Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model

Tin Van Huynh, **Duc-Vu Nguyen**, **Kiet Van Nguyen**, **Ngan Luu-Thuy Nguyen**, and **Anh Gia-Tuan Nguyen**
University of Information Technology, Vietnam National University Ho Chi Minh City
16521827@gm.uit.edu.vn,{vund,kietnv,ngannlt,anhngt}@uit.edu.vn

***Abstract*—In recent years, Hate Speech Detection has become one of the interesting fields in natural language processing or computational linguistics. In this paper, we present the description of our system to solve this problem at the VLSP shared task 2019: Hate Speech Detection on Social Networks with the corpus which contains 20,345 human-labeled comments/posts for training and 5,086 for public-testing. We implement a deep learning method based on the Bi-GRU-LSTM-CNN classifier into this task. Our result in this task is 70.576% of F1-score, ranking the 5th of performance on public-test set.**

***Index Terms*—hate speech detection, Bi-GRU-LTSM-CNN, Vietnamese, Social Media Text**

I. Introduction

Due to the rapid development of the internet, the number of users on social networks have increased significantly. Data generated from wallet social networks also has exponentially grown. Users' commenting or posting are difficult to control. Therefore, a tool that categorizes posts and comments is essential. This is a primary major that VLSP Shared Task 2019 open with the first task - **Hate Speech Detection on Social Networks** with the purpose of detecting Vietnamese social media text according to predefined labels.

Recently, **Hate Speech Detection** has been studied by researchers in the field of natural language processing through the Shared Task SemEval 2019 - Task 5: Multilingual detection of hate speech against immigrants and women in Twitter [1] which mainly solve problems of predicting hate speech on social media in English and Spanish. In addition, Zhang and Luo evaluated on the largest collection of hate speech datasets based on Twitter [2] on deep neural models.

In this task, we focus on a solution for predicting hate speech on Vietnamese which is a low-resource language for natural language processing. In particular, we have implemented deep learning to classify comments or posts on social networks. The problem is stated as:

- **Input**: Given a Vietnamese post/comment on social network.
- **Output**: One of three labels (HATE, OFFENSIVE, or CLEAN) which is predicted by our system.

Table I shows several examples for this task.

In this paper, our contributions in this paper are presented as follows.

- Firstly, we implemented three different models based on neural networks such as TextCNN, Bi-GRU-CNN and Bi-GRU-LSTM-CNN to solve the VLSP Shared Task: Hate Speech Detection on Vietnamese social media text.
- Secondly, we achieved the best result on this task accounting for 70.576% on the public test, ranking the 5th in the Hate Speech Detection task on social networks.

The organization of the paper is as follows: in the section 2 we will discuss related works on the topic and related models, the third section we will talk about the data set we have, the section 4 and 5 is pre-processing and the proposed method, the section 6 is our experiment and the section 7 is the conclusion and the future work.

II. Related work

Deep neural network models have been widely used to improve performance of a different natural language processing (NLP) tasks. [3] have demonstrated the effectiveness of combining pre-processing and the CNN-GRU network where the network consists of an word embedding layer, CNN-1D, 1D max-pooling, GRU, global max pooling and a softmax layer. Zhang et al. have empirically illustrated that CNN perfectly works in classification of text [4]. RNNs shown in [5] and Bi-LSTMs shown in [6] also give better performance in text classification. Besides, there are many traditional machine learning [7] such as Random Forests, SVM, Gradient Boosted Decision Trees, Logistic Regression and Deep Neural Networks with the well-known word embedding Glove [8] were used to recognize hate speech in Tweets. Besides, we also take some other combination models for classification, for example, Bi-RNN [9], Bidirectional-GRU [10], Bidirectional-LSTM [11], Bi-LSTM-CNN [12] and Bi-LSTM-CRF [13]. Facebook Artificial Intelligence Research (FAIR) developed a pre-trained word embedding which is very good to text-classification model involving out-of-vocabulary words [14].

III. Dataset

We use a dataset which the VLSP Shared Task 2019 provide, containing posts or comments from the social network Facebook which are annotated with three different

Table I
SEVERAL EXAMPLES FOR VIETNAMESE HATE SPEECH

No.	Comment/Post	Label
1	Thương tụi mày quá không biết tụi mày có thương tao ko :(	Clean(0)
2	Thi đấu thể thao chuyên nghiệp ở trong nước bạc bẽo vl	Offensive(1)
3	Không ai rãnh mà nói chuyện với mày đâu thằng ngũ	Hate(2)

classes (Hate, Offensive and Clean).

- **HATE** (Hate Speech): a comment or post is identified as hate speech if it (1) targets individuals or groups on the basis of their characteristics; (2) demonstrates a clear intention to incite harm, or to promote hatred; (3) may or may not use offensive or profane words. For example: “Assimilate? No they all need to go back to their own countries. #BanMuslims Sorry if someone disagrees too bad.". See the definition of Zhang et al. [2]. In contrast, “All you perverts (other than me) who posted today, needs to leave the O Board. Dfasdfdasfadfs" is an example of abusive language, which often bears the purpose of insulting individuals or groups, and can include hate speech, derogatory and offensive language.

- **OFFENSIVE** (Offensive but not hate speech): a post or comment may contain offensive words but it does not target individuals or groups on the basis of their characteristics. For instance, “WTF, tomorrow is Monday already."

- **CLEAN** (Neither offensive nor hate speech): normal comments or posts on social networks, it does not contain offensive or hate speech. For example, “She learned how to paint very hard when she was young".

Table II
STATISTICS OF THE DATASET LABELS

	Clean(0)	Offensive(1)	Hate(2)
Frequency	18,614	1,022	709
Percentage	91.49	5.02	3.49

IV. TEXT PRE-PROCESSING

We use several simple techniques in text pre-processing in all models for this task as follows.

- Converting all words to lower case.
- Removing extra white spaces, punctuation marks.
- Replacing all numbers with "number".
- Word tokenization using the pivy library [15].

V. BI-GRU-LSTM-CNN MODEL FOR VIETNAMESE HATE SPEECH DETECTION

In this section, we propose a deep neural model for the prediction of hate speech on social media text. Figure 1 shows the architecture of our network. The basic architecture in this paper is Convolutional Neural Network (CNN) with 1D convolutions.In addition, we also study about two other deep neural models which are Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). The details of all these neural models are presented in next sub-sections. In this model, there are several common parts:

- **Word embedding layer**: The input is a matrix of 220x300 dimensions. In particular, each sentence has only 220 words and each word is represented by a 300 dimensional word embedding. Pre-training word level vector already is a kind of word representations for deep neural network models since Word2Vec [16]. In our experiments, we choose FastText [17] as our pre-training model.

- **CNN-1D layer**: We use a 1D spatial drop out with 0.2 dropout rate. It can prevent the model from over-fitting and to get better generalizations.

- **Bidirectional LSTM**: The model uses two parallel blocks of Bidirectional Long Short Term Memory (Bi-LSTM) where the term Bidirectional is that the input sequence is given to the LSTM in two different ways. LSTM is a variation of a recurrent neural network that has an input gate, an output gate, a forget gate and a cell. In our experiment, we used two parallel bidirectional LSTM blocks having 112 units for each. We used sigmoid and tanh for recurrent activations and hidden units respectively.

- **Bidirectional GRU**: Different from LSTMs, gated recurrent units (GRU) is without output gate. [18] introduced firstly in 2014. In addition, GRUs have an update gate and a reset gate which is responsible of combining new input with the previous one. Finally, the update gate is responsible of how much the previous memory is required to be saved.

VI. EXPERIMENTAL RESULTS

In this section, we describe our experiments and results for the task. Evaluation for this task is based on a metric of the F1-score. We show the results of our experiments for Vietnamese hate speech detection task in Table III. In particular, the Bi-GRU-LSTM-CNN achieved the best performance among three different models that we tried to conduct experiments.

Table IV shows the 5th rank of performance on the public-test set. As a result, our rank in this task is the 5th with

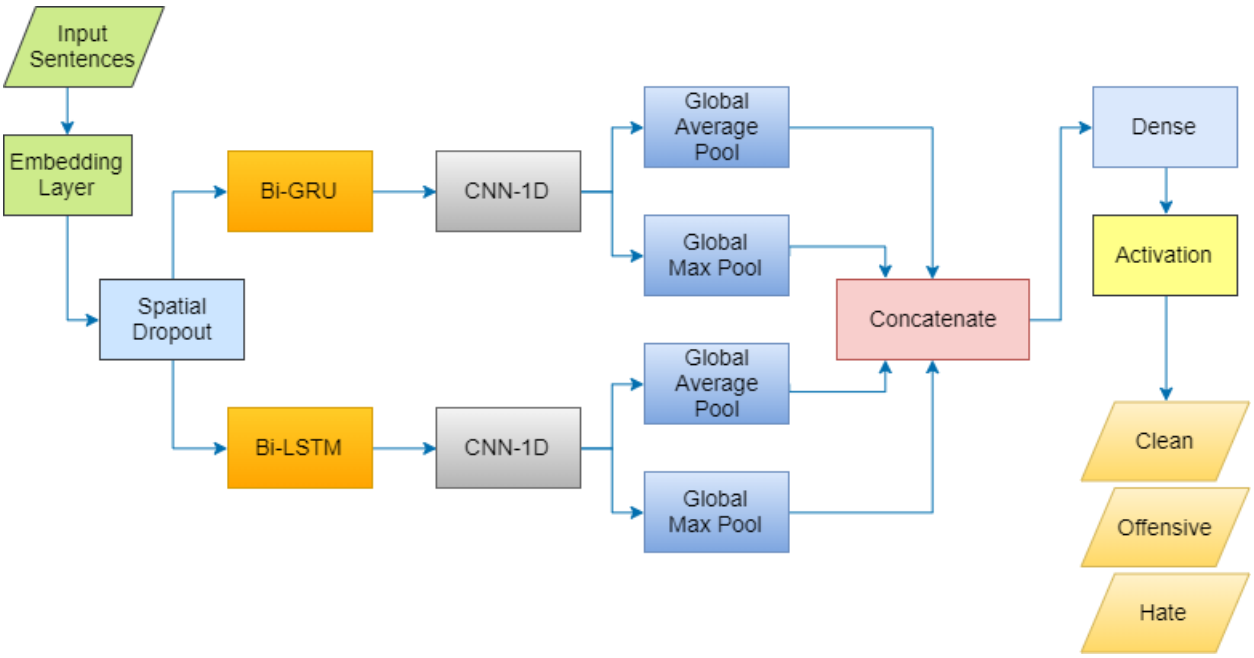

Figure 1. The Bi-GRU-LSTM-CNN architecture for Hate Speech Detection

Table III
F1-SCORES OF OUR EXPERIMENTS ON THIS TASK

Model	F1-score
TextCNN	56.512
Bi-GRU-CNN	69.293
Bi-GRU-LSTM-CNN	70.576

70.576% of F1-score. The results were not significantly different from other teams on the public-test set. However, our results only ranked the 11th on the private-test set.

Table IV
RESULTS OF THE TOP 5 ON PUBLIC-TEST SET

Rank	Team	F1-score
1	Try hard	73.019
2	HH_UIT	71.432
3	titanic	70.747
4	ABCD	70.582
5	**HUYNH TIN**	**70.576**

VII. CONCLUSION AND FUTURE WORK

In this paper, we have described our approach to solve the hate speech detection task proposed at the VLSP Shared Task 2019. We develop the system using supervised approach for classifying three different labels. We participate in this and evaluate the performance of our system on this dataset. Our official result is 70.576% of F1-score, ranking the 5th of the scoreboard on the public-test set.

In the future work, we plan to address this problem in different ways to improve the performance. We will investigate directions both in traditional machine learning and types of deep neural network models for this problem. In addition, we also analyze experimental results on this task to select the efficient approach such as the hybrid approach which combines supervised method and rule heuristic to improve the result of detecting hate speech on social media text.

ACKNOWLEDGMENT

We would like to thank the VLSP Shared Task 2019 organizers for their really hard work and providing the Vietnamese Hate Speech Detection dataset for our experiments.

REFERENCES

[1] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., Sanguinetti, M.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019). Association for Computational Linguistics.

[2] Zhang, Z., Luo, L.: Hate speech detection: A solved problem? the challenging case of long tail on twitter. CoRR abs/1803.03662 (2018), http://arxiv.org/abs/1803.03662

[3] Z. Zhang, D. Robinson, and J. Tepper, "Detecting hate speech on twitter using a convolution-gru based deep neural network," in European Semantic Web Conference. Springer, 2018, pp. 745–760.

[4] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding Convolutional Neural Networks for Text Classification. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 56–65. Association for Computational Linguistics.

[5] Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. arXiv preprint arXiv:1605.05101.

[6] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional lstm with two-dimensional max pooling," arXiv preprint arXiv:1611.06639, 2016.

[7] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma, "Deep learning for hate speech detection in tweets," in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 759–760.

[8] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of EMNLP-2014, pages 1532–1543, Doha, Qatar, October.
[9] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. Trans. Sig. Proc., 45(11):2673–2681, November 1997.
[10] Zhiyao Duan Rui Lu. Bidirectional GRU for sound event detection. 2017.
[11] Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. 2016.
[12] Jason P. C. Chiu and Eric Nichols. Named entity recognition with bidirectional lstm-cnns. 2015. cite arxiv:1511.08308.
[13] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. 08 2015.
[14] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," arXiv preprint arXiv:1607.04606, 2016.
[15] Pyvi library, link: https://pypi.org/project/pyvi.
[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. CoRR, abs/1301.3781, 2013.
[17] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. LREC.
[18] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.