IMPROVING THE QUALITY OF THE VIHSD DATASET THROUGH PREPROCESSING TECHNIQUES

Cao Dinh Duy Ngoc, Ngo Huynh Truong, Nguyen Thi Mai Lien, Nguyen Thi Thu Huong, and Nguyen Van Kiet

Abstract—The increasing use of social media platforms has led to a rise in hate speech and offensive language targeting other users, which is one of the negative side effects. Negative online content can have harmful effects on individuals using social media, causing issues related to mental health, social isolation, and even physical harm. To address this problem, we aim to improve the performance of the Vietnamese Hate Speech Detection (ViHSD) dataset by fine-tuning it, as proposed by Son T. Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen [1] . After fine-tuning the dataset, we conducted experiments using various models, including machine learning, neural networks, transfer learning, and hybrid models. We obtained more promising results (measured by F1 score) compared to the original task.

Index Terms—Social Media, Vietnamese Hate speech, ViHSD, performance, pre-precessing

1 INTRODUCTION

In recent years, social media has become an integral part of our daily lives. Social media platforms such as Facebook, Youtube, Instagram, and TikTok have grown significantly in Vietnam, creating a large and diverse online community. However, along with this development, hate speech and offensive language have become severe issues that social media users face. This has been proven to be harmful to the well-being of other users (Mohan et al., 2017; Anjum et al., 2018) [2]. Such behavior can sometimes be considered as online bullying, cyber threats, or cyber harassment.

In the context of online hate speech, detecting and preventing such language has become an important task. To address this issue, we rely on the Vietnamese Hate Speech Detection (ViHSD) dataset - a manually labeled dataset for automatically detecting offensive language on social media. This dataset contains over 30,000 comments categorized into three labels: CLEAN, OFFENSIVE, or HATE.

The main objective of this research is to optimize the ViHSD dataset to improve the performance of classifying hate speech on social media. By implementing various data preprocessing methods, we aim to enhance the accuracy of the models.

We evaluate the performance of the fine-tuned ViHSD dataset using various types of models, including machine learning, neural networks, transfer learning, and hybrid models. Comparing the results with the original dataset demonstrates the effectiveness of the applied preprocessing and optimization methods.

This research addresses the issue of hate speech on social media in Vietnam by providing a fine-tuned and improved ViHSD dataset. The results of this research will provide valuable information for researchers and technology developers to build more robust and accurate systems for automatic hate speech detection on social media.

The paper is organized into several main sections: Section 1 introduces the problem, Section 2 discusses related research works, Section 3 discusses the process of fine-tuning the dataset. Section 4 describes the research methods, Section 5 presents the results and evaluation, and finally, the conclusion and future directions are provided. Each section will provide detailed information about fine-tuning the ViHSD dataset, the achieved results, and analyses based on the results.

2 RELATED WORKS

In recent years, detecting inappropriate language on social media has received significant attention, but most studies have focused on widely spoken languages such as English [3] or French [4]. However, in Vietnam, despite having a social media user base of 71 million people, there is still a limited number of research studies in this area [[1], [5]]. The VLSP-HSP dataset used in the HSD Shared Task in the VLSP Campaign 2019: Hate Speech Detection for Social Good [5] does not provide much information about the labeling process and data evaluation. On the other hand, the ViHSD dataset, which was introduced in the paper "A Large-Scale Dataset for Hate Speech Detection on Vietnamese Social Media Texts" [1] consists of 33,400 manually annotated comments and achieved a Macro F1-score of 62.69% when using the BERT model.

There are several prominent Machine Learning models [6] and Deep Neural Networks [7]. These methods showed promising results but required large-scale labeling, a significant challenge for low-resource languages like Vietnamese. Therefore, in recent years, models such as BERT, BERTology, and BERT-based transfer learning have been employed to detect offensive and hateful behavior in the online language [[8], [9]]. The ViHOS dataset, introduced in the paper "ViHOS: Hate Speech Spans Detection for Vietnamese," utilizes strong baselines such as BiLSTM-CRF [10], XLM-R [11], and PhoBERT [12]. The best-performing model, XLM-RLarge, achieved a notable F1-score of 77.70%. Although the results are not yet outstanding, it is evident that these models have been fine-tuned and effectively trained on the language, surpassing traditional deep learning approaches.

3 Dataset

In this study, we utilized comments extracted from the ViHSD dataset [1], also known as The Vietnamese Hate Speech Detection dataset. ViHSD is a prominent dataset in recent Vietnamese language research, consisting of 27,624, 3,514, and 2,262 comments labeled as CLEAN, HATE, and OFFENSIVE, respectively. All comments in the ViHSD dataset are public comments collected from popular social media platforms in Vietnam. Before analysis, we performed several data preprocessing steps: converting to lowercase, Unicode normalization, removing emojis, standardizing Vietnamese diacritics, removing unnecessary characters, and normalizing "Teencode" (a form of writing Vietnamese using numbers and symbols). The purpose of these preprocessing steps was to standardize and clean the data, thereby enhancing the performance and accuracy of subsequent processing tasks. Additionally, we performed word segmentation using the VNCoreNLP tool [12] to suit certain models.

After preprocessing the data, we removed samples with empty "text" and eliminated duplicates. Subsequently, we balanced the data using undersampling, reducing the CLEAN label to match the total of HATE and OFFENSIVE labels. After applying all these operations to the original datasets (train, validation, test), we obtained a total of 11,130 samples, including 5,509 CLEAN labels, 2,177 OFFENSIVE labels, and 3,444 HATE labels. The distribution among the three sets is 7,961; 956; and 2,213 samples for train, val, and test, respectively, with an approximate ratio of 7:1:2.

4 Models

4.1 Logistic Regression

The Logistic Regression [13] model is a machine learning model primarily used for classification problems. The model utilizes the logistic function, also known as the sigmoid function, to map the output to a probability value between 0 and 1. This sigmoid function allows Logistic Regression to handle nonlinear relationships between the input features and the target variable. Logistic Regression can be extended to handle multi-class classification problems using approaches like One-vs-Rest (OvR) or Multinomial Logistic Regression. OvR constructs multiple binary classifiers, treating each class as positive and the rest as negative. Multinomial Logistic Regression directly models the probabilities of each class.

4.2 Support Vector Machine

The Support Vector Machine (SVM) [14] model is a machine learning model primarily used for classification and regression problems. SVM aims to find the best hyperplane to separate data points belonging to different classes in the feature space. The SVM model is trained on a training set to find the optimal hyperplane. The training process of SVM typically involves optimizing an objective function through optimization algorithms such as Sequential Minimal Optimization (SMO) or gradient descent.

4.3 LSTM

LSTM (Long Short-Term Memory) [15] is a Recurrent Neural Network (RNN) type designed to process sequential data with long-term dependencies. The LSTM model emulates the ability to retain long-term memory and learn long-range dependencies within texts.

4.4 BERT

BERT [16] is a natural language processing (NLP) model based on the BERT (Bidirectional Encoder Representations from Transformers) architecture. One notable feature of BERT is its ability to understand the context and meaning of each word in a sentence by considering both the preceding and succeeding parts of that word. This allows BERT to capture subtle nuances and semantic relationships between words accurately. BERT has achieved impressive results in various language tasks, surpassing many previous traditional models. With its strong contextual understanding and rich language representation, BERT plays a crucial role in applications such as text classification, machine translation, information extraction, and many other fields in natural language processing.

4.5 XLM-RoBERTa

XLM-RoBERTa (Cross-lingual Language Model - RoBERTa) [17] is a state-of-the-art language model that provides advanced language representation and understanding across multiple languages. It is an extension of RoBERTa based on the BERT architecture. One notable aspect of XLM-RoBERTa is its language transferability, producing universal language representations and knowledge sharing from one language to another. This helps improve processing performance and leverage learned knowledge from one language to another.

4.6 PhoBERT

PhoBERT [18] is an advanced and powerful language model in the field of natural language processing, trained on Vietnamese data. It is built upon the RoBERTa model and trained on a large amount of Vietnamese data from various sources. This enables the model to effectively understand and work with the Vietnamese language, with the ability to comprehend context and meanings of words in Vietnamese.

4.7 PhoBERT + CNN

PhoBERT-CNN is a natural language processing (NLP) model that combines PhoBERT with a convolutional neural network (CNN) [19]. PhoBERT-CNN uses PhoBERT to embed Vietnamese text into semantic-based vector representations and then applies a CNN to capture contextual and structural features of sentences.

4.8 XLM-RoBERTa + CNN

XLM-RoBERTa-CNN is a natural language processing (NLP) model that combines XLM-RoBERTa with a convolutional neural network (CNN). XLM-RoBERTa-CNN utilizes XLM-RoBERTa to embed Vietnamese text into semantic and contextual vector representations. Then, a CNN is applied to extract contextual and structural features of the sentences.

4.9 CNN + LSTM

The model combines two different neural network architectures: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network. The CNN architecture is used to learn spatial features from the input data. CNN has the ability to extract information from neighboring elements and images, enabling it to detect important patterns and features. After being processed by the CNN, the output is fed into the LSTM network. LSTM is a type of neural network based on Recurrent Neural Networks (RNN), designed to handle sequential data and address the issue of information transmission in long sequences.

5 EXPERIMENTS AND RESULTS

5.1 Experiment settings

After preprocessing the dataset using the steps described in Section 3, we proceeded with word segmentation to enhance compatibility with specific models, following established linguistic conventions. The word-segmented dataset was then utilized in our experimental evaluations, involving a range of models including Logistic Regression, Support Vector Machines (SVM), PhoBERT, PhoBERT + CNN, Long Short-Term Memory (LSTM), and CNN + LSTM.

Furthermore, we conducted additional experiments employing the syllable-level dataset, focusing on models such as XLM-R, mBERT, and XLM-R + CNN. This allowed us to explore the impact of different linguistic units on the performance of the models.

For the experimental setup, we trained the transformer models for 5 epochs, while the LSTM and hybrid models were trained for 20 epochs. To maintain consistency, the sequence length was uniformly set to 100 across all models, while the batch size was adjusted to accommodate the hardware capabilities, ranging from 16, 32, 64, to 128. Additionally, the learning rate was fine-tuned to promote convergence, with values of 1e-04, 5e-05, and 5e-06 being employed for different models based on empirical observations.

5.2 Evaluation metrics

In this problem, we focus on two main evaluation metrics, namely Accuracy and F1-macro. However, during the evaluation process, we place more emphasis on the F1-macro metric. The emphasis on F1-macro is due to the class imbalance within the dataset. When a dataset is imbalanced, evaluating solely based on Accuracy can lead to misunderstandings. Accuracy can be high when the model focuses only on predicting the majority class and disregards the minority class. This does not accurately reflect the model's classification ability.

By emphasizing F1-macro, we ensure that the model performs well across all classes, including the minority classes. This implies that the model has the ability to classify accurately and consistently across different instances, ensuring fairness and reliability in the evaluation process.

5.3 Experiment results

STT	Model	Validation	Preprocessed test	
		Accuracy	Accuracy	F1
1	Logistic Regression	69.67	71.17	63.47
2	SVM	70.29	71.35	63.03
3	**XLM-R**$_{large}$	**75.21**	**76.01**	**71.71**
4	XLM-R$_{base}$	72.07	74.42	69.81
5	PhoBERT$_{large}$	72.91	72.93	67.33
6	PhoBERT$_{base}$	67.78	68.91	59.29
7	mBERT$_{cased}$	69.77	71.62	67.28
8	Phobert$_{base}$ + CNN	72.38	73.84	69.82
9	XLM-R$_{base}$ + CNN	71.92	73.70	69.98
10	CNN + LSTM	68.30	45.05	39.15
11	LSTM	67.61	68.32	57.73

TABLE 1: Experimental results of the models on the preprocessed test set.

Table 1 presents the results of the different models on the test set. The results indicate no significant difference among the models, with the highest performance observed in the transformer models, followed by the hybrid models, neural networks, and basic machine learning models.

5.4 Results analysis

After evaluating the preprocessed test set, we further evaluated the models on the original test set to compare the results with the baseline models in the original paper. For the original test set, we only performed word segmentation on the models mentioned in Section 5.1, and used standard input for the remaining models.

STT	Model	Original Test	
		Accuracy	F1
1	Logistic Regression	82.74	58.11
2	SVM	83.91	58.91
3	**XLM-R-large**	**82.69**	**63.40**
4	XLM-R-base	80.21	60.37
5	PhoBERT-large	81.36	59.01
6	PhoBERT-base	79.97	54.17
7	mBERT-cased	74.31	54.67
12	PhoBERT-base + CNN	79.60	58.93
13	XLM-R-base + CNN	78.25	59.53
10	CNN + LSTM	58.91	34.72
11	LSTM	57.38	34.99

TABLE 2: Results on the Original Dataset for Each Model.

In this comparison, we assess the performance of different models against the highest-performing model from the original paper, mBERT. The results reveal that the majority of models exhibit lower Accuracy and F1 scores compared to the original model, except for the XLM-R-large model, which displays a slightly higher F1 score of 0.71%. This disparity can be attributed to the unsuitability of the input data in the original test set for these models. These models were trained on a dataset that underwent preprocessing steps outlined in Section 3, which may account for the variations in performance.

Additionally, when comparing the results of the models on the original test set and the preprocessed test set, we observe that the results on the preprocessed test set exhibit improved balance in both accuracy and F1 score. The closer results of these two metrics indicate that the dataset has mitigated the imbalance. This enhanced balance is crucial to ensure a fair and reliable evaluation of the models' performance. By minimizing the impact of data heterogeneity, the preprocessed test set provides a more accurate reflection of the models generalization capabilities, enabling a more unbiased and robust comparison and assessment.

With the preprocessed dataset, the highest achieved result is 71.71% with the XLM-R-large model in terms of F1 score, which is 9.02% higher than the highest-performing model on the original dataset.

#	Comment	True	Predict
1	sao không đền bằng rau muống (Why not compensate with water spinach?)	1	0
2	chi em ma chui lon ke cho người ta nghe người ta cuoi hay ho gi (If sisters argue with each other and then tell others about it to make them laugh, what's so great about it?)	0	1
3	đm đen méo chơi nữa (Fuck, things are going fucking bad. I'm not fucking playing anymore)	0	1
4	<person name> vãi loinmm (Damn pussy)	1	2

TABLE 3: Some incorrect predictions on the test set by the XLM-R-large model.

Table 3 highlights some errors observed in the XLM-R-large model. The most notable errors involve semantic or contextual issues in the Vietnamese language. In comment #1, the sentence contains no explicit offensive words, but it is written in a context that mocks or criticizes an individual or organization. This is a common challenge in Vietnamese datasets. In comment #2, there are words written without accents ("lon"), which can be understood as "cãi lộn" (lash out) or "cái lồn" (pussy). These words can confuse even human annotators. Apart from model-related errors, there are also errors resulting from annotators during the labeling process. In comment #3, the sentence contains the word "đm" (fuck), but the annotators assigned it a clean label, leading to confusion for the model. Similarly, in comment #4, the data sample includes pre-words associated with the OFFENSIVE label ("vãi loinmm" - "Daim pussy"), which can cause confusion between the two labels.

6 Conclusion and future work

Our research focused on enhancing the prediction accuracy for comments marked as HATE and OFFENSIVE by fine-tuning the ViHSD dataset. To accomplish this, we conducted extensive experiments employing different models. The XLM-R-large model stood out, attaining the highest F1 score of 71.71%. This achievement demonstrates the effectiveness of our approach in improving the performance of comment classification. By refining the ViHSD dataset and leveraging advanced models like XLM-R-large, we have made significant strides in accurately identifying and categorizing hateful and offensive comments.

Moving forward, we plan to enhance the performance of the dataset by implementing more stringent guidelines and re-evaluating the labeling process to attain a higher level of consensus. We aim to improve the dataset's suitability for offensive comment detection tasks. Moreover, we will explore alternative methodologies to address this issue. One potential avenue is to extract significant phrases or segments from comments classified as HATE or OFFENSIVE, which can enhance the reliability and overall performance of the system. By incorporating these approaches, we anticipate achieving even greater accuracy and effectiveness in identifying and addressing offensive content.

References

[1] Son T Luu, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. A large-scale dataset for hate speech detection on vietnamese social media texts. In *Advances and Trends in Artificial Intelligence. Artificial Intelligence Practices: 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2021, Kuala Lumpur, Malaysia, July 26–29, 2021, Proceedings, Part I 34*, pages 415–426. Springer, 2021.

[2] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. The impact of toxic language on the health of reddit communities. In *Advances in Artificial Intelligence: 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings 30*, pages 51–56. Springer, 2017.

[3] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[4] Delphine Battistelli, Cyril Bruneau, and Valentina Dragos. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365, 2020.

[5] Xuan-Son Vu, Thanh Vu, Mai-Vu Tran, Thanh Le-Cong, and Huyen Nguyen. Hsd shared task in vlsp campaign 2019: Hate speech detection for social good. *arXiv preprint arXiv:2007.06493*, 2020.

[6] Hang Thi-Thuy Do, Huy Duc Huynh, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen, and Anh Gia-Tuan Nguyen. Hate speech detection on vietnamese social media text using the bidirectional-lstm model, 2019.

[7] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[8] Zaki Mustafa Farooqi, Sreyan Ghosh, and Rajiv Ratn Shah. Leveraging transformers for hate speech detection in conversational code-mixed tweets. *arXiv preprint arXiv:2112.09986*, 2021.

[9] Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Vihos: Hate speech spans detection for vietnamese. *arXiv preprint arXiv:2301.10186*, 2023.

[10] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[11] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[12] Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. Vncorenlp: A vietnamese natural language processing toolkit. *arXiv preprint arXiv:1801.01331*, 2018.

[13] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[14] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational*

Linguistics, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[18] Dat Quoc Nguyen and Anh Tuan Nguyen. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online, November 2020. Association for Computational Linguistics.

[19] Keiron O'Shea and Ryan Nash. An introduction to convolutional neural networks, 2015.