1,169	1,831
Groups	1,406	639	932
Religion/Creed	8	40	40
Race/Ethnicity	1,336	1,635	1,723
Politics	229	433	1,429

This represents the data distribution for each target when applying the proposed method.

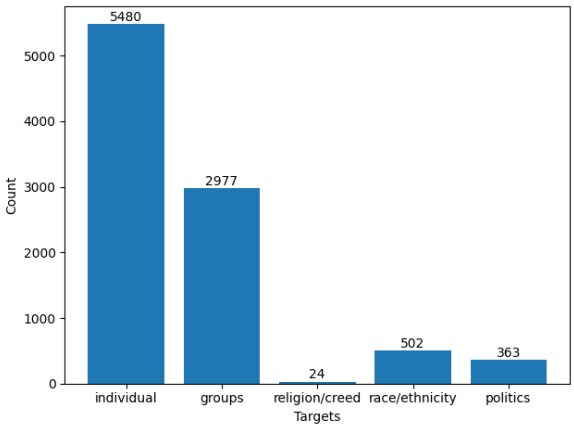

Figure 1: Original ViTHSD training set

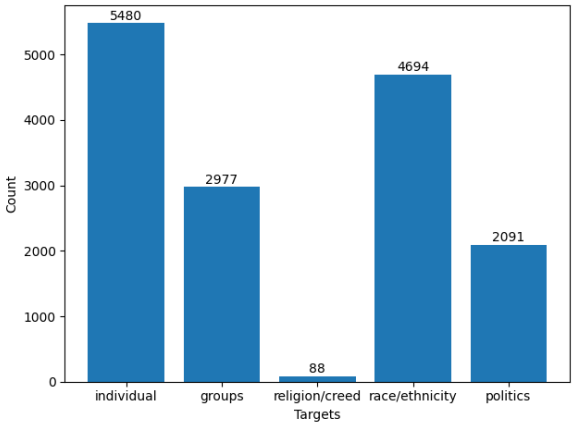

Figure 2: ViTHSD training set after augmentation

5.1.2 Normalization data

Using the method outlined in section 4.2, we illustrate the dataset before and after normalization. Below are some typical examples of what we have achieved.

Original	Normalization
trẻ trâu đóng film hài hả	trẻ trâu đóng phim hài hả
nói j thì nói nghe giọng khàn khàn éo thể nào iu dc	giọng khàn khàn đéo thể yêu
tr ơi nhìn thầy giống con póng quá à	trời thầy bóng à

Table 4: Example of sentences before and after normalization

5.2 Baseline model

Since this report aims to improve the results of the previous work, ViTHSD, the same configuration will be used again to bring more objective comparison results.

According to the preceding work, we will unfreeze the last four layers of BERT for the training parameters. The parameters are then passed to the Dropout layer and the probabilities are distributed to the five Dense layers corresponding to the five objectives of the dataset including Individual, Group, Religion, Race, and Politics. On each objective, there will be four levels including: Normal, Clean, Offensive and Hate is based on the ViHSD project. Use the Softmax function to calculate the distribution probability for the four levels: Normal, Clean, Offensive, and Hate per Dense layer. The predicted hatred levels for each target are the levels with the highest probabilities in each Dense layer across all five Dense layers. If a target is predicted to have the level "Normal," it means the target is not mentioned in the comments. Correspondingly, for the remaining predicted levels "Clean," "Offensive," and "Hate," the degree of hatred for each target will be determined. We utilize both multilingual and monolingual pre-trained BERTology to fine-tune BERT for the Vietnamese language model.

Model	Pre-Trained	#Parameter
XLM-R	xlm-roberta-base	279M
ViSoBERT	uitnlp/visobert	97M
PhoBERT	phobert-base	135M

Table 5: Representative Pre-trained BERTology models

To enhance the model's ability to extract valuable features from comments for detecting hate speech, we integrate Pre-trained BERTology with a Bi-GRU-LSTM-CNN architecture.

In this approach, Bi-GRU-LSTM-CNN leverages two powerful sequence processing models, Bi-GRU and Bi-LSTM, to extract valuable features from comments. Subsequently, a CNN layer is applied to analyze the information extracted by Bi-GRU and Bi-LSTM before forwarding it to Dense layers for classification.

We replace the embedding layer with pre-trained BERT layers to utilize BERT as a robust language model for text representation. Additionally, the final layer is replaced with a combination of five Dense layers to handle the probability distributions of hostility levels across five targets.

Finally, we construct predictive models for all Pre-trained BERTology models as mentioned in Table 5. The experimental results of our models are presented in Section 5.4

5.3 Evaluation metrics

Hate speech detection with targeted goals is considered a multi-label task, similar to ABSA tasks. We have referred to and, for the purpose of fair comparison with previous works, will evaluate the performance of the model in two stages: The target and the target with polarity (P is the predicted target and T is the true target); Precision, recall, and F1-score.

$$\text{Precision} = \frac{|P \cap T|}{|P|}$$

$$\text{Recall} = \frac{|P \cap T|}{|T|}$$

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5.4 Model performance result

We use three datasets, including: Normalization ViTHSD (only normalized), Normalization & Augmentation ViTHSD (normalized then augmented), and Augmentation & Normalization ViTHSD (augmented then normalized), to apply to four representative models from previous works in order to compare the results with the Original ViTHSD dataset.

The tables below illustrate the comparison results of the models with the mentioned datasets.

5.5 Error analysis

The data in the **"Religion/Creed"** category experienced a moderate increase after augmentation. Initially, the dataset contained only 24 rows of data. After the augmentation process, this number grew to 90 rows, indicating a noticeable yet not overly substantial expansion in the dataset size for this specific category.

A single data instance can have multiple targets, so when augmenting for a specific target, it is crucial to ensure that the instance is labeled with only one unique target. As a result, the number of data instances eligible for augmentation is limited. This limitation poses a challenge when trying to achieve a larger augmented dataset, especially if the original dataset is small. Consequently, newly generated sentences may overlap, leading to redundancy and the risk of overfitting. Evidence for this issue can be observed in the loss function graph, which performs well both before and after applying the proposed method, yet the final results remain suboptimal.

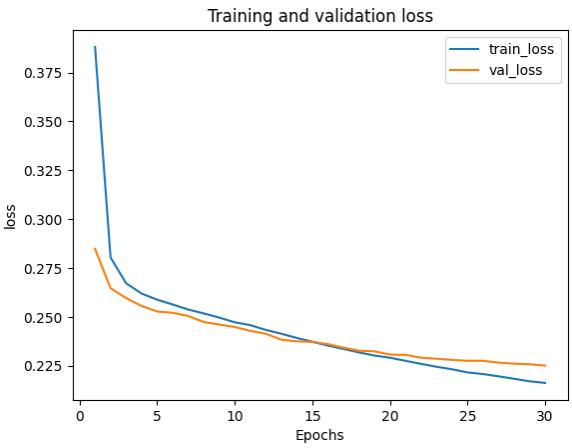

Figure 3: Loss before in XLM-R Model

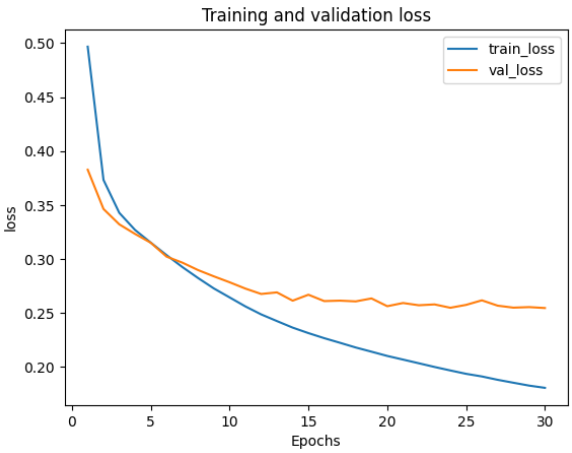

Figure 4: Loss after in XLM-R Model

Table 6: Performance comparison of models on Original and Normalization datasets [5.4]

Model	Dataset	Target Only	Target + Level
		F1-Score	F1-Score
Bi-GRU-LSTM-CNN	Original	62.57	40.89
	Normalization	60.13	40.16
XLM-R Bi-GRU-LSTM-CNN	Original	63.46	38.17
	Normalization	66.72	40.93
ViSoBERT Bi-GRU-LSTM-CNN	Original	56.98	39.06
	Normalization	60.55	38.06
PhoBERT Bi-GRU-LSTM-CNN	Original	54.55	37.89
	Normalization	60.15	38.83

Table 7: Performance comparison of models on Original and (Normalization & Augmentation) datasets [5.4]

Model	Dataset	Target Only	Target + Level
		F1-Score	F1-Score
Bi-GRU-LSTM-CNN	Original	62.57	40.89
	Normalization & Augmentation	59.65	40.91
XLM-R Bi-GRU-LSTM-CNN	Original	63.46	38.17
	Normalization & Augmentation	54.91	34.15
ViSoBERT Bi-GRU-LSTM-CNN	Original	56.98	39.06
	Normalization & Augmentation	58.24	38.15
PhoBERT Bi-GRU-LSTM-CNN	Original	54.55	37.89
	Normalization & Augmentation	58.57	38.17

Table 8: Performance comparison of models on Original and (Augumentation & Normalization) datasets [5.4]

Model	Dataset	Target Only	Target + Level
		F1-Score	F1-Score
Bi-GRU-LSTM-CNN	Original	62.57	40.89
	Augmentation & Normalization	60.12	39.06
XLM-R Bi-GRU-LSTM-CNN	Original	63.46	38.17
	Augmentation & Normalization	54.49	34.48
ViSoBERT Bi-GRU-LSTM-CNN	Original	56.98	39.06
	Augmentation & Normalization	56.79	36.88
PhoBERT Bi-GRU-LSTM-CNN	Original	54.55	37.89
	Augmentation & Normalization	58.45	37.45

6 Conclusion

In this report, we have outlined the normalization [4.2] and augmentation [4.1] methods for the initial dataset. Additionally, we have adjusted the Original ViTHSD dataset based on the methods described and created three new datasets, including Normalization, Normalization & Augmentation, and Augmentation & Normalization. We then applied these to four representative models selected based on the baseline from previous ViTHSD work [5.2] to compare the results based on the evaluation criteria [5.3] provided. We rely on the obtained results to objectively assess the methods applied to the ViTHSD work. From this, we highlight the strengths and limitations of the improvements made to the original dataset.

Data augmentation has certain limitations [5.5] for the Original ViTHSD dataset, as for each target, we only selected comments that belong to that specific target and ignored cases that may belong to multiple targets. However, the statistics for the five targets in the dataset are highly imbalanced, and the number of comments that satisfy the given condition becomes even smaller.

References

[1] Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. "Empirical Study of Text Augmentation on Social Media Text in Vietnamese". In: (2020). arXiv: 2009 . 12319 [cs.CL]. URL: https://arxiv.org/abs/2009.12319.

[2] Thanh-Nhi Nguyen, Thanh-Phong Le, and Kiet Van Nguyen. "ViLexNorm: A Lexical Normalization Corpus for Vietnamese Social Media Text". In: (2024). arXiv: 2401.16403 [cs.CL]. URL: https://arxiv.org/abs/2401.16403.

[3] Cuong Nhat Vo et al. "ViTHSD: Exploiting Hatred by Targets for Hate Speech Detection on Vietnamese Social Media Texts". In: (2024). arXiv: 2404.19252 [cs.CL]. URL: https://arxiv.org/abs/2404.19252.

[4] Jason Wei and Kai Zou. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks". In: (Nov. 2019). Ed. by Kentaro Inui et al., pp. 6382–6388. DOI: 10.18653/v1/D19-1670. URL: https://aclanthology.org/D19-1670.