Computational Social Networks

RESEARCH Open Access

A review: preprocessing techniques and data augmentation for sentiment analysis

Huu-Thanh Duong[1*] and Tram-Anh Nguyen-Thi[2]

*Correspondence:
thanh.dh@ou.edu.vn
[1] Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam
Full list of author information is available at the end of the article

Abstract

In literature, the machine learning-based studies of sentiment analysis are usually supervised learning which must have pre-labeled datasets to be large enough in certain domains. Obviously, this task is tedious, expensive and time-consuming to build, and hard to handle unseen data. This paper has approached semi-supervised learning for Vietnamese sentiment analysis which has limited datasets. We have summarized many preprocessing techniques which were performed to clean and normalize data, negation handling, intensification handling to improve the performances. Moreover, data augmentation techniques, which generate new data from the original data to enrich training data without user intervention, have also been presented. In experiments, we have performed various aspects and obtained competitive results which may motivate the next propositions.

Keywords: Machine learning, Semi-supervised, Sentiment analysis, Preprocessing techniques, Data augmentation, Vietnamese processing, Limited training data

Introduction

An enormous and rapid growth of the Web technologies has changed the way to buy and reply the reviews of the bought products. The customers can reply their reviews rapidly, the providers can receive an abundance of the customers' reviews. These help to understand the customers' expectation, evaluate the advantages and disadvantages of the products, also predict the product trends to satisfy the customers' expectation quickly as possible as. However, this cannot be perform manually, it imagines that many employees follow the customers' replies about a product, read and analyze the hundreds or thousands of the replies to evaluate the degree of customers' satisfaction for making next strategies about the products or taking a direction in development. This is costs both human resources and money a lot, reaches the customers' expectation is slow and easy to miss ones. A perfect solution for this problem is sentiment analysis which promises commercial benefits.

Sentiment analysis is an essential task to detect the sentiment polarities in the text applied widely in e-commerce system, blogs, social media. Its main task groups the document into various polarities. Based on automatic prediction, the traders can make decision easier, and also plan the direction to develop their business.

Working with sentiment analysis faces many challenges. The first one is text structure, Hussein [1] made a survey on sentiment analysis challenges by comparing many past studies; the authors showed types of text structure for sentiment analysis: (i) structured sentiments are format sentiment text; (ii) unstructured sentiments are informal and free text; (iii) semi-structured sentiments are between format structured text and unstructured text. The most difficult is working with unstructured sentiment, the writer is not required to comply with any constraints: using slang terms, wrong spelling, wrong grammar structure, etc. These have made it hard to analyze text structure, especially detecting negation text being big challenge impacting sentiment detection and text structure evaluation. Another challenge is to select the most relevant techniques or approaches to classify the sentiment polarities.

According to Medhat et al. [2], there are three main approaches for sentiment analysis: lexicon-based approach, machine learning approach, hybrid approach. Lexicon-based approach replies on the emotional lexicons to detect customers' emotions, its main drawbacks are to depend on the context and languages. Actually, sentiment analysis is text classification problem which can apply machine learning classifiers in emotional polarities, Soleymani et al. [3] made a survey summarized sentiment analysis methods, including in text, it showed many previous researches in supervised and unsupervised learning.

The traditional approach is usually supervised learning, supervised classifiers are used such as Naive Bayes, SVM, logistic regression, ensemble of voting classifiers, also investigating on feature selection for retaining useful features and ignoring redundant features to improve the performing approach. This depends on the size and quality of the pre-labeled datasets which are scarce and unavailable for a certain application, they are tedious to collect, expensive and time-consuming to build, depend on domain adaptation and ineffectively handle unseen data. Especially, the Vietnamese training data are not abundant and lack so much preventing many propositions in research team. This motivates us to develop an effective solution to text classification in general and sentiment analysis in particular.

Related works

Our study is based on semi-supervised learning and used many preprocessing techniques to normalize the data such as emoji icons replacement, elongated characters removal, negation handling, intensification handling. Furthermore, we have investigated to enhance more training data automatically in order to improve the performance of the models with the limited training data.

Preprocessing techniques are frequently used in natural language processing to prepare text that is going to be classified. Especially, reviews in e-commerce system, blogs and social media are informal, so they contain so much noisy information, unnecessary in detecting the sentiment. Those will clean text, normalize text and only keep useful information, Symeonidis et al. [4] and Effrosynidis et al. [5] summarized the preprocessing techniques and performed experiments to prove they improve significantly the accuracy of classifiers. Fernández-Gavilanes et al. [6] used sentiment lexicon to create by means of an automatic polarity expansion algorithm and some natural language processing techniques such as detecting of polarity conflicts or concessive subordinate clauses.

Singh and Kumari [7] evaluated the effects of preprocessing on twitter data and indicated the improvement of the classifiers. They removed URLs, hashtags, user mentions, punctuation, stop words and replaced slang words with actual words using *n*-gram. Similarly, Jianqiang and Xiaolin [8] also evaluated these on five twitter datasets, including expanding acronyms, replacing negation, removing URLs, numbers, stop words.

Negation handling is used to remove negative forms to reduce ambiguities of the classified sentences, this is one of important preprocessing techniques applied widely in sentiment analysis. AL-Sharuee et al. [9] used SentiWordNet 3.0 to prepare the underlying text for further processing and handle common linguistic forms such as intensifiers, negation, contrast. Next phase, they proposed binary ensemble clustering by assembling the results of a modified *k*-means algorithm, where the selected features are adjectives and adverbs in all documents.

Emoji icons have boomed in e-commerce system, blogs and social media, these express so much sentiment, Fernández-Gavilanes et al. [10] also constructed a novel emoji sentiment lexicon using an unsupervised sentiment analysis based on the definitions given by emoji creators Emojipedia and created lexicon variants thanks to the sentiment distribution of the informal texts. Wang and Castanon [11] conducted analyses to examine the relationship between the emotional icons and sentiment polarities, they confirmed a few emotional icons are strong signals of sentiment polarity, but a group the emotional icons conveys complicated sentiment in detecting sentiment polarities.

Data augmentation is a spotlight in recent years, from a limited training data will automatically generate more training data as considered semi-supervised learning. Sennrich et al. [12], Sugiyama and Yoshinaga [13] used back translation technique to generate training data to improve performance of translation model. Fadaee at al. [14] also proposed a novel approach that augments the training data to improve translation quality substantially, this targets low-frequency words to generate new sentence pairs containing rare words, synthetically created context. Kobayashi [15] proposed the contextual augmentation. They stochastically replace words with other words predicting by a bidirectional language model at word positions, language models improved with a label-conditional architecture which allows the model to augment sentences without breaking the label-compatibility.

Query expansion (QE) is also an effective solution to get more data, Azad et al. [16] made a survey the QE techniques in information retrieval (IR), its purpose is to reformulate the original query to enhance the IR effectiveness. This can be applied by putting the original query into a specific search engine and selecting the most relevant retrieval results as new query.

Şahin and Steedman [17] based on dependency tree to remove dependency links (crop) and move the tree fragments around the root (rotate) for new data, this proposal was inspired from two techniques augmenting data in image processing as cropping and rotating, their experiments showed improvement for majority of low-resource languages.

Wei and Zou [18] applied some easy data augmentation (EDA) techniques, namely synonym words, random swap, random insert, random delete to generate new data. Although these techniques are easy to implement, not depending on any external resources, they improve the performances substantially.

In this paper, we aim to improve the accuracies of the models with limitation of training samples by using the preprocessing techniques to normalize and clean data, also enhancing training samples automatically for the original samples. In order to evaluate our approach, we have applied the well-known multiple classifiers such as logistic regression, support vector machine, and ensembles of classifiers such as one-vs-one, one-vs-all.

Our main contribution has evaluated the effects of preprocessing techniques and data augmentation for Vietnamese. We have summarized the preprocessing techniques, investigated data augmentation techniques and experimented to examine the possibility to generate training data automatically applied for Vietnamese to improve the accuracies of the algorithms. This is the necessary and meaningful step due to the limitation of the Vietnamese dataset, this enriches the Vietnamese dataset and saves the time and cost to build the pre-labeled dataset of a certain domain.

In the rest of this paper is organized as follows: Sect. 3 presents our background and approach, the experiments is presented in Sect. 4, and Sect. 5 is the conclusions and future works.

The approach

Preprocessing techniques

Most of recent studies in sentiment analysis focus on the user-generated texts which have been based on habit and are informal, hence it is necessary to clean, normalize language, also remove noisy information to be classified.

Vietnamese segmentation this is always a required step to work with Vietnamese, for example "“đây là điện thoại tuyệt vời”" (this is a wonderful phone) is tokenized "“đây là điện_thoại tuyệt_vời”" (using pyvi[1] library). Unlike English, words are separated by whitespaces and punctuations, Vietnamese words may contain many tokens and they must be processed, if not, the meaning of the sentence can be much different from the original expectation.

Lowercase is a classic preprocessing technique converting all texts into lowercase form. The same words are merged, so the dimensionality of the problem is reduced, for example "tốt" (good) and "Tốt" (Good) is the same dimensionality. This techniques have been widely used by many researchers [19–21].

Stop words removal stop words are function words, they are usually less meaning words and do not contain any sentiment, but appear high frequencies in texts. They should be removed to reduce the dimensionality and the computational cost, also improve the performance. The set of these words is not totally predefined depending on the application. In our experiments, they are determined stop words list based on term frequencies and inverse document frequencies weights in the collected datasets.

Elongated characters removal some characters are elongated one or more times in a lexicon to emphasize sentiment, this can lead to increase unnecessary dimensionality because the classifiers treat them as different words, even they may be ignored due to low frequency. So the elongated characters removal transforms the word to the source

[1] https://pypi.org/project/pyvi/.

word in order to merge them in the same dimensionality. For example "quuuuáaaa" is replaced by "quá" (so), "tuyệt_vờiiiii" is replaced by "tuyệt_vời" (wonderful). The experiments of Symeonidis [4] proved the improvement of this one.

Abbreviations or wrong-spelling lexicons replacement abbreviations and wrong-spelling words become a habit and are usually used in reviews of social media or e-commerce system, they should also merge into the source word, for example dth -> dễ_thương (cute); iu, êu -> yêu (lovely); omg -> oh my god; k -> không (not); sd -> sử dụng (use); ote, okay, oki, uki, oke -> ok; tẹc vời, tẹt zời, tẹc zời, toẹt vời -> tuyệt_vời (wonderful); sức sắc, xức sắc, xúc sắc, xs, xs -> xuất_sắc (excellent); wá, qá -> quá (so). Currently, a list of abbreviation and wrong-spelling lexicons have prepared manually for our experiments based on observing the reviews in social media and our collected datasets. Kim [19] corrected the common spelling mistakes by using AutoMap. Symeonidis et al. [4] also mentioned this technique in a comparison of preprocessing techniques.

Emotional icons replacement emotional icons have been widely used in reviews and denotes users' sentiment. Wang and Castanon [11] analyze and compare sentiments of tweets with and without emotional icons in order to provide the evidences the importance of emotional icons in expressing the sentiment in social media. In our case, the positive and negative icons are, respectively, replaced by "pos" and "neg" lexicons, for example: :) is replaced by "pos" lexicon or :(is replaced by "neg" lexicon.

Punctuation removal some punctuations (excluding underscore _ is used for Vietnamese segmentation) usually do not affect the sentiment, it should be removed to reduce noise, for example: "quá đẹp!, yêu điện_thoại này!" (so beautiful!, love this phone!) will be "quá đẹp yêu điện_thoại này". However, some punctuations contain sentiment, so it might decrease the accuracy of classification in those cases such as :), :D, ;), < 3 are positive icons which affect sentiment in reviews. In our works, this one will be applied after emotional icons replacement. Kim [19] also removed punctuation, URLs, stop words not containing any sentiment to improve performance.

Numbers removal normally, numbers do not contain any sentiments, it is necessary to remove them, but this should be performed after emotional icon replacement, wrong-spelling replacement because any of them contain numbers such as :3, < 3, 8|, 8-), etc.

Part of Speech (POS) handling POS tagging is an essential problem in natural language processing to assign part of speech to each words in a sentence as noun, verb, adjective, pronoun. This is helpful to increase semantic in text. In our works, POS tagging is used to retain words containing the sentiment, namely nouns, adjectives, verbs, adverbs. Symeonidis et al. [4] also applies POS tag and keeps nouns, verbs, adverbs in experiments. For example of our case "điện_thoại đó đẹp quá, tôi rất hài_lòng " (that phone is so beautiful, I am so pleased), the part of speech for each words is "điện_thoại/N đó/P đẹp/A quá/R, tôi/P rất/R hài_lòng/A" (N: noun, P: pronoun, A: adjective, R: adverb), the sentence becomes "điện_thoại đẹp quá, rất hài_lòng". In order to do this, we also used pyvi [1] library for POS tagging.

Negation handling is a challenge in sentiment analysis, for example "sản_phẩm không tốt" (the product is not good) used terms to vectorize, if not considering "không" (not) term, it might evaluate this is a positive sentence instead of a negative one. Normally, when detecting a negation lexicon (không (not), chẳng (not), chưa (not yet), etc) following by a positive or negative lexicons, those phrases should be

replaced by antonyms of next lexicon, for example the phrase "không tốt" (not good) is replaced by "xấu" (bad) as an antonym of "tốt" (good) based on a certain wordnet. However, based on the experiments of Symeonidis et al. [4], replacing negations with antonyms only logistic regression algorithm of SS-Twitter dataset and Convolutional Neural Networks for SemEval dataset beat the baseline. Even, Xia et al. [22] presented many machine learning algorithms fail replacing negations with antonyms.

Our works are based on the negation terms (không (not), chẳng (not), chưa (not yet)) to detect the negation, Fernández-Gavilanes et al. [6] also estimated negation scope based on some negator forms (not, no, never, neither). Our case has no Vietnamese wordnet being strong enough for negation replacement, so if detecting the negation following by a positive lexicon, then replacing by "not_pos" lexicon, it also detects the negation following be a negative lexicon, then replace by "not_neg" lexicon. After that, in order to show affectation of lexicons, we append "pos" and "neg" lexicons whenever appearing a positive and negative lexicon, respectively. In our experiments, this improves significantly accuracy of classifiers.

For example the sentence "thiết_kế của điện_thoại không đẹp, nhưng nó có hiệu năng tốt" (this phone design is not nice, but its performance is good), "không đẹp" (not nice) is a negation phrase, "đẹp" (nice) is a positive lexicon, so it is replaced by "not_pos" lexicon, and "tốt " (good) is also a positive lexicon, so the sentence becomes "thiết_kế của điện_thoại **not_pos**, nó có hiệu_năng tốt**pos**".

Intensification handling intensifier lexicons such as "rất" (very), "quá" (too), "hơi" (a bit), "khá" (pretty) aim to emphasize, increase or decrease the semantic meanings of the lexicons which precede or follow them. This is also so necessary to detect the degree of customers' satisfaction. Fernández-Gavilanes et al. [6] applied intensification treatment as a preprocessing technique, they used the parsing to determine which semantic orientation altered.

For our works, if the program detects an increasing intensifier lexicon preceding or following by a positive or negative lexicon, then appending "strong_pos" or "strong_neg" lexicon, respectively. Otherwise, if detecting a decreasing intensifier lexicon preceding or following by a positive or negative lexicon, then appending "pos" and "neg" lexicon, respectively. This one will be applied after negation handling.

For example the sentence "thiết_kế đẹp, cấu_hình mạnh, tôi rất hài_lòng" (nice design, strong configuration, I'm very pleased), "rất" is an increasing intensifier lexicon and follows by a positive lexicon as "hài_lòng" (pleased), so the sentence becomes "thiết_kế đẹp, cấu_hình mạnh, tôi rất hài_lòng **strong_pos**". For intensification handling, we have also prepared a list of intensifier lexicons used frequently in Vietnamese, grouped into increasing (rất, quá, lắm) and decreasing (tạm, khá, hơi, cũng được) semantic.

Other techniques relate to morpheme of word not using such as stemming, lemmatizing since Vietnamese is an inflexionless language, words are only one form.

Data augmentation

The original data augmentation is used in image classification by increasing image data such as rotate, translate, scale, add noise, etc. Similarly, data augmentation has

also applied for text classification by increasing text data based on various techniques. In text, data augmentation is more complex, many studies have been investigated to get new data, also improve quality of new data without user intervention. This helps to enhance the original training data to increase accuracies of models. However, it notes that data augmentation is only useful for a small dataset.

Firstly, we present EDA techniques which was introduced by Wei and Zou [18] and apply them to Vietnamese.

Synonym replacement the words are not stop words, their synonym words have obtained randomly and replaced them for a new sentence. *Random Swap* will swap *n* times two non-stopword words randomly. *Random Insert* will find a random synonym of a non-stopword word and insert this randomly *n* times in the sentence. *Random Delete* will randomly remove each word in the sentence with probability *p*. These processes have repeated many times until having the expected training data. For example the sentence: "*tuyệt_vời! tôi rất hài_lòng*" (great! I'm so pleased) has applied these techniques which may generate new sentences as follows:

- Synonym Replacement: "tuyệt_vời! tôi rất **ưng_ý**" ("ưng_ý" is a synonym word of "hài_lòng").
- Random Swap: "**hài_lòng**! **hài_lòng**! tôi rất **tuyệt_vời**" (swapping "tuyệt_vời " and "hài_lòng").
- Random Insert: "tuyệt_vời! tôi rất **mãn_nguyện** hài_lòng" ("mãn_nguyện" is also a synonym word of "hài_lòng").
- Random Delete: "tuyệt_vời! rất hài_lòng" (deleting "tôi" word).

Although these methods may increase the meaningless sentences, but they can increase the accuracies of classifiers in experiments. Depending on the dataset size to determine the repeated times because too much augmented data can lead to overfitting issue. The biggest disadvantage of these methods is not reserving meaning concerning the context of the sentences, so we present more complex approaches retaining the meaning as the original sentence.

Back translation aims to obtain more training samples based on the translators, many research teams have used to improve translation models [12–15, 23]. This technique is resolved by using the translators to translate the original data to a certain language, after that taking the translated data into the independent translator to translate back to the original language. Normally, the data of back translation will be never totally exact the same as the original data. English is one of languages having many training datasets for translation, others lack training datasets for translation models. So, English was utilized as the intermedia language to get more data.

For example the sentence "Mình rất thích mua máy ở tiệm này", Google translator translates it to English: "I really like to buy the device at this store", taking this translated sentence to Google translator to translate back to Vietnamese as: "Tôi thực sự muốn mua thiết bị tại cửa hàng này".

This approach is simple to understand and helpful to augment data retaining the meaning of the original data, but it needs effective translators. In experiments, we

have used Google Translation API to translate the original data in Vietnamese into English, and translate back to Vietnamese for augmented data.

Syntax-Tree Transformations is a rule-based approach to generate new data. From the original data, a syntactic parser builds a syntax tree, then using some syntactic grammars transform this tree to the transformed tree which is used to generate new sentence form. There are many syntactic transformations such as moving active voice to passive voice.

For example, the sentence "tôi thích điện_thoại này" (I like this phone) is parsed into "tôi/P thích/V điện_thoại/N này/P" (P: pronoun, N: noun, V: verb) and transforms to "điện_thoại này được thích bởi tôi" (this phone is liked by me). The generated data still retains the meaning of the original data, but this approach is costly time in calculation, especially Vietnamese which is complex in sentence structures.

Classifiers

Based on the experiments [24], we choose the best classifiers for our experiments, namely logistic regression, SVM and ensembles of classifiers as OVO and OVR.

Logistic regression (LR) is a statistical approach to determine relationship between the dependent variable y and a set of independent variables $\boldsymbol{x}$. In order to predict the label of a data point, this is based on the probability of logistic function and a predefined threshold belongs to [0, 1]. The logistic function is often used as sigmod function.

Support vector machine (SVM) is a strong classifier to find the hyperland which divides the dataset into various groups in multi-dimensional space, this must have the same distance between it and two hyperlands which contain the nearest data points belongs to two groups, respectively. For non-linearly separable dataset, SVM used kernel functions to transform the data points from non-linearly separable space into linearly separable space. Our experiments use RBF (radial basis function kernel) kernel as follows:

$$k(x, x') = exp(-\gamma \left\|x - x'\right\|_2^2), \gamma > 0, \tag{1}$$

where γ indicates how far the influence of a data point in calculation of a certain hyperland. Data points, which are low γ values are far from or high γ values which close to a separation hyperland, are considered in calculation.

One-vs-one (OVO) and one-vs-all (OVA) are ensembles of binary classifiers for multi-class problem. Each iteration of OVO takes a pairwise of classes and applies the binary classifier to indicate the label of a data point, the final label is determined based on majority voting of iterations. For OVR, the computational cost is lower, if having c classes then OVO needs to execute $c(c-1)/2$ iterations, about OVR takes only c iterations. For each iteration, the binary classifier determines whether a data point belongs to that label, the final label is determined based on a probability.

Table 1 Datasets are used for validation

No.	Datasets	Polarities	Size
1	Dataset 1	Positive	9280
		Negative	6870
2	Dataset 2	Strong positive	2380
		Positive	2440
		Negative	2380
3	Dataset 3	Positive	15000
		Negative	15000
4	Dataset 4	Positive	5000
		Negative	5000

Table 2 The samples of the datasets

No.	Datasets	Sample	Polarity
1	Dataset 1	chất_lượng sản_phẩm tuyệt_vời, đẹp, dùng ổn (*the product quality is wonderful, beautiful, stable*)	Positive
		sai mẫu_mã, sản_phẩm kém chất_lượng ... thất_vọng (*wrong sample, low quality ... disappointed*)	Negative
2	Dataset 2	chưa bao_giờ cảm_thấy hài_lòng với note như lần này. tuyệt_vời về mọi thứ (*I have never been pleased with note phone like this time. everything is wonderful*)	Strong positive
		tốt trong tầm giá, nhưng pin chưa đủ dùng. (*reasonable price, but it's not enough battery to use*)	Positive
		điện_thoại mới mua được 2 ngày đã xuất_hiện lỗi màn_hình (*the phone has just bought 2 days ago, the monitor is occurring error*)	Negative
3	Dataset 3	pizza ngon. phục_vụ nhanh nhiệt_tình. Thích quán này nhiềuu hihi (*pizza is delicious, service are quick i like this restaurant a lot hihi*)	Positive
		chờ vừa lâu mà đồ_ăn vừa dở, chẳng có gì là đặc_sắc! (*long waiting, bad food, nothing's good!*)	Negative
4	Dataset 4	không_gian nhà_hàng khá thoải_mái, món ăn rất ngon và vừa_miệng, hợp túi_tiền. nằm tại địa_điểm khá đẹp tại trung_tâm sài_gòn (*restaurant space is comfortable, food is delicious, price is reasonable. the location is at the center of SaiGon City*)	Positive
		phục_vụ lâu, ngồi chờ hơn 30 phút mới ra được đĩa cơm sườn. chất_lượng món ăn bình_thường, không tương_xứng với giá_cả (*long waiting for a dish of rice, the quality's normal, unreasonable price*)	Negative

The experiments

Data preparation and feature extraction

We have prepared four Vietnamese datasets as short reviews on watch, phone, food collecting from the internet and previous studies (see Sect. 6). Table 1 shows the size of

Table 3 The F1 scores of various experiments

Datasets	Classifiers	(1)	(2)	(3)	(4)	(5)	(2)+(3)	(2)+(4)	(2)+(5)
Dataset 1	LR	0.828	0.847	0.834	0.835	*0.830*	0.861	*0.862*	*0.861*
	SVM	*0.829*	*0.856*	0.829	0.837	0.812	0.854	0.861	0.838
	OVO	*0.829*	0.850	*0.838*	*0.839*	0.818	*0.863*	*0.862*	0.852
	OVR	*0.829*	0.850	*0.838*	*0.839*	0.818	*0.863*	*0.862*	0.852
Dataset 2	LR	0.700	0.739	0.706	0.704	*0.695*	0.743	0.743	*0.740*
	SVM	0.662	0.705	0.698	0.696	0.669	0.739	0.736	0.731
	OVO	*0.706*	*0.744*	*0.713*	*0.717*	0.681	*0.750*	*0.750*	0.736
	OVR	0.693	0.726	0.699	0.702	0.675	0.730	0.729	0.725
Dataset 3	LR	0.790	*0.820*	0.793	0.791	*0.786*	*0.827*	*0.824*	*0.824*
	SVM	0.789	0.818	0.793	0.791	0.782	0.824	0.821	0.819
	OVO	*0.793*	0.818	*0.794*	*0.793*	0.783	0.825	0.821	0.820
	OVR	*0.793*	0.818	*0.794*	*0.793*	0.783	0.825	0.821	0.820
Dataset 4	LR	0.779	0.820	0.790	*0.791*	*0.772*	*0.826*	0.822	*0.818*
	SVM	0.784	0.817	0.788	0.786	0.766	0.819	0.822	0.808
	OVO	*0.785*	*0.822*	*0.789*	0.787	0.766	0.824	*0.826*	0.808
	OVR	*0.785*	*0.822*	*0.789*	0.787	0.766	0.824	*0.826*	0.808

(1) Without preprocessing techniques (baseline results), (2) with preprocessing techniques, (3) back translation, (4) Syntax-Tree transformation, (5) EDA

each dataset used for validation and Table 2 presents some representative samples of each dataset.

Positive and negative lists contain Vietnamese positive and negative lexicons, respectively, including English lexicons used usually in Vietnamese sentences such as happy, nice, good, bad, etc. Negation list is to detect negation in the sentence such as "không" (not), "chưa" (not yet). And intensification list has grouped into increasing intensifiers ("rất" (very), "quá" (so), "lắm", "cực kỳ" (extremely)) and decreasing intensifiers ("tạm" (pretty), "khá" (pretty), "cũng" (also)).

In order to extract features, we used $tf \times idf$ weight which is widely used in natural language processing and has high score in text classification. The dimensionality is represented by unigram and bigram. F1 score is used to evaluate the performances of the approaches, it is the average of precision and recall metrics which reaches the best score at 1 and worst score at 0.

The experiments

As earlier mentioned, the approach is based on semi-supervised learning with a limited training data to reduce efforts to build a pre-labeled dataset. In order to prove the effects of preprocessing techniques and data augmentation techniques, we get *100 reviews* of every datasets for training, and all remaining ones will be used for validation data. Various experiments have been performed with the well-known classifiers.

Preprocessing techniques

The first column of Table 3 is the F1 scores of the classifiers without using preprocessing techniques (the baseline results), and the second one is with preprocessing techniques.

The performances of all datasets which are applied with preprocessing techniques improved well.

Nguyen-Nhat [21] has also showed the effects of these techniques, even using only one review for training. This proves that they are the reasonable techniques for Vietnamese sentiment analysis problem to normalize data before feeding to the classifiers.

Back translation

Back translation technique has generated 100 of new reviews for every polarities of training data, Google translator has been applied for this one, each review is translated between Vietnamese (vi) and English (en). The third column of Table 3 is the F1 scores of the classifiers to execute back translation technique to generate new reviews without using preprocessing techniques, the performances of the classifiers are improved a little bit for four datasets compared to the baseline results.

The sixth column of Table 3 incorporated the preprocessing techniques to normalize data. Later, applying back translation techniques to generate 100 of new samples more, the scores have been improved much better than the baseline results, also only preprocessing techniques or back translation results.

The reason is the normalized data help the results of translator better, so the new sample also obtained better quality. From that, it improves the performances of the classifiers rather than only using techniques independently. The results show that this is a promising approach to enhance training samples and improve the performance for Vietnamese sentiment analysis problem.

Syntax-tree transformation

For Syntax-Tree transformation, we have generated new reviews by converting each sentence of a review which is in active voice to passive voice. This has also generated 100 of new reviews for every polarities.

The F1 scores of this technique which are placed at the fourth column of Table 3 are also better than the baseline results, and the results of incorporating the preprocessing techniques presented at the seventh column of Table 3. The same as back translation techniques, these have boosted the performances of the classifiers. The normalized data help POS tagging works better leading to the quality of the converted sentences better. Thus, this approach is also suitable for Vietnamese sentiment analysis problem.

EDA

As above discussion, these above approaches have still retained meaning and grammar structure of the original data, but they are complex and need external resources. In other hand, EDA techniques are totally simple, easy to understand, do not need other predefined datasets or external resources, they still obtain promising performances in English [18].

Our works perform ten iterations for every EDA techniques, so every 100 reviews of the original training data will generate 4000 of new reviews. The fifth column of Table 3 is the results of these techniques performing without preprocessing techniques, but most of the scores are worse than the baseline results, only logistic regression of dataset1 is better a little bit. However, we have investigated one more

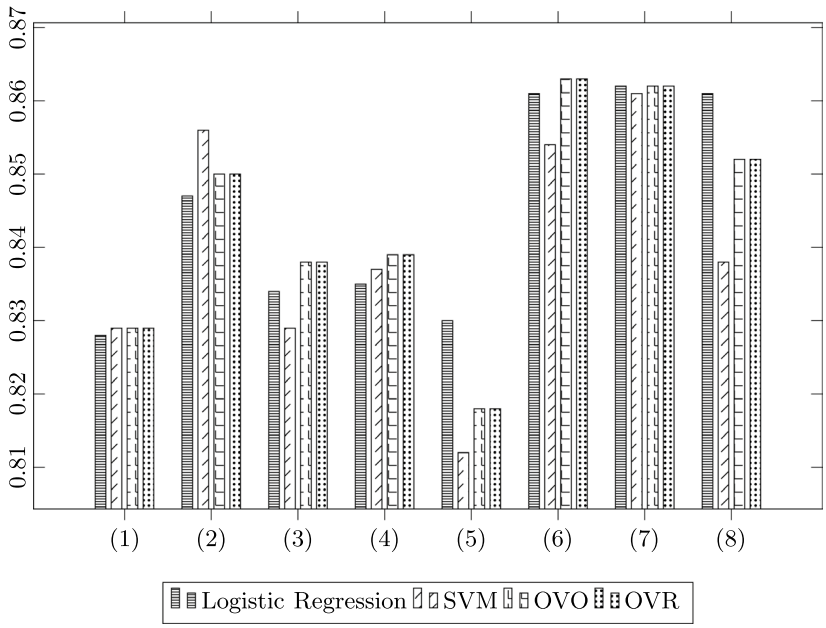

Fig. 1 The experimental results of dataset 1

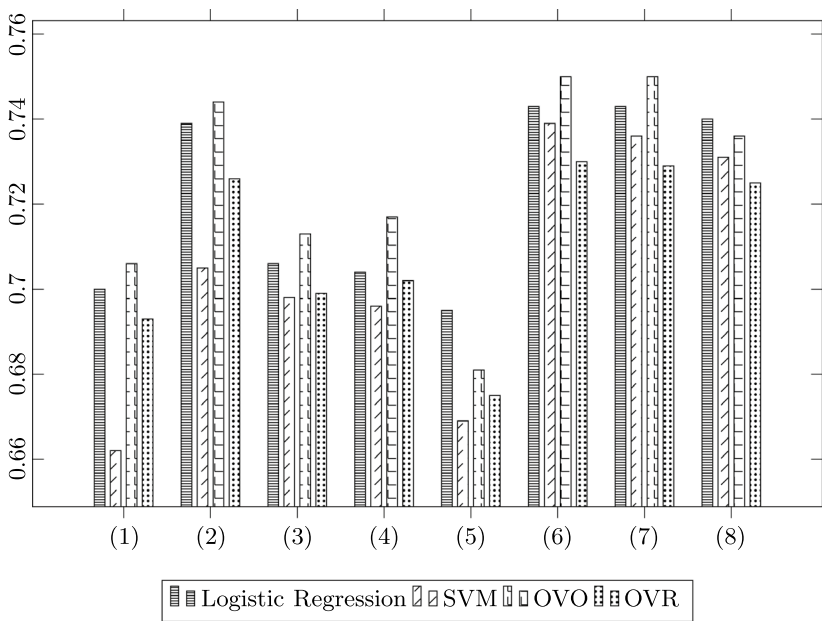

Fig. 2 The experimental results of dataset 2

experiment by incorporating these techniques with preprocessing techniques (the eighth column of Table 3), almost results of the datasets have been better than the baseline and preprocessing techniques (the second column), excluding the dataset 4 which is only better than the baseline results, but a little bit worse than the preprocessing techniques. These indicate that these techniques depend on the dataset, also

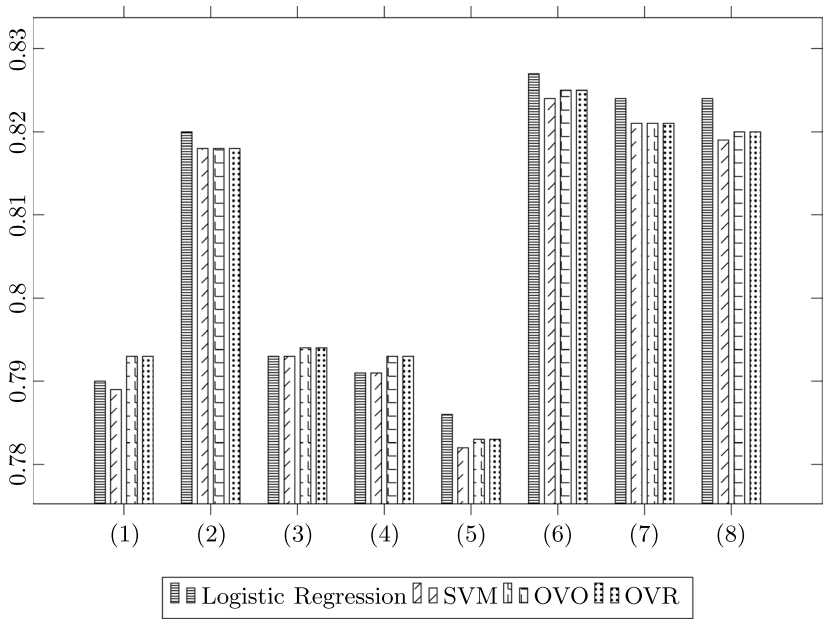

Fig. 3 The experimental results of dataset 3

the applied classifiers. Briefly, this is a potential solution to enrich training samples to boost the performances of Vietnamese sentiment analysis problem.

Some discussions may present worse results: synonym replacement or random insert technique depends on the way to get the synonym words, none of them are

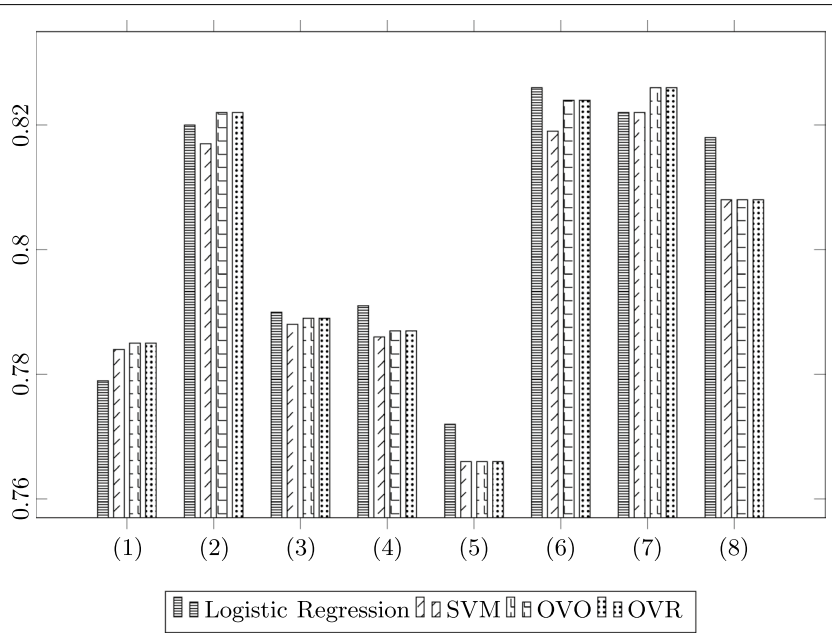

Fig. 4 The experimental results of dataset 4 (F1 scores)

suitable in a certain domain context, for example "tôi hài_lòng điện_thoại" (I am pleased the phone), "hài_lòng" (pleased) word has many Vietnamese synonym words such as ["mãn_nguyện", "thoả_mãn", "bằng_lòng", "ưng_ý"], replacing "hài_lòng" by "hài_lòng" by "ưng_ý" (contented) is better than "thoả_mãn" (satisfied) in this context. About delete random technique sometimes deletes some words containing sentiment in the sentence.

In short, the visualizable results of the experimented datasets [Figs. 1, 2, 3, 4] show the preprocessing techniques are the robust indicators to improve the performances in Vietnamese sentiment polarity. Moreover, data augmentation is a promising solution to make an abundance of training samples to boost the accuracies of the classifiers. Based on our experiments, back translation and Syntax-Tree transformation are the reasonable approaches and the EDA techniques have the potential to improve Vietnamese sentiment polarity.

Conclusions and future works

In this paper, we have based on semi-supervised learning for Vietnamese sentiment analysis, summarized the preprocessing techniques to normalize data and augmentation data techniques to generate new training data from the limited original training data. We have performed many experiments to present the effects of these techniques applying in Vietnamese. For most of experimented datasets, the accuracies of classifiers are improved. Moreover, ensembles of data augmentation techniques have also experimented and competitive results obtained. These experimented results show the approaches are reasonable and suitable for Vietnamese, this saves cost and time to build a pre-labeled dataset and gradually reach domain independence.

We can see that the performances of these techniques have depended on the original training data which are used to generate new data, and it will be better if the predefined data which serve preprocessing are collected large enough. Therefore, we will investigate to process slang words, concessive words, collect the intentional wrongspelling and abbreviation words in social media, select the better synonym words, and also propose the novel approaches of data augmentation techniques to obtain new training samples being more qualification, especially in Vietnamese context. In short, these results are promising scores, it can motivate our team and other propositions to improve the scores for this attractive problem in future.

Abbreviations

LR: Logistic regression; SVM: Support vector machine; OVO: One-vs-one; OVR: One-vs-all; W2V: Word2Vec; CBOW: Continuous Bag of Word; QE: Query expansion; IR: Information retrieval; POS: Part of speech; EDA: Easy data augmentation.

Acknowledgements

The authors would like to thank Dr. Vinh Truong Hoang for useful discussions, suggestions and also our students Dang-Khoa Nguyen-Nhat, Anh-Duy Pham-Lu for supporting in collecting the experimented datasets, some related resources and performing a few initial experiments.

Authors' contributions

H.-T. D studied the past studies, reviewed, tested data augmentation techniques and wrote this manuscript. T.-A. N-T reviewed the preprocessing techniques, apart of data augmentation techniques and collected the datasets for the experiments. Both authors read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
For four datasets of the experiments, Dataset 1 got from a contest of Vietnamese sentiment analysis prediction in comments (https://www.aivivn.com/contests/1), dataset 2 was built by B.-T. Nguyen-Thi [20] , dataset 3 and dataset 4 are getting from train and test datasets obtaining by streetcodevn.com (https://streetcodevn.com/blog/dataset?fbclid=IwAR28leEOyqBbLlgUdu29fETpfd-UX2QKzmnQAkLml9HIHWxMy-HKePiMjAo).

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Faculty of Information Technology, Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam. [2] Department of Fundamental Studies, Ho Chi Minh City Open University, 97 Vo Van Tan, Ward 6, District 3, Ho Chi Minh City, Vietnam.

Received: 21 March 2020 Accepted: 12 October 2020
Published online: 06 January 2021

References

1. Hussein DME-DM. A survey on sentiment analysis challenges. J King Saud Univ Eng Sci. 2018;30(4):330–8.
2. Medhat W, Hassan A, Korashy H. Sentiment analysis algorithms and applications: a survey. Ain Shams Eng J. 2014;5(4):1093–113.
3. Soleymani M, Garcia D, Jou B, Schuller B, Chang S-F, Pantic M. A survey of multimodal sentiment analysis. Image Vis Comput. 2017;65:3–14.
4. Symeonidis S, Effrosynidis D, Arampatzis A. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. Expert Syst Appl. 2018;110:298–310.
5. Effrosynidis D, Symeonidis S, Arampatzis A. A Comparison of Pre-processing Techniques for Twitter Sentiment Analysis. In: Kamps J., Tsakonas G., Manolopoulos Y., Iliadis L., Karydis I. (eds) Research and Advanced Technology for Digital Libraries. TPDL. Lecture Notes in Computer Science, vol. 10450. Cham: Springer; 2017.
6. Fernández-Gavilanes M, Àlvarez-López T, Juncal-Martínez J, Costa-Montenegro E, González-Castaño FJ. "GTI: An Unsupervised Approach for Sentiment Analysis in Twitter," in Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver; 2015. pp. 533–538.
7. Singh T, Kumari M. Role of text pre-processing in Twitter sentiment analysis. Procedia Comp Sci. 2016;89:549–54. https://doi.org/10.1016/j.procs.2016.06.095.
8. Jianqiang Z, Xiaolin G. Comparison research on text pre-processing methods on Twitter sentiment analysis. IEEE Access. 2017;5:2870–9. https://doi.org/10.1109/ACCESS.2017.2672677.
9. AL-Sharuee MT, Liu F, Pratama M. Sentiment analysis: an automatic contextual analysis and ensemble clustering approach and comparison. Data Knowl Eng. 2018;115:194–213.
10. Fernández-Gavilanes M, Juncal-Martínez J, García-Méndez S, Costa-Montenegro E, González-Castaño FJ. Creating emoji lexica from unsupervised sentiment analysis of their descriptions. Expert Syst Appl. 2018;103:74–91.
11. Wang H, Castanon JA. "Sentiment expression via emoticons on social media," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara. 2015; pp. 2404-2408, https://doi.org/10.1109/BigData.2015.7364034.
12. Sennrich R, Haddow B, Birch A. "Improving Neural Machine Translation Models with Monolingual Data," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, vol 1: Long Papers, Berlin. 2016; pp. 86–96, https://doi.org/10.18653/v1/P16-1009.
13. Sugiyama A, Yoshinaga N. "Data augmentation using back-translation for context-aware neural machine translation," in Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019), Hong Kong. 2019; pp. 35–44, https://doi.org/10.18653/v1/D19-6504.
14. Fadaee M, Bisazza A, Monz C. "Data Augmentation for Low-Resource Neural Machine Translation," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol 2: Short Papers. Vancouver. 2017; pp. 567–573, https://doi.org/10.18653/v1/P17-2090.
15. Kobayashi S. "Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans. 2018; pp. 452–457.
16. Azad HK, Deepak A. Query expansion techniques for information retrieval: a survey. Inf Process Manage. 2019;56(5):1698–735.
17. Şahin GG, Steedman M. "Data Augmentation via Dependency Tree Morphing for Low-Resource Languages," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels. 2018; pp. 5004–5009. https://doi.org/10.18653/v1/D18-1545.
18. Wei J, Zou K. "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification," in ICLR 2019-7th International Conference on Learning Representations, 2019.
19. Kim K. An improved semi-supervised dimensionality reduction using feature weighting: application to sentiment analysis. Expert Syst Appl. 2018;109:49–65.
20. Nguyen-Thi BT, Duong HT. A Vietnamese sentiment analysis system based on multiple classifiers with enhancing lexicon features. In: Duong T, Vo NS, Nguyen L, Vien QT, Nguyen VD, editors. Industrial networks and intelligent systems INISCOM, vol. 293., Lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications EngineeringCham: Springer; 2019.
21. Nguyen-Nhat D-K, Duong H-T. One-Document Training for Vietnamese Sentiment Analysis. In: Tagarelli A, Tong H, editors. Computational Data and Social Networks, vol. 11917. Cham: Springer International Publishing; 2019. p. 189–200.

22. Xia R, Xu F, Zong C, Li Q, Qi Y, Li T. Dual sentiment analysis: considering two sides of one review. IEEE Trans Knowl Data Eng. 2015;27(8):2120–33. https://doi.org/10.1109/TKDE.2015.2407371.
23. Xia M, Kong X, Anastasopoulos A, Neubig G. Generalized Data Augmentation for Low-Resource Translation, in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence. 2019; pp. 5786–5796. https://doi.org/10.18653/v1/P19-1579.
24. Duong H-T, Truong Hoang V. "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), Phuket. 2019; pp. 23–28, https://doi.org/10.1109/KST.2019.8687509.

Publisher's Note