International Journal of Scientific and Technical Research in Engineering (IJSTRE)
www.ijstre.com Volume 6 Issue 5 ǁ Sept-Oct 2021.
ISSN: 2581-9941

Challenges In Classification Of Vietnamese Sentiment

Hoa Tran Thi Thieu
Ha Tinh University, Viiet Nam
tranthithieuhoa@gmail.com

Abstract: *Sentiment classification is a field of study in text classification, natural language processing, and data mining. Sentiment classification is also a text classification process, in which a document is analyzed the content that users express their views on an object or a certain content, then give decide whether that content belongs to the group of positive or negative views or neutral views.*

On over the world, there have been many effective studies on the problem of opinion classification, most commonly on written documents in English. For Vietnamese language, there are also many reseraches on this field, but the effect is not really high. Due to the features and specific characteristics of the Vietnamese language with diverse grammar, many polysemes, and different meanings in different contexts, these form objective and subjective causes in the classification process. In this study, the author analyzes and conducts experiments on the data, thereby points out the challenges and difficulties in dealing with the problem of sentiment classification in Vietnamese language.

Keywords - *Sentiment classification Vietnamese , text classification, sentiment analysis.*

I. INTRODUCTION

Today, when the development of technology 4.0 has penetrated into every corner of human life along with the increasing demand for spiritual life, users often express and share their views and sentiments through the activities of experiencing with an entity or an object. Those sentiments are expressed and showed through text comments on websites, social networking sites, blogs or forums, ...etc. In addition, those views and sentiments are also hidden in contextual textual contents that simply cannot be recognized by reading or extracting. Management orientation based on such information, managers, businessmen and other social objects want to get information about users' views to implement plans for their activities.

Analysis sentiment is the research process that calculates people's opinions, attitudes, feelings and opinions towards an entity, and then processes human opinions about that entity, analyses viewpoints, identifies trends of opinion expressed in a document and analyzes it. Entities can represent individuals, organizations, events, or topics. Issues considered will be covered by the most views. The goal of sentiment analysis is to find opinion trends, determine which trend expressed opinions are, and thereby categorize their polarizing opinions toward that entity.

To perform sentiment classification, it is common to map a text to a known topic in a finite set of topics based on the text's semantics. In terms of text organization structure, the text consists of a set of related words that create the semantic content of the text. Words of each text are very diverse and rich due to the characteristics of polyseme and diverse words in the language. In Vietnamese documents, it is possible that a document has a small number of words, but the number of words that need to be revised is high, because it must include all the words under consideration. Then to analyze the text, it is necessary to read the content and analyze it, then use an algorithm to classify the text. Currently, researchers around the world have focused on researching many fields related to sentiment such as classification techniques, techniques for selecting features for text in classification, and processing problems of natural language nature through machine learning. Most of the research related content is popular on natural English language. For Vietnamese, this is a problem with many challenges and difficulties to solve due to the characteristics of the Vietnamese language. We will study this issue in this article.

II. LEVEL OF SENTIMENT CLASSIFICATION

The Since sentiment classification is a special case of text classification, sentiment classification can be performed at different levels such as word level, phrase level, sentence level and document level. For document level, sentence level or phrase level, they do not provide the necessary detailed sentiment on all aspects of the entity, but they are necessary when constructing application. Therefore, in order to classify sentiments in practice, we need to rely on the word level.

The word level is the level used to analyze and classify sentences and whole texts. The word level defines and involves the analysis of the polarity of words. There are two common methods used to classify sentiments at the word level, namely the lexical method (Lexical) [1] and the corpus-based method (Corpus) [2].

The phrase level involves polarizing phrases such as noun phrases, verb phrases, prepositional phrases, etc. To conduct the classification, the terms are first analyzed, then use those phrases and perform a contextual classification such as negative, positive, neutral or both. Finally, through the statistical method to extract the directional semantics, use the lexical score from the dictionary to predict the polarity of the phrase in the sentence. Machine learning techniques together with feature selection methods are used to polarize the content [3].

Sentence level is used to classify sentiment at sentence level to classify sentiments expressed in each sentence. Firstly, determine whether the sentence is subjective or objective. If the sentence is subjective, then the classification will determine whether the sentence expresses a positive or negative sentiment. The polarity of each subjective sentence was determined by applying word methods. The process of classifying each sentence is examined using machine learning to analyze the polarity of a phrase and combine them by combining the effects of conjunctions to determine the polarity of that sentence. .

First, use the tagging method to extract adjectives and adverbs, then use information retrieval and information reciprocity techniques to evaluate the sentinental orientation of the extracted phrases. Finally calculate the average semantic orientation for the phrases to value them through the obtained accuracy.

The document level [4] aims to categorize sentiments for a document as positive or negative. To do this we consider the entire document as a basic unit of information. Processing sentiment classification for a document consists of three steps. First, using the tagging method to extract adjectives and adverbs, then using information retrieval and information reciprocity techniques to evaluate the emotional orientation of the extracted phrases. Finally calculate the average semantic orientation for the phrases to value them through the obtained accuracy.

III. SOME CHALLENGES IN SENTIMENT CLASSIFICATION OF VIETNAMESE LANGUAGE

1. Challenges in Vietnamese text preprocessing

Vietnamese language is classified as an isolated type, that is, each sound (syllable) is pronounced separately, without distortion, monosyllable and expressed in a script, this feature is showed clearly in all including phonetics, lexical, and grammar. The basic difference between English and Vietnamese is that Vietnamese word and English word is different both in terms of lexicalization and morphology. According to author Dinh Dien [5], "Words are composed of morphemes and Vietnamese words are composed of Vietnamese morphemes". Due to the rich and diverse characteristics of Vietnamese, in order to classify sentiments in Vietnamese, at the present, researchers do not consider non-standard cases of Vietnamese texts, but only deal with key issues such as:

- Normalization of lowercase: In Vietnamese, lowercase and uppercase have the same meaning, but depending on the context and grammatical structure, the text content will be shown in uppercase or lowercase. Therefore, if you want to process documents, this normalization must be done in the preprocessing stage, the techniques of standardizing uppercase and lowercase letters in Vietnamese are not too difficult, but it takes a long time because there are many different context usage and grammatical rules. This also affects the time in Vietnamese classification.

- Normalization of punctuation marks, processing punctuation marks at the end of sentences, between sentences, and chains containing many continuous full stops: Vietnamese grammar uses many types of punctuation marks such as a full stop ".", exclamation mark "!", colon ":", semicolon ";" , etc. along with the

characteristics of sentence structure in the text, the number of these punctuation marks is very large in the structure of paragraphs. The semantics of the text are also different when different punctuation marks are attached, which makes the processing and identifying of text meaning more difficult in sentiment classification.

- Dealing with cases of abbreviations and special characters: The habit of abbreviating and using special characters in Vietnamese is quite a lot, the abbreviations are rich, diverse, without standards and according to the feelings and conventions of each user. These cases make the processing of Vietnamese documents very complicated and time consuming. It is necessary to define all the abbreviations, identify them in different contexts and make a list of special words to process. User's daily habit of abbriviation makes text processing and semantic guessing very time consuming and difficult in different contexts. This factor also affects the time to classify Vietnamese in general and causes difficulties in classifying sentiments in Vietnamese in particular.

- In addition, when pre-processing documents also handle cases of reduplication, handling cases such as law on language. The preprocessing of text data is carried out with paragraph segment, sentence segmentn and spelling normalization.

2. Challenges in separating Vietnamese words

Vietnamese is composed of extended Latin characters, Vietnamese has a common feature with Southeast Asian phonographic languages so it is difficult to define the boundaries between words. In Vietnamese, space (space) does not mean separating words, but only means separating syllables from each other, the separation of words must also be based on other factors such as content, context. , etc. Therefore, in order to classify Vietnamese documents, it is necessary to conduct word segment. Word segment is a word processing process that determines the boundaries of words in a sentence. Segmenting Vietnamese words is a difficult task and has many obstacles to carry out, while the results of text classification depend a lot on the results of the word segment process in Vietnamese documents.

The word segment problem in general and the word segment problem in Vietnamese in particular is a difficult and complex problem that needs solving many tasks. At present, there are basically three main approaches to the word segment problem: the dictionary-based approach and the statistical-based approach, and a hybrid approach based on both of the above approaches. The problem of word segment in Vietnamese has been studied a lot, each method has its own advantages and disadvantages, uses different techniques and gives relatively positive results. However, all methods have to deal with the basic problems when performing word segment. One of the important and complex problems that the word segment problem needs to solve is ambiguity handling. The handling of ambiguity in Vietnamese word segment is divided into two types: Overlapping Ambiguity and Combination Ambiguity. Due to the characteristic of Vietnamese language that the minimum vocabulary is mostly monosyllabic words (one syllable), so the phenomenon of Overlapping Ambiguity occurs commonly. This is also the reason that when processing word separation, the overlapping ambiguity must be handled first, the overlapping ambiguity process takes a lot of time, effort and even processing technique which affect the quality of classification.

In addition, another difficulty that also needs to be well solved is the identification of unknown words (including Vietnamese or foreign proper names), if this information is not understood, the word segment will be difficult. Therefore, this problem also needs to be solved effectively because in the text there are not only the existence of pure words in the dictionary, but also other units of information.

3. Challenges in Vietnamese dictionary and corpus

Monolingual dictionaries are an extremely important and necessary resource in text classification. The process of sentiment classifying in general and separating word segment in particular will use information taken from the dictionary to analyze morphology (analysis of words/phrases) and analyze the meaning of words. Therefore, the dictionary plays a very important role in preprocessing or segmenting Vietnamese words, but currently Vietnamese does not have a large and complete dictionary to meet all the words in Vietnamese. This difficulty is also the disadvantage of the Maximum Matching method [6], in this method its accuracy depends entirely on the completeness and accuracy of the dictionary, it requires a big dictionary enough to perform comparisons during word segment. Another difficulty also related to the dictionary is that it cannot solve the problem of ambiguity in word segment and cannot recognize unknown words when the dictionary is not big enough and cannot covers all words because only words that exist in the dictionary are properly segmented.

For Vietnamese, up to now, there has not been a standard corpus which is large enough and tested and evaluated to be able to be used for all tests. This difficulty has been shown in the Weighted finit–state

Transducer – WFST method [7]. The difficulty of this method is the construction of a standard corpus, which is very complicated and time consuming. As for the MMSeg word segment method [8], this is a widely used and highly effective method, with less difficulty but it is essential to know how to apply the rules flexibly in dealing with the problem of ambiguity in word segment, use dictionaries flexibly, which is not accessible for everyone. In addition, the word segment methods all use machine learning techniques, so they depend on sufficient training data, which is a difficult problem in practice such as the recognition of unknown words (including proper name words in Vietnamese or foreign languages), if this information is not understood, the word segment will be affected and the results will not be high.

4. Vietnamese is a sentimental language

Another difficulty in classifying Vietnamese's sentiment is that Vietnamese is a language rich in sentiment and polyseme, and Vietnamese semantics is very rich and diverse. The expression of each individual's point of view is also diverse, with different ways of writing in terms of words and expressions and the way people use language. Due to the polysemy in language, in fact a word in a comment can express a positive opinion in one situation but can be negative in another, especially in the case of metaphors, the sarcastic way of saying, it is easy to cause the misunderstanding. Sometimes a sentiment on an issue or part of a problem can also mislead the sentiment of a mining system. At the level of paragraph or text file expressing sentiments, there is a situation where there are many conflicting sentiments in the same paragraph or document, i.e. the same comment on the same issue but the article includes both positive and negative sentiments or contain many conflicting ones in the text. It is these problems that make it difficult to parse and mine the sentiments.

5 Remove the Stopwords

The word stopwords are function words or auxiliary words, adjuncts and formal words. In Vietnamese, there are many stopwords which are often used in all contexts; such words as “be”, “that”, “of”, “especially”, …etc, linking words, quantitative words “with”, “together” “each”, “every”, etc., are not distinguishable in classification. In addition, there are many other words that are also not valid in the text classification. The removal of these stopwords will effectively increase the text meaning and also reduce the large number of features in Vietnamese in Vietnamese text classification models.

6. Feature selection

Text features are the categories in the text. Feature selection aims to shorten the dimensionality of the feature space, the nature of the feature selection process is to reduce the dimensionality of the feature vector by removing unimportant feature components. In fact, one cannot consider all the words of the language, but rather uses a set of words drawn from a (large enough) set of documents under consideration texts(called a corpus).

For Vietnamese, Vietnamese features include Vietnamese language parameters such as number of words, number of syllables, uni-gram, bi-gram information, .. specifically with the research [9] has shown the Vietnamese parameters are as follows: number of words: 40.181 words. (The most used and widely used words); Number of syllables: 7.729 syllables. In which 81.55% of the syllables are also single words. 70.72% of compound words have 2 syllables. 13.59% of compound words have 3.4 syllables. 1.04% of compound words have 5 or more syllables. Thus, Vietnamese has a very large number of features, in order to classify sentiments, it is necessary to select appropriate features. In principle, it includes all words in the language, so with 40.181 words in Vietnamese, the number of spatial dimensions is very large, making the classification problem difficult to handle and unable to bring high efficiency. In fact, the features must be selected in order to shorten the dimensionality of the feature space by removing the unimportant feature components but still ensure the accuracy of the text content.

IV. VERIFICATION EXPERIMENT

In order to assess the complexity of the classification of sentiments in Vietnamese, we conduct an assessment of the sentiments of restaurant and hotel service users based on comments from websites [10-12] on two sentence level and document level.

Experimental data has 602 text files assembled for machine learning construction and testing. After word segment and word removal, the number of words is 52.132 words. The corpus to be modeled is a matrix containing TF*IDF of words of size 602×52132 elements. We use 70% of the acquired data as training data and use the remaining 30% as test data.

The experimental process was conducted in accordance with the sentiment classification process. First, the collected data was preprocessed text, then performed word segment. In our research experiments, we have used the Pointwise method [13] to solve the Vietnamese word segment problem. Next, we rely on the list of stopwords referred from the website to remove the stopwords. The process of extracting the feature set and representing the text is used formula to calculate the value based on the Terms Frequency and Inverse Document Frequency methods [14]. Finally, we use the algorithm Support Vector Machines - SVM [15] to solve for the trained system.

The classification result for the sentence level has an accuracy of 78% and the document level reaches 70%.

V. CONCLUSION

The problem of sentiment classification is always a difficult and complex problem due to many factors involved, especially with Vietnamese documents, the classification of sentiment has many obstacles. As we have analyzed above with the language features, objective and subjective difficulties related to the Vietnamese language classification process, the process of classifying Vietnamese sentiment faces many difficulties in different stages from preprocessing to word extraction or feature selection.

From the experimental results, the accuracy of sentence level classification is 78% higher than that of document level - 70% because the sentence classification with a short structure, quite clear and simple content. For document level classification, the document structure consists of many sentences combined, each paragraph of text represents different sentiments including positive and negative ones, even containing many conflicting sentiments making it difficult to analyze and categorize sentiments for the entire document. This shows that the document level sentiment classification process is quite complicated and has many difficulties.

With the relevant experimental results, it once again proves that the Vietnamese dictionary elements are incomplete, the corpus is not large enough, the context is not covered, and the objective causes from language are Vietnamese language is sentimental with many polyseme. The process of classifying categorical data falls into the case of ambiguous textual content, implications, slurs, words containing figurative and literal meanings, making the system difficult to classify. However, with the cooperation of the research community, we hope that these difficulties will soon be thoroughly resolved in the classification of Vietnamese sentiments in particular and the classification of texts in general.

REFERENCES

[1] M. Taboada, J. Brooke, "Genre-based paragraph classification for sentiment analysis," presented at the Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, London, United Kingdom, 2009.

[2] X. Wan, "Co-training for cross-lingual sentiment classification," presented at the Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1, Suntec, Singapore, 2009.

[3] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, Vancouver, British Columbia, 2005.

[4] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL", presented at the Proceedings of the 12th European Conference on Machine Learning, 2001.

[5] Đinh Điền. Building and exploiting an electronic English - Vietnamese bilingual corpus. Thesis of PhD in Comparative Linguistics. University of social sciences and humanities – Ho Chi Minh City. 2-2005.

[6] Chooi-Ling Goh, Masayuki Asahara, Yuji Matsumoto. (2004). Chinese Word Segmentation by Classification of Characters. In Proceedings of Third SIGHAN Workshop.

[7] Richard Sproat, Chilin Shih. Corpus – based Methods in Chinese Morphology and Phonology. Lecture notes for LSA Summer Institute, Santa Barbara. 2001.

[8] . Chih-Hao Tsai ," MMSeg: A Word Identification System for Mandarin Chinese Text Based on two Variants of the Maximum Matching Algorithm" 2000

[9] http://viet.jnlp.org/home

[10] https://docs.google.com/spreadsheets/d/1FlzllKpR_8ipxRmzlJl9JhwgME860_WM4UbHMGgM3ylU/edit#gid=1242070509

[11] https://www.booking.com/reviews.vi.html

[12] https://www.trivago.vn/

[13] Lưu Tuấn Anh, Yamamoto Kazuhide. Applying the Pointwise method to the word segment problem for Vietnamese. Natural Language Processing Laboratory Department of ElectricalEngineeringNagaoka University of Technology 940-2188, Nagaoka City, Niigata, Japan.

[14] . Yang, Y. and Pedersen ,J,O. , A comparative Study On Feature Selection in Text Categorization . In Proceedings of the 14th International Conference on Machine Learning(ICML), (1997), 412-420.

[15] J. Platt, "Sequential minimal optimization: A fastalgorithm for training Support Vector Machines", Technical Report MSR-TR-98-14, MicrosoftResearch, 1998