A data augmentation technique based on text for Vietnamese sentiment analysis

Thien Ho Huong*
thienhh.188i@ou.edu.vn
Ho Chi Minh City Open University
Ho Chi Minh City, Vietnam

Vinh Truong Hoang
vinh.th@ou.edu.vn
Ho Chi Minh City Open University
Ho Chi Minh City, Vietnam

ABSTRACT

Online opinions are used as a data source that contains relevant information about customer sentiments toward a product or service. This can be used to make a specific decision for customers and management. In order to achieve the good models for sentiment analysis, we require a large human-labeled data which is costly to obtain. This paper proposes an approach based on text data augmentation based on product reviews in Vietnamese language. Several basic techniques are applied to generate more comments by random insertions, substitutions. The experimental results demonstrate the efficiency of the proposed approach.

KEYWORDS

sentiment analysis, product reviews, text augmentation, Vietnamese language, natutal language processing, text mining

ACM Reference Format:
Thien Ho Huong and Vinh Truong Hoang. 2020. A data augmentation technique based on text for Vietnamese sentiment analysis. In IAIT '20: International Conference on Advances in Information Technology, July 01–03, 2020, Bangkok, Thailand. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/XXXXXX.XXXXXX

1 INTRODUCTION

Throughout the era of digitization, a growing number of people are expressing their opinions on the Web, for example throughout public reviews. These comments are important for running businesses or services because they provide recommendations for enhancement. Therefore, the decisions of customers rely heavily on reviews [1]. However, manually analyzing those comments is time-consuming and it is more challenging to generalize the results. Sentiment Analysis (SA) is a research topic which aimed at recognizing and analyzing the components of a person's opinion in machine learning. In the last decade, SA has been received many attentions and it is widely applied in various application such as market analysis [2], analyze the reviews related to products [3, 4], political communication [5], social medias [6, 7].

Sentiment analysis can be considered as a text classification problem which is essentially based on computational linguistics Natural Language Processing (NLP). It is easier for mining text in longer documents than in short texts due to the context of semantic understanding. The sentiments of an aspect can be divided into different categories depending on the purpose of sentiment classification such as: positive, negative and neutral classes. Thus, it is very simple to collect large quantities of unlabeled data from online social networks but very costly to fully label texts into categories. However, the classification results mainly reply on supervisory information and required a large scale labeled data to train the model. Data augmentation method is one of the most techniques used for tackling this issue for generating more data. It is widely applied in computer vision [8] by using simple techniques such as flip, rotate, crop, scale or color texture features [9] to transform the original images. Due to the meaning of words, grammar diversity of language and context so data augmentation in NLP is still a challenging problem.

Several augmenting training data with semi-supervised learning have been proposed for dealing with limited training data in literature. Lu et al. [10] apply the propagation method to generate unlabeled data via a weighted undirected graph. Lee et al. [11] combined two learning approaches on supervised and unsupervised in order to handle small number of labeled sentiments. Shakeel et al. [12] proposed a data augmentation strategy and a multi-cascaded model for improved paraphrase detection in short texts. This approach is based on binary relations over the set of texts which apply graph theoretic concepts to generate paraphrase and non-paraphrase pairs. Jason Wei et al.[13] present a basic technique for text data augmentation, namely Easy Data Augmentation (EDA) including synonym replacement, random insertion, random Swap, and random Deletion. Additionally, the similarity word replacement method based on vector word embedding space, and random noise method is investigated in this study. The work in [14] applied synonym replacement method but it restricted the words to nouns, adjectives and adverbs. A score is computed for each synonym and the best score is chosen for the selected words. The experimental results showed that in some cases replacing

*Both authors contributed equally to this research.

IAIT '20, July 01–03, 2020, Bangkok, Thailand

ACM ISBN 978-1-4503-XXXX-XXXXXXX. . . $15.00
https://doi.org/10.1145/XXXXXX.XXXXXX

verbs or prepositions with synonym words make the ungrammatical sentences and can be wrong in contextual meaning. However, this phenomenon does not occur in case of nouns, adjective and adverbs synonym replacement.

Sentiment analysis for Vietnamese language has received significantly less attention in the literature. The authors in [15] applied the basic data augmentation technique by using synonym replacement and random swap in semi-supervised learning context. In this paper, we further investigate this problem by using several methods of word replacement by synonym or similar words for Vietnamese text. The word embedding is based on the cosine distance for measuring the similarity [16]. The rest of this paper is organized as follows. Section 2 illustrates our proposed method. Then, section 3 describes the experimental results. Finally, the conclusion is discussed in section 4.

2 TEXT DATA AUGMENTATION APPROACH

The overall of the proposed text data augmentation approach is illustrated in figure 1. Online comments on products are used as a data source to make various management decisions. These reviews are often given in short sentences by different ways of expression. So, the text Pre-processing is a main step to reduce noise from those comments. The contents consist of more characters which are not meaningful so that it will be removed from the training dataset. These basic comment pre-processing are tokenization, removing URLs, removing hashtag, removing email, removing @user, emoticons handling, removing numbers, lowercasing, removal of duplicate letters, punctuations removal. In this study, we focus on words preprocessing which include segmentation, stopwords removal, negation handling.

Segmentation for Vietnamese language is an essential task due to its grammatical complexity. In Vietnamese, a word can have completely distinct meaning when it is an individual position or combining with another word. For example, the meaning of the word "đất" (sand) and "nước" (water) when they are combined is "đất nước" (country). Therefore, we need a strong enough segmentation tool to do this. In this study, the pyvi library is applied for Vietnamese word segmentation. Moreover, the removal of stopwords is applied to remove words that are less meaningful for sentiment analysis. Many Vietnamese stop words are considered as: "thì" (to be), "nhưng" (but), "là" (to be), "vì" (because). These stopwords are created manually based on the frequency term in the data by determining their TF-IDF score. For negation handling, we have built a list of negation words based on the study [17]. Some negative words such as không (not), chẳng (not), chưa (not yet), chả (not), đâu (not), đâu có (not at all), nào (any), nào có (any), khỏi (without), ứ (not). We first determine the negative words of each sentence and then combine sentiment words in [18]. An addition of the symbol NOT_ is realized to the tokens by positive or negative lexicons.

All comments will be tokenized to individual words in order to characterize by feature vectors. The two extraction methods are considered such as: BOW (Bag of Words), TF-IDF (Term Frequency-Inverse Document Frequency) since they are simple and efficient for representing textual data[19]. The Term Frequency (TF) is a frequency of word appearance and the number of times a word appears in a document, divided by the total number of words in that document. Where t is a word in a document, $f(t,d)$ is a frequency of word occurring in that document and T is a total number of words in the document.

$$TF(t) = \frac{f(t,d)}{T} \tag{1}$$

There are some words that appear in most documents but it has no meaning for sentiment recognition. For example "thì" (to be), "mà" (yet), "nhưng" (but) etc. The Inverse Document Frequency (IDF) is computed as the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears:

$$IDF(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{2}$$

where N is a total number of documents and denominator is the number of documents containing the word t. In case, the word disappears in any document then the equation will become invalid so we need to set the value of the denominator as 1. The TF-IDF is finally calculated as below:

$$TF - IDF(t,d,D) = TF(t) \times IDF(t,D) \tag{3}$$

The text data augmentation is applied by using four techniques [13] to generate more comments:

(1) Words Replacement: many works applied WordNet to replace the synonym words but to the best of our knowledge, WordNet does not perform well for Vietnamese text. So, the similar words based on cosine distance of word embedding vectors, Word2vec[16] are used to replace words. The pre-trained model [20] is applied in our experiments.
(2) Words Insertion: this technique is used to find and insert the synonym or similar word at the end of each sentence.
(3) Words Swapping: this technique will be implemented n times, where n is a subtraction of the tokenized number of each sentence.
(4) Words Deletion: a new sentence will be created after deleting some words including verbs, adverbs, prepositions.

Table 1 illustrates an example of a comment "Nhân viên phục vụ nhiệt tình và lịch sự" (Enthusiastic and polite service staff) by using four data augmentation techniques. When these techniques are applied the grammar and meaning of the sentences might be changed totally but its general sentiment does not change. It is worth noting that our scope is to focus on sentiment of sentences. The grammatical structure and global context is not considered in this case.

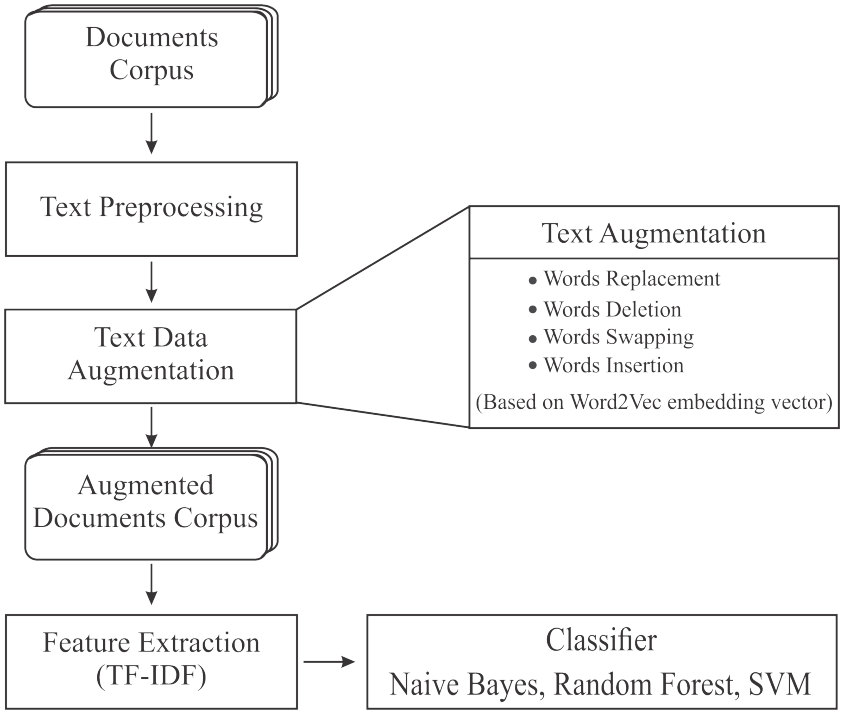

Figure 1: Overall of the proposed text data augmentation approach.

Table 1: Sentences generated for the comment "Nhân viên phục vụ nhiệt tình và lịch sự" (Enthusiastic and polite service staff) by using four data augmentation techniques.

Techniques	Sentence
None	Nhân_viên phục_vụ nhiệt_tình và lịch_sự
Words Replacement	Nhân_viên cao_cấp hoan_nghênh và lịch_sự
Words Insertion	Nhân_viên phục_vụ nhiệt_tình và lịch_sự chuyên_viên cao_cấp
Words Deletion	Nhân_viên nhiệt_tình lịch_sự
Words Swapping	nhiệt_tình Nhân_viên và phục_vụ lịch_sự

3 EXPERIMENTAL AND RESULTS

We use the dataset provided in [15] order to evaluate our proposed approach. Dataset 1 and dataset 2 are the collection of Vietnamese comments on food which were crawled on streetcodevn.com. Dataset 3 is collected from an AI contest for Vietnamese sentiment analysis. The characteristic of these datasets is illustrated in table 2. All comments are divided into emotional polarity by two-class classification problems (positive and negative).

For classification, there are many classifiers state-of-the-art in natural language processing. In the studies [21], [22] the authors showed comparative study of classifier to evaluate the effectiveness of proposed text data augmentation, so we apply experiment by three well-known classification algorithms such as Naive Bayes, Random Forest, Support Vector Machine. All experiments were implemented by Python library and on a PC configured with Intel core I7 and 8 Gigabyte of memory.

After the preprocessing step, the data augmentation is applied to generate more comments. Table 3 presents the number of comments and words corresponding to each dataset before and after augmentation step. The number of comments increased four times. The number of words augments nearly 7 millions words for dataset 1. Three common classifiers for text classification [21] including Naive Bayes (NB), Random Forest (RF) and Support Vector Machine (SVM) are used to recognize sentiments. Table 4 presents the classification performance on two scenarios of with and without text augmentation. The average accuracy of NB classifiers achieves at 84% in both cases before and after augmentation. The similar

Table 2: Characteristic of dataset used in experiments.

No	Name	Emotional polarity	Number of comments	Labels
1	Dataset 1	Positive	15,000	Categorical
		Negative	15,000	
2	Dataset 2	Positive	5,000	Categorical
		Negative	5,000	
3	Dataset 3	Positive	7,383	Binary
		Negative	8,690	

Table 3: The number of comments and words before and after data augmentation step.

	Before augmentation			After augmentation		
Data Source	Dataset 1	Dataset 2	Dataset 3	Dataset 1	Dataset 2	Dataset 3
Number of comments	30,000	10,000	16,073	120,000	40,000	64,292
Number of words	2,962,235	1,003,237	347,733	10,438,550	3,534,413	1,287,089

observation is acquired for SVM classifiers. The RF classifier demonstrates its efficiency for handling high-dimensional feature vectors in case data augmentation is applied. Therefore, the best accuracy is 95% given by the RF classifier and it improves nearly 10%.

Table 4: The classification performance (%) of two scenarios: before and after text augmentation.

	Before augmentation			After augmentation		
Name	NB	RF	SVM	NB	RF	SVM
Dataset 1	82	84	86	83	97	86
Dataset 2	82	83	85	84	97	89
Dataset 3	88	91	91	85	92	88
Average	84	86	87	84	95	87

4 CONCLUSION

We presented a method based on data augmentation techniques for sentiment analysis. The experimental results on three datasets have been shown the efficiency of the proposed approach. We improve about 10% on the product comments in Vietnamese language by using simple techniques for replacing words and random insertions associated with several common classifiers. The extension of this work is now continuing to build the sentiment lexicon for Vietnamese language to enhance the performance of augmentation method.

REFERENCES

[1] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2):1–135, 2008.

[2] Dattatray P. Gandhmal and K. Kumar. Systematic analysis and review of stock market prediction techniques. Computer Science Review, 34:100190, November 2019.

[3] Tanjim Ul Haque, Nudrat Nawal Saber, and Faisal Muhammad Shah. Sentiment analysis on large scale Amazon product reviews. In 2018 IEEE International Conference on Innovative Research and Development (ICIRD), pages 1–6, Bangkok, May 2018. IEEE.

[4] Ana Valdivia, Emiliya Hrabova, Iti Chaturvedi, M. Victoria Luzón, Luigi Troiano, Erik Cambria, and Francisco Herrera. Inconsistencies on TripAdvisor reviews: A unified index between users and Sentiment Analysis Methods. Neurocomputing, 353:3–16, August 2019.

[5] Martin Haselmayer and Marcelo Jenny. Sentiment analysis of political communication: combining a dictionary approach with crowdcoding. Quality & Quantity, 51(6):2623–2646, November 2017.

[6] Mohammad A. Hassonah, Rizik Al-Sayyed, Ali Rodan, Ala' M. Al-Zoubi, Ibrahim Aljarah, and Hossam Faris. An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter. Knowledge-Based Systems, 192:105353, March 2020.

[7] Zulfadzli Drus and Haliyana Khalid. Sentiment Analysis in Social Media and Its Application: Systematic Literature Review. Procedia Computer Science, 161:707–714, 2019.

[8] Luis Perez and Jason Wang. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. arXiv:1712.04621 [cs], December 2017. arXiv: 1712.04621.

[9] H. Duong and V. T. Hoang. Data Augmentation Based on Color Features for Limited Training Texture Classification. In 2019 4th International Conference on Information Technology (InCIT), pages 208–211, 2019.

[10] X. Lu, B. Zheng, A. Velivelli, and C. Zhai. Enhancing Text Categorization with Semantic-enriched Representation and Training Data Augmentation. Journal of the American Medical Informatics Association, 13(5):526–535, September 2006.

[11] Vivian Lay Shan Lee, Keng Hoon Gan, Tien Ping Tan, and Rosni Abdullah. Semi-supervised Learning for Sentiment Classification using Small Number of Labeled Data. Procedia Computer Science, 161:577–584, 2019.

[12] Muhammad Haroon Shakeel, Asim Karim, and Imdadullah Khan. A multi-cascaded model with data augmentation for enhanced

paraphrase detection in short texts. Information Processing & Management, 57(3):102204, May 2020.

[13] Jason Wei and Kai Zou. 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, 2019. Association for Computational Linguistics.

[14] Praveen Giridhara, Chinmaya Mishra, Reddy Venkataramana, Syed Bukhari, and Andreas Dengel. A Study of Various Text Augmentation Techniques for Relation Classification in Free Text:. In Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods, pages 360–367, Prague, Czech Republic, 2019. SCITEPRESS - Science and Technology Publications.

[15] Dang-Khoa Nguyen-Nhat and Huu-Thanh Duong. One-Document Training for Vietnamese Sentiment Analysis. In Andrea Tagarelli and Hanghang Tong, editors, Computational Data and Social Networks, volume 11917, pages 189–200. Springer International Publishing, Cham, 2019.

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs], September 2013. arXiv: 1301.3781.

[17] Bùi Thanh Hoa. Nhóm hư từ mang ý nghĩa phủ định trong tiếng Việt. Tạp Chí Ngôn Ngữ, page 9, 2014.

[18] Xuan-Son Vu and Seong-Bae Park. Construction of Vietnamese SentiWordNet by using Vietnamese Dictionary. page 4.

[19] Ravinder Ahuja, Aakarsha Chug, Shruti Kohli, Shaurya Gupta, and Pratyush Ahuja. The Impact of Features Extraction on the Sentiment Analysis. Procedia Computer Science, 152:341–348, 2019.

[20] Xuan-Son Vu. Pre-trained word2vec models for vietnamese, 2016.

[21] Huu-Thanh Duong and Vinh Truong Hoang. A Survey on the Multiple Classifier for New Benchmark Dataset of Vietnamese News Classification. In 2019 11th International Conference on Knowledge and Smart Technology (KST), pages 23–28, Phuket, Thailand, January 2019. IEEE.

[22] Yassine Al Amrani, Mohamed Lazaar, and Kamal Eddine El Kadiri. Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia Computer Science, 127:511–520, 2018.