Automated Hate Speech Detection on Vietnamese Social Networks

Quang Pham Huu, Son Nguyen Trung, Hoang Anh Pham
R&D Lab, Sun* Inc
{pham.huu.quang, nguyen.trung.son, pham.hoang.anh}@sun-asterisk.com

***Abstract*—Nowadays, the internet plays an important role in our everyday life. It provides us useful information, knowledge, news, and a free space to share and exchange our personal opinions with other people all over the world through some platforms such as the social network. While there are various advantages of social network, its freedom sometimes bring to us a lot of trouble. Some people use the social network for some immoral aims such as harass, racist, and offend others which must be detected and removed immediately. With the rapid development of the social network, number of content uploaded on it is dramatically enormous and becoming larger which can not control effectively by human. We proposed a novel method for solving this problem by a multi-class classification model to classify content into 3 labels: HATE, OFFENSIVE, and CLEAN. With the Vietnamese dataset of the competition VLSP-SHARED Task, our experimental results have the first position on the contest table.**

***Index Terms*—Vietnamese Hate Speech Detection, Natural Language Processing, Text Mining, Supervised Learning, Social Networks.**

I. Introduction

Recently, Hate speech is becoming a popular problem in social network, which is a real challenge for website's administrators. However, improving quality of contents and making a friendly environment for users is one of the most important purposes of social network, which will increase their number of users significantly, and bring a good profit to their organizations. Furthermore, blocking and punishing harmful behaviors on the internet are also necessary actions of every government and authorities to protect the safety of their countries and regions. Many successful platforms such as Facebook, Twitter, and Instagram are showing their efforts to solve this issue with many policies, technologies implemented, but until now, they have not come up with a complete idea yet.

The challenge is it is hard to determine if the content is a violation or not without the whole information of context such as other related contents or stories. In the case of Vietnamese, that burning issue is much more difficult since the diverse vocabulary and complex grammar of our language. Furthermore, noise and shortage of the imbalance dataset provided by the contest also make this task become more tricky.

In this paper, we introduced effective steps of text prepossessing combine with traditional machine learning techniques to addressing this problem. Our experimental reached 0.6675 and 0.6197 F1-score on the test data and private test data respectively, which help us on leading the ranking table of VLSP-Shared task. Our notable method on this research are:

- The text preprocessing stage that can use to improve considerably our method on the a noise Vietnamese dataset.
- The complete pipeline for hate speech detection task on Vietnamese content.

In the following sections, we will provide related background work, methods of implementations, results and analysis of findings.

II. Related work

As we mentioned above, hate speech detection is a problem incurred from the dramatically rise of the number content on internet which gradually become out of controlling ability of human. It has been a actively research field from late 2010s when platforms such as social network, forum, blogs became popular with people. The earliest approaches used a Logistic Regression Model to address the short messages classification task introduced by Waseem and Hovy (2016) [4], and Davidson $et.\ al.$ (2017) [1]. While Waseem and Hovy proposed an method based on character-level features, word-level features, part of speech, and some other related meta-data of messages are used on Davidson $et.\ al.$ (2017)'s research.

With the successful of Neural Network at contests recent years in many tasks of computer vision, natural language processing, and recommendation. Some state-of-the-art methods also implemented a Deep Neural Network into Hate Speech Detection task. Convolution Neural Networks (CNNs) suggested by Park and Fung (2017) [2] on binary classification task reach some successful with word and charecter level embedding. The classification for type of hate speech is still solved by a Logistic Regression classifier used n-gram features. Zhang, Robinson, and Tepper (2018) [5] proposed a pre-train word embedding by Gated Recurrent Units (GRUs) instead of directly training on this task for more understanding in language of our model.

The state-of-the-art approach of Rizoiu $et.\ al.$ (2019) [3] improved the idea of using pre-train model by using Bidirectional-Long Short Term Memory (Bi-LSTM) model to training carefully the word embedding model only on the hate speech domain. After that, transfer learning step is applied on three other various datasets which help model to learn faster and improve the accuracy on this task significantly.

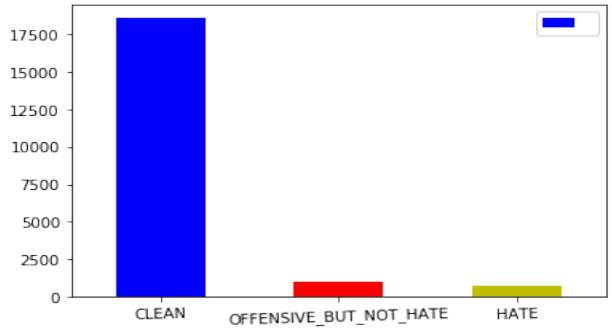

Figure 1. Data distribution

III. Proposed Method

A. Dataset

Dataset used is downloaded from the VLSP competition crawled data from posts and/or comments on Facebook, annotated by many annotators. The dataset includes 25431 items with three classes (HATE, OFFENSIVE, CLEAN) whose meaning are explained below.

- *Hate speech (HATE)*: Messages contained racist, insult about other person or organization, targets individuals or groups on the basis of their characteristics. The messages demonstrates a clear intention to incite harm, or to promote hatred also listed into this class. The contents of these messages may not contain offensive or profane words but it is counted as Hate Speech if their purpose are trying to attack directly other people. For instance, the sentence "Tôi không thích bọn da đen" ("I do not like a negro") in dataset is labeled as HATE.
- *Offensive speech (OFFENSIVE)*: Messages contain offensive words but for emphasizing and joking purpose are counted as OFFENSIVE.
- *Neither offensive nor hate speech (CLEAN)*: Messages do not include any offensive language, hate speech or do not have the harmful content are CLEAN messages.

Our visualization on dataset (Figure 1) illustrates how much data in each class in the dataset published by the VLSP's organizers. In general, there is an extremely enormous gap between the number of CLEAN speeches and the 2 other categories which tends to lead to very bad result for some methods especially Deep Learning approaches. This observation makes us to focus much more on text preprocessing stages and applying classical machine learning methods to solve this dataset to avoid overfiting phenomenon.

Since the dataset was crawled directly from users' activities on social network, it has some notable properties:

- 18614 items clean, 1022 offensive, 709 hate.
- There are a lot of abbreviations and emoji mixed in almost all messages.
- The numbers and characters are written together with no space between them in short message.
- Special charecters such as: #,$,%,& exist in messages with a high frequency.
- Messages contain many foreign language word.
- Break mark such as dot or semicolon are not used with clearly and consistently purpose.
- Link and Mobile Number are hidden because of user's privacy.

B. Data preprocessing

With these properties outlined above, we proposed an effective technical solution for text preprocessing stage works as follows:

- Unicode normalization to standard format (NFC). Normal form C (NFC) first applies a canonical decomposition, then composes pre-combined characters again.
- Converts all text to lowercase.
- Abbreviations are replaced manually by their origin words in each message. Words with different typing but have the same meaning are replaced with one unique representation called normalization step. For instance: all words such as Hà-Nội, HNoi, Hanoi are replaced by Hà Nội. We have built a dictionary of abbreviation which is the most important factor of pre-processing stage because data taken from social networks often contain many strange words such as teen-code which are not written in a regular rule. This dictionary contains over 200 words taken both from train and test data.
- Remove special characters such as b"xefxb8x8f", "u200d", etc.
- Convert email addresses to MAIL label, number to NUMBER label.
- Normalize the cases of repeated characters such as "ghetttt quaaaa diii".
- Convert all emoji characters to the unique EMOJI label.
- Remove all mixed words and numbers, such as J2prime, 200gram, etc.
- Remove all punctuation existed in messages.

It is also notable that, we do not use a separate model for tokenizing Vietnamese words. After several experimental, we realized that method does not work well on social network context. So, for competing in this contest, we try to tokenize sentence to words merely by spaces.

C. Feature Extraction

In this stage, we have experimented a lot of algorithms such as TF-IDF, but finnaly Bag-of-word (BOW) is chosen based on its greatest result. We used BOW with max features = 35000 and n-gram in the range between = (1, 5).

D. Classification Method

As we listed in the Related Work session, recently, the hate speech detection problem often be solved by supervised machine learning method. Therefore, we approached the problem through current popular supervised algorithms with the aim is trying to find the best algorithm fit for this year problem. The algorithms used are listed below:

- Logistic Regression
- Support-vector machine

Table I
THE RESULT OF OUR METHOD IN TRAINING TIME

Model	Validation score (%)	Model	Validation score (%)
1	68.0	11	69.8
2	68.1	12	68.0
3	67.8	13	67.8
4	67.6	14	69.3
5	67.2	15	69.4
6	67.4	16	71.2
7	70.2	17	69.8
8	67.5	18	68.0
9	69.3	19	64.4
10	70.4	20	65.9

- Neural Network
- Nearest Neighbors
- Decision Tree
- Naive Bayes
- Random Forest

We conducted many experiments and evaluations to find the most efficient algorithm.

IV. Experiments

A. Training

Through many experimental, the Logistic Regression algorithm is used as the main training algorithm. We use Grid search and k-fold cross validation with k = 5 to find the best hyper-parameter for the algorithm. Finally, the parameter C = 10, l2 penalties are used.

B. Evaluation Metrics and Results

Macro F1-score is used as the evaluation metric for this shared-task. The F1 score is a weighted average of the precision and recall. Precision is the number of correct positive results divided by the number of all positive results returned by the classifier, and recall is the number of correct positive results divided by the number of all samples that should have been identified as positive.

$$F1 = 2\frac{1}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}} = 2\frac{\text{precion} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

F1-macro is calculated by calculating the F1-score for each label and finding their unweighted mean.

$$Macro\ F1 = \frac{F1_{class0} + F1_{class1} + F1_{class2}}{3} \quad (2)$$

We used stratified shuffle split for split dataset into train and test data with 90% and 10% of total amount in orderly. This split process is repeated 20 times to obtain 20 training datasets and 20 validation datasets, respectively. The results of 20 training times are presented in Table 1.

Finally, we ensemble predictions from these 20 models to produce the final result. The final results is presented in Table 2.

Table II
THE EVALUATION RESULT OF OUR METHOD

	Public test (%)	**Private test (%)**
Macro F1	67.76	61.97

V. Conclusion

In this paper, we describe and proprose our approach to solve the Vietnamese hate speech detection task in the evaluation campaign of VLSP 2019. For future work, we would like to use some deep learning approaches for this task which includes the solution for the imbalance data to improve the performance.

Acknowledgement

We would like to thank our colleagues at ***Sun Asterisk Inc*** for their helpful advice and expertise. We also want to thank the VLSP organizers for their appreciate hard work in maintaining the conference every single year and providing dataset for this evaluation contest.

References

[1] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017.

[2] Ji Ho Park and Pascale Fung. One-step and two-step classification for abusive language detection on twitter. *CoRR*, abs/1706.01206, 2017.

[3] Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. Transfer learning for hate speech detection in social media. *CoRR*, abs/1906.03829, 2019.

[4] 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, San Diego, California, June 2016. Association for Computational Linguistics.

[5] Ziqi Zhang, D. Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. 03 2018.