Empirical Study of Text Augmentation on Social Media Text in Vietnamese

Son T. Luu
University of Information Technology
VNU-HCM, Vietnam
sonlt@uit.edu.vn

Kiet Van Nguyen
University of Information Technology
VNU-HCM, Vietnam
kietnv@uit.edu.vn

Ngan Luu-Thuy Nguyen
University of Information Technology
VNU-HCM, Vietnam
ngannlt@uit.edu.vn

Abstract

In the text classification problem, the imbalance of labels in datasets affect the performance of the text-classification models. Practically, the data about user comments on social networking sites not altogether appeared - the administrators often only allow positive comments and hide negative comments. Thus, when collecting the data about user comments on the social network, the data is usually skewed about one label, which leads the dataset to become imbalanced and deteriorate the model's ability. The data augmentation techniques are applied to solve the imbalance problem between classes of the dataset, increasing the prediction model's accuracy. In this paper, we performed augmentation techniques on the VLSP2019 Hate Speech Detection on Vietnamese social texts and the UIT - VSFC: Vietnamese Students' Feedback Corpus for Sentiment Analysis. The result of augmentation increases by about 1.5% in the F1-macro score on both corpora.

1 Introduction

In recent years, the growth of hate speech has become a crime, not only face-to-face action but also online communication (Fortuna and Nunes, 2018). The development of social network nowadays had made this situation worse. The threading of harassment comments and harassment speech makes the user stop expressing their opinions and looking up for other ideas (Vu et al., 2019). Fortuna and Nunes (2018) defined hate speech as the language attacking, diminishing, and inciting violence or hate against individuals or groups based on their characteristics, religion, nations, and genders. To solve this problem, many datasets are constructed to detect and classify user comments on social network sites such as Facebook, Twitter, and Facebook in many languages[1].

The HSD-VLSP dataset (Vu et al., 2019) provided by the VLSP 2019 Shared task about Hate speech detection on Social network[2] contained nearly 25,000 comments and posts of Vietnamese Facebook users and has three labels. However, the distribution of three classes in the dataset is imbalanced. Besides, the UIT-VSFC dataset (Nguyen et al., 2018) that was used for predicting the feedback from students contained about 16,000 sentences and was annotated for two different tasks: sentiment analysis and topic analysis. Same as the HSD-VLSP dataset, the distribution of labels on the UIT-VSFC dataset is also imbalanced. We use the data augmentation techniques to generate new comments that belong to minority classes from the original dataset to tackle those restrictions. We conduct experiments on the augmented dataset and compare it with the original dataset to indicate data augmentation effectiveness. Those augmentation techniques include synonym replacement, random insertion, random swapping, and random deletion (Wei and Zou, 2019).

[1] http://hatespeechdata.com/
[2] https://www.aivivn.com/contests/8

The rests of the paper are structures as below. Section 2 introduces recent works in hate speech detection. Section 3 gives an overview of two datasets include the HSD-VLSP dataset and the UIT-VSFC dataset. Section 4 presents the methods and models used in our paper. Section 5 shows our experiment results when applied to the text augmentation techniques. Section 6 concludes the paper.

2 Related works

Duyen et al. (2014) conducted an empirical study about the sentiment analysis for Vietnamese texts based on machine learning to study the influences on the models' accuracy. However, besides the impact of the model's ability, and the feature selection such as word-based, syllable-based, and extracting essential words, the imbalance in the dataset also affects the result. The imbalance in label distribution happens regularly (Ali et al., 2015) when one class seems to be more interested than the other. For example, in social media networks, the abusive and hateful comments are often hidden by the users or administrators, since the clean comments take the majority part. The VLSP2019 hate speech dataset (Vu et al., 2019) and the UIT-VSFC dataset (Nguyen et al., 2018) also suffer the imbalance in class distribution. The detail of those datasets is showed in Section 4.

Wang and Yang (2015) provided a novel method for enhancing the data used for behavior analysis using social media texts on Twitter. Their approaches include using the lexical embedding and frame-semantic embedding. The obtained results showed that using the data augmentation brings significantly better results than no data augmentation (using Google New Lexical embedding brings 6.1% improvement in F1-score and using additional frame-semantic embedding from Twitter brings 3.8% improvement in F1-score.

Ibrahim et al. (2018) presented different data augmentation techniques for solving the imbalance problem in the Wikipedia dataset and an ensemble method used for the training model. The result achieved a 0.828 F1-score for toxic and nontoxic classification, and 0.872 for toxicity types prediction.

Rizos et al. (2019) introduced data augmentation techniques for hate speech classification. The authors 's proposed methods increased the result of hate speech classification to 5.7% in F1-macro score.

Finally, Wei and Zou (2019) provided EDA (Easy Data Augmentation) techniques used to enhance data and boost performance on the text classification task. It contains four operations: synonym replacement, random insertion, random swap, and random deletion. In this paper, these operations are applied to the HSD-VLSP 2019 dataset and the UIT-VSFC dataset to increase the classification models' ability.

3 Datasets

3.1 The HSD-VLSP dataset

The hate speech dataset was provided by the VLSP 2019 shared task about hate speech detection for social good (Vu et al., 2019). The dataset contains a total of 20,345 comments and posts crawled from Facebook. Each comment is labeled by one of three labels: CLEAN, OFFENSIVE, and HATE. Table 1 showed the overview information about the dataset.

	Num. comments	**Avg. word length**	**Vocab. size**
CLEAN	18,614	18.69	347,949
OFFENSIVE	1,022	9.35	9,556
HATE	709	20.46	14,513
Total	20,345	18.28	372,018

Table 1: Overview of the HSD-VLSP dataset

According to Table 1, the number of CLEAN comments take a majority part in the dataset, the number of OFFENSIVE comments and HATE comments are much fewer. Thus, the distribution of labels in the dataset is imbalanced.

3.2 The UIT-VSFC dataset

The Vietnamese Students' Feedback Corpus for Sentiment Analysis (UIT-VSFC) by Nguyen et al. (2018) are used to improve the quality of education. The dataset contains nearly 11,000 sentences and consists of two tasks: sentiment-based classification and topic-based classification. The sentiment-based task comprises three labels: positive, negative, and neutral. The topic-based task comprises four labels corresponding to lecturer, training program, facility, and others. Table 2 describes the overview about the UIT-VSFC training set.

	Num. comments	**Avg. word length**	**Vocab. size**
Total	11,426	10.2	117,295
Sentiment based task			
Positive	5,643	8.2	46,807
Negative	5,325	12.6	67,193
Neutral	458	7.1	3,295
Topic based task			
Lecturer	8,166	9.7	79,854
Training program	2,201	12.2	27,039
Facility	497	12.3	6,130
Others	562	10.9	4,272

Table 2: Overview of the UIT-VSFC training set

According to Table 2, the number of data in the neutral label is lower than positive and negative on the sentiment-based task. So is the topic-based task when the *facility* and *others* labels are much lower than the two remain labels. In brief, the imbalance data happened on the neutral label for the sentiment-based task, and the *facility* and the *other* labels for the topic-based task.

4 Our proposed method

4.1 The augmentation techniques

In this paper, we implement the EDA techniques introduced by Wei and Zou (2019). Those techniques will get a sentence as input and perform one of these following operations to generate new comments:

- **Synonym replacement (SR)**: This operation creates a new sentence by randomly choosing n words from the input sentence and replaces them by their synonyms, excluding the stop words. In our experiments, we use the Vietnamese wordnet[3] from Nguyen et al. (2016) for synonym replacement and the Vietnamese stopword dictionary[4] for removing stop words in the sentence.

- **Random Insertion (RI)**: This operation generates new data by first finding a random word in the input sentence, which is not a stop word, then taking its synonym and putting it into the sentence's random position. The synonyms are taken from the Vietnamese wordnet[3].

- **Random Swap (RS)**: This operation makes a new sentence by choosing two random words in the input sentence and swap their position.

- **Random Deletion (RD)**: This operation creates a new sentence by accidentally deleting p words in the sentence (p is the probability defined before by the user).

According to Wei and Zou (2019), n indicates the number of changed words for SR, RI, and RS methods, which calculated as $n = \alpha * l$, where α is the percentage of replacement word in the sentence and l is the length of the sentence. For the RD method, the probability of deletion words p equal to α. The α is defined by the user.

Table 3 shows examples of data between original and after augmented by EDA techniques in the HSD-VLSP dataset.

[3] https://github.com/zeloru/vietnamese-wordnet

[4] https://github.com/stopwords/vietnamese-stopwords

Comments	Type
Original: con này xấu trai vl *(this guy is f*cking ugly)* **Augmented**: con xấu trai vl	RD
Original: Đcm nản vl *(This is f*cking bored)* **Augmented**: Đcm nhụt chí vl	SR
Original: Đume đau răng vl *(Toothache got damn hurt!)* **Augmented**: Đume răng đau vl	RS
Original: Đm Lắm chuyện vl *(F*uck those curious guys)* **Augmented**: Đm thứ Lắm chuyện vl	RI

Table 3: Several example of the augmented data on the HSD-VLSP dataset

4.2 The classification model

Aggarwal and Zhai (2012) defined the text classification problem as a set of training data $D = \{X_1, ..., X_N\}$, in which each record is labeled with a class value drawn from a set of discrete classes indexed by $\{1..k\}$. The training data used to construct a classification model. With a given test dataset, the classification model is used to predict a class for each instance in the test dataset. Our paper used the Text-CNN model (Kim, 2014) for the HSD-VLSP dataset and the Maximum Entropy model (Nigam et al., 1999) for the UIT-VSFC dataset to study the effectiveness of data augmentation on those two datasets. In practice, the idea of Logistic Regression is maximizing the cross-entropy loss of the actual label in the training dataset (Jurasky and Martin, 2000), which is the same as the Maximum Entropy model (Nigam et al., 1999). Thus, we use the term Maximum Entropy instead of Logistic Regression in our results.

5 Empirical results

5.1 Experiment configuration

For the HSD-VLSP corpus, we use cross-validation with five folds for the Text-CNN model and the Maximum Entropy model. Following the same manner in the previous study (Luu et al., 2020), for each fold, we keep the test set and enhance the training set with EDA techniques.

For the UIT-VSFC dataset, we used the data divided into the training, development, and test sets by Nguyen et al. (2018). Then we run the EDA techniques on the training set and use the test set to evaluate the result.

5.2 Data augmentation result

We first applied the EDA techniques on the entire original HSD-VLSP dataset to enhance the data on HATE and OFFENSIVE labels. Table 4 describes the information about the HSD-VLSP dataset after making data augmentation.

	Num. comments	Avg. word length	Vocab. size
CLEAN	18,614	19.3	360,958
OFFENSIVE	13,823	11.3	157,517
HATE	11,051	23.6	260,841
Total	43,488	17.9	779,316

Table 4: The augmented HSD-VLSP courpus

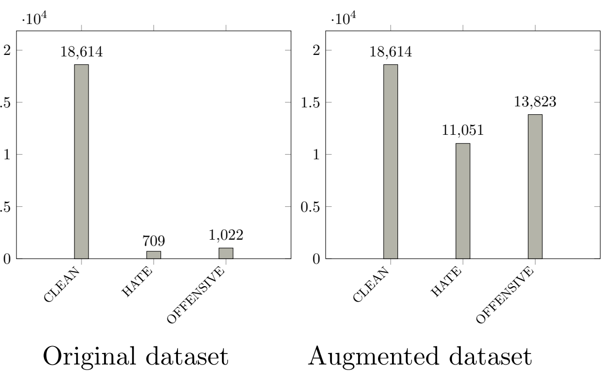


Figure 1: Number of comments on before and after augmentation in the HSD-VLSP dataset

It can be inferred from Table 4 that after applying the EDA techniques, the number of data and the vocabulary size on the HATE and OFFENSIVE labels and increased significantly (Words calculate the vocabulary size, and we use the pyvi[5] for tokenizing). Figure 1 illustrates the distribution of three classes before and after us-

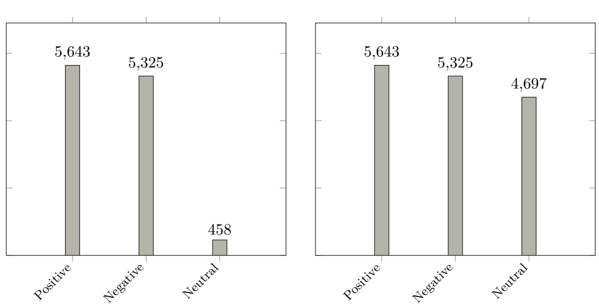
[5] `https://pypi.org/project/pyvi/`

ing data augmentation techniques on the HSD-VLSP dataset. After using EDA techniques, the data on three labels are well-distributed.

Besides, we apply the EDA on the UIT-VSFC training set to enhance the data on the neutral label for the sentiment-based task, and on *facility* and *other* labels for the topic-based task. Table 5 describes the UIT-VSFC training set after enhanced. Comparing with the original dataset as described in Table 2, the number of comments and the vocabulary size of the neutral label on the sentiment-based task increased significantly. Same as the sentiment-based task, the number of comments and vocabulary size of the *facility* and *other* labels are also dramatically increased.

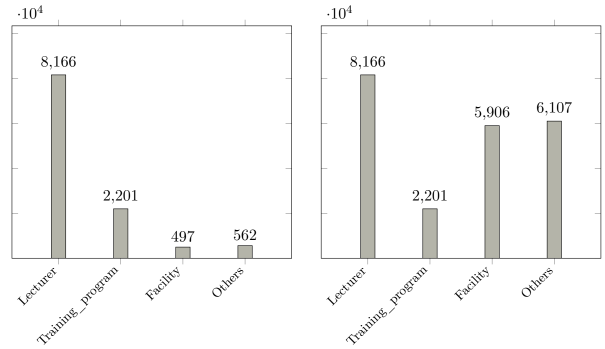
	Num. comments	**Avg. word length**	**Vocab. size**
Sentiment-based task			
Positive	5,643	8.2	46,807
Negative	5,325	12.6	67,193
Neutral	4,697	8.1	38,349
Total	15,665	9.7	152,349
Topic-based task			
Lecturer	8,166	9.7	79,854
Training program	2,201	12.2	27,039
Facility	5,906	13.7	81,299
Others	6,107	13.3	54,722
Total	22,380	10.8	242,914

Table 5: The augmented UIT-VSFC training set



Original training set Augmented training set

Figure 2: The distribution of the sentiment-based task's labels of the UIT-VSFC dataset before and after enhanced



Original training set Augmented training set

Figure 3: The distribution of the topic-based task's labels of the UIT-VSFC dataset before and after enhanced

Figure 2 and Figure 3 illustrate the UIT-VSFC training dataset before and after enhanced data on sentiment based task and topic based task respectively. For the two tasks, after augmentation the distribution of data between labels are balanced.

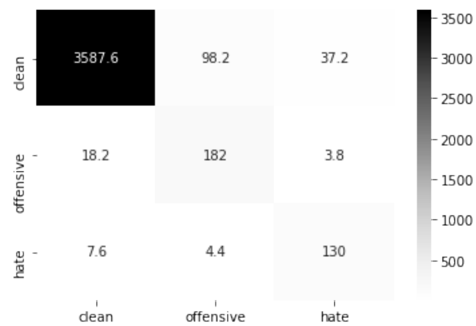
5.3 Model performance results

We implement the Text-CNN model on the entire original HSD-VLSP dataset and the augmented HSD-VLSP dataset. Table 6 shows the result by F1-macro score. Comparing with the original results (Luu et al., 2020), the accuracy of the HSD-VSLP dataset after using augmented techniques are higher than the original dataset. According to Figure 4, the number of right prediction on the *offensive* and the *hate* labels are increased.

Methodology	**F1-macro (%)**
Text-CNN (original) (Luu et al., 2020)	83.04
Text-CNN (augmented)	**84.80**
Maximum Entropy (original) (Luu et al., 2020)	64.58
Maximum Entropy (augmented)	**75.27**

Table 6: Empirical result by the Text-CNN model on the HSD-VLSP dataset

The original dataset



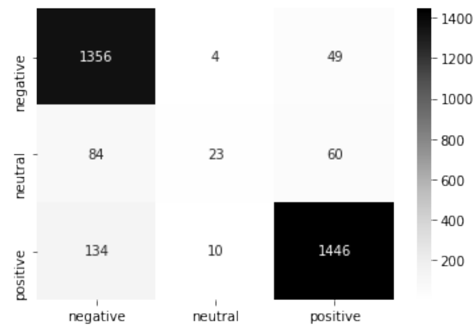
The augmented dataset

Figure 4: Confusion matrix of Text-CNN model before and after enhanced data on the HSD-VLSP

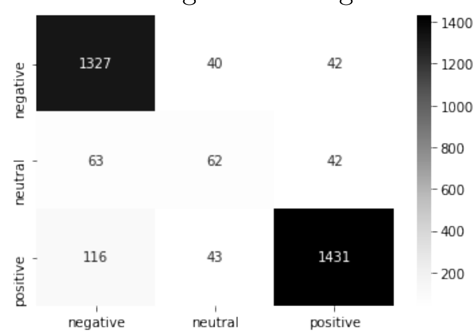
Methodology	F1-micro (%)	F1-macro (%)
Sentiment-based task		
Maximum Entropy (original)	87.94	68.47
Maximum Entropy (augmented)	**89.07**	**74.32**
Text-CNN (original)	**89.82**	75.57
Text-CNN (augmented)	89.38	**77.16**
Topic-based task		
Maximum Entropy (original)	84.03	71.23
Maximum Entropy (augmented)	**86.03**	**74.87**
Text-CNN (original)	86.63	75.23
Text-CNN (augmented)	86.32	74.86

Table 7: Empirical result of the UIT-VSFC dataset

Besides, Table 7 shows the result of the UIT-VSFC dataset on the sentiment-based and the topic based tasks, respectively, before and after enhanced data. The original F1-micro score of the UIT-VSFC on both sentiment-based and topic-based tasks are referenced from (Nguyen et al., 2018).



The original training set

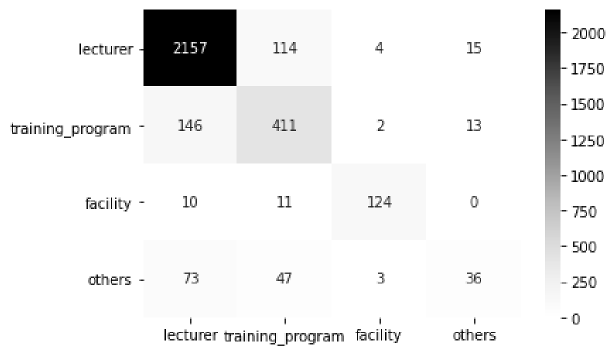


The augmented training set

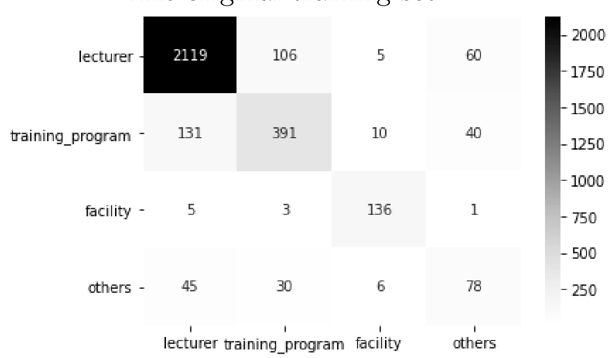
Figure 5: Confusion matrix of the Maximum Entropy model on the UIT-VSFC dataset before and after enhanced data for the sentiment-based task

According to Table 7, for the sentiment-based task, the UIT-VSFC dataset, after enhanced on the training set, gave better results than the original training set by the Maximum Entropy model on both F1-micro and F1-macro scores. The Text-CNN model gave better results by the F1-macro score when the training data are enhanced. For the topic based task, the result of Maximum Entropy are better after enhanced data. The Text-CNN results after augmented data, in contrast, are not as better as the original data.

In addition, Figure 5 illustrates the confusion matrix of the UIT-VSFC dataset trained by the Maximum Entropy model before and after enhanced data for the sentiment-based task, and Figure 6 indicates the confusion matrix of the UIT-VSFC dataset for the topic based task trained by the Maximum Entropy model. According to Figure 5, the ability of true prediction on the neutral label is increased after enhanced data. Nevertheless, according to Figure 6, the results before and after augmented data are just slightly different. Indeed, the enhanced data does not affect much on the performance result of the topic based task.



The original training set



The augmented training set

Figure 6: Confusion matrix of the Maximum Entropy model before and after enhanced data on the UIT-VSFC dataset for the topic-based task

Overall, for the HSD-VLSP hate speech dataset, the data augmentation techniques increase the models' performance. For the UIT-VSFC corpus, the data augmentation increased models' performance on the sentiment-based task by both Maximum Entropy and Text-CNN, while it does not impact the topic-based task.

5.4 Error analysis

According to Figure 6, on the UIT-VSFC dataset on the topic based task, the prediction of the *training_program* label seems to be inclined to the *lecture* label, and *others* label seem to be inclined to the *training_program* and the *lecturer* labels. Table 8 listed examples of those cases. It can be inferred from Table 8 that, most of cases the model predicted wrong to the *lecture* label because the texts have words related to lecture such as: teacher, teaching, lesson, and knowledge. So does the *training_program* label with the appearance of words related to training program topic such as: subjects, requirements, and outcomes.

No.	**Texts**	**True**	**Predict**
1	cô nhiệt tình, giảng bài hiệu quả (English: The teacher is so enthusiastic and teaches very well)	1	0
2	tiến độ dạy hơi nhanh (English: The teaching process is fast)	1	0
3	sinh viên khó tiếp thu kiến thức (English: Student feel difficult to understand the knowledge)	3	0
4	các yêu cầu của môn cần ghi rõ (English: The subject's requirements should be well described)	3	1

Table 8: Error analysis in the test set of the UIT-VSFC dataset on topic-based task. Label description: 0 - lecturer, 1 - training program, 2 - facility, 3 - others

6 Conclusion

The imbalance in the datasets impact the performance of the machine learning models. Therefore, this paper focuses on the techniques that decreased the skewed distribution in the dataset by enhancing minority classes' data. We implemented the EDA techniques on the VLSP hate speech and the UIT-VSFC datasets and studied data augmentation's effectiveness on the imbalanced dataset. The results show that, when the data on the minority labels are increased, the model's ability to predict those labels is higher. However, the data augmentation techniques pull down the accuracy of other labels. Therefore, it is necessary to consider whether it is appropriate to apply the data augmentation techniques in a specific problem.

In the future, we will construct the lexicon-based dictionary for sentiment analysis in the Vietnamese language, especially the abusive lexicon-based words like Hurtlex (Bassignana et al., 2018) for hate speech detection to improve the ability of the machine learning model. We will also implement modern techniques in text classification such as the BERT model (Devlin et al., 2019) and the attention model (Yang et al., 2016) to increase the performance. Furthermore, in the hate speech detection problem, we will construct a new dataset which is more diverse in data sources and more balanced among classes.

Acknowledgments

We would like to give our great thanks to the 2019 VLSP Shared Task organizers for providing a very valuable corpus for our experiments.

References

Charu C. Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. In Charu C. Aggarwal and ChengXiang Zhai, editors, *Mining Text Data*, pages 163–222. Springer.

Aida Ali, Siti Mariyam Hj. Shamsuddin, and Anca L. Ralescu. 2015. Classification with class imbalance problem: a review. In *SOCO 2015*.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CLiC-it*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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. Association for Computational Linguistics.

N. T. Duyen, N. X. Bach, and T. M. Phuong. 2014. An empirical study on sentiment analysis for vietnamese. In *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, pages 309–314.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4), July.

M. Ibrahim, M. Torki, and N. El-Makky. 2018. Imbalanced toxic comments classification using data augmentation and deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Daniel Jurasky and James H Martin. 2000. Speech and language processing: An introduction to natural language processing. *Computational Linguistics and Speech Recognition. Prentice Hall, New Jersey.*

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.

S. T. Luu, H. P. Nguyen, K. Van Nguyen, and N. Luu-Thuy Nguyen. 2020. Comparison between traditional machine learning models and neural network models for vietnamese hate speech detection. In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6.

Phuong-Thai Nguyen, Van-Lam Pham, Hoang-Anh Nguyen, Huy-Hien Vu, Ngoc-Anh Tran, and Thi-Thu Ha Truong. 2016. A two-phase approach for building vietnamese wordnet. In *Proceedings of the 8th Global WordNet Conference. Bucharest, Romania*, pages 259–264.

K. V. Nguyen, V. D. Nguyen, P. X. V. Nguyen, T. T. H. Truong, and N. L. Nguyen. 2018. Uit-vsfc: Vietnamese students' feedback corpus for sentiment analysis. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 19–24.

Kamal Nigam, John Lafferty, and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 workshop on machine learning for information filtering*, volume 1, pages 61–67. Stockholom, Sweden.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th*

ACM International Conference on Information and Knowledge Management, CIKM 19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Xuan-Son Vu, Thanh Vu, Mai-Vu Tran, Thanh Le-Cong, and Huyen T M. Nguyen. 2019. HSD shared task in VLSP campaign 2019: Hate speech detection for social good. In *Proceedings of VLSP 2019*.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2557–2563, Lisbon, Portugal, September. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6383–6389, Hong Kong, China, November. Association for Computational Linguistics.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, June. Association for Computational Linguistics.