DOI: 10.15849/IJASCA.230720.21

Int. J. Advance Soft Compu. Appl, Vol. 15, No. 2, July 2023
Print ISSN: 2710-1274, Online ISSN: 2074-8523
Copyright © Al-Zaytoonah University of Jordan (ZUJ)

Sentiment Analysis by Lexical Analysis Combined with Machine Learning

Van Lam Ho, Vo Le Minh, Tran Xuan Viet, Nguyen Ngoc Dung, Trang Thi Ho

Department of Information Technology, Quy Nhon University, Vietnam.
e-mail: hovanlam@qnu.edu.vn

ExploraScience QuyNhon, Binh Dinh, Vietnam
e-mail: voleminh10t2 @gmail.com
Quyhoa National Leprosy Dermatology Hospital, Binh Dinh, Vietnam.
e-mail: thstranxuanviet@gmail.com;
Department of Information Technology, Quy Nhon University, Vietnam.
e-mail: nguyenngocdung@qnu.edu.vn
Department of Computer Science & Information Engineering TamKang University, Taiwan.
Email: hothitrang@mail.tku.edu.vn

Abstract

This paper aims to sentiment analysis users evaluate products using lexical analysis method combined with machine learning. This study analyze the emotions of users through analyzing their comments and evaluations for information posted or shared about services and products at the Explorascience QuyNhon. The first we retrieve the data of the user's product review comments and then build a Vietnamese emotional dictionary using vocabulary-based methods by calculating the semantic value of words or phrases in documents, finally, using a machine learning model to analyze and evaluate emotions with two problems: classifying sentences with feelings or without emotions and classifying sentences with positive or negative emotions. With input is a set of raw Vietnamese comments of users our methods will return outputs are Vietnamese comments which have been classified into three categories: without, positive or negative emotions. The input data will be a set of Vietnamese comments that are then evaluated and then put into processing Vietnamese errors with accents, processing emoticons, processing stop words collectively referred to as preprocessing. After the preprocessing has been standardized, the system begins to extract the characteristics of each sentence based on the emotion dictionary and the factors affecting the emotions in the sentence. From the characteristics obtained, subjective classification and emotional classification of comment sets to finally output the set of comments are classified into three categories: without emotions, with positive emotions or with negative emotions by machine learning model.

Keywords: *Sentiment analysis, lexical analysis, machine learning, classifying sentences, sentiment classifying.*

1 Introduction

With the growth of social media on the Internet such as forums, blogs, Facebook, Google plus, Twitter, Instagram; and social networks are increasingly influential not only with

Received 5 May 2023; Accepted 5 July 2023

businesses but also with society as a whole; so emotion analysis has developed rapidly and become the main field of study in natural language processing and applied to every fields in the real world and at the Explorascience QuyNhon is a case study. Besides, machine learning is a field of Artificial Intelligence, which is a technique that helps computers learn on their own without setting up decision rules. Normally, a computer program needs rules to be able to execute a certain task, but with machine learning, computers can automatically execute the task upon receiving input data. In other words, machine learning means that computers can think on their own like humans. Another approach argues that machine learning is a method of drawing lines that represents the relationship of a data set [2].

In this paper, we have combined word analysis in natural language processing with machine learning models [3] to solve the problem of detecting users' emotions through analyzing their comments and reviews for information posted or shared about the services, products at the Scientific Discovery Center.

To carry out this work, we collect data, process data and build a Vietnamese Emotion Dictionary to calculate the semantic value of words or phrases in the document, calculate semantic values based on a set of words and their semantic values and are called words emotional dictionary.

Currently, there is no officially published emotional dictionary for Vietnamese. We used the English emotional dictionary SO-CAL (Dictionaries for the Semantic Orientation CALculator) by Maite Taboada [4] and translated it into Vietnamese. The SO-CAL Emotional Dictionary [5] has about 6600 words divided into five small dictionaries: noun dictionary, verb dictionary, adjective dictionary, verb dictionary and intensifier. Each dictionary includes a list of emotional words and accompanying SO values.

Based on the constructed emotion dictionary, we build machine learning models [6] to the analysis of emotional evaluation relies on comments by classifying sentences with or without emotions and classifying sentences with positive or negative emotions.

To categorize whether a sentence contains emotions or not emotions (often referred to as subjective classification) is a problem that a sentence when written or spoken will usually have a certain speaking purpose: narrative (used to describe, tell or introduce a thing or thing), questioning (used to ask), asking (used to suggest, ask), exclamation (used to express emotions), ... so categorizing the purpose of the sentence would make subjective categorization easier and more accurate.

If the sentence contains only ordinary words that do not carry emotions and words that carry positive or negative emotions, then the classification of emotions will be easily solved. However, in practice, sentences also have negation words, words that increase the level of semantics (amplifiers), words that reduce the level of semantics (Downtoners), verbs that are defective,... and assessing the effect of these words on the emotions in the sentence, their combination with the emotional words in the sentence so that the most accurate conclusion is whether the sentence carries negative or positive emotions is also studied in this paper.

The remainder of this paper is organised as following. Section 2 introduces the job to construct the Vietnamese Emotion Dictionary using a vocabulary-based method by calculating the semantic value of words or phrases in a document. Section 3 show algorithm uising to build machine learning models to the analysis of emotional evaluation. The experimental results of using lexical analysis method combined with machine learning are described in Section 4. Section 5 is conclusion of the paper.

2 Construct The Vietnamese Emotion Dictionary

The SO-CAL dictionary [4] includes 5 small dictionaries: noun, verb, adjective, adverb and reinforcement dictionary. The number of words in the dictionary of nouns, verbs, adjectives and adverbs is 1142 words, 903 words, 2252 words, 745 words, respectively, and with each word is accompanied by an integer representing the corresponding SO value in the dictionary, ranges from -5 for every negative to +5 for very positive and none of the words have a SO value of 0. The words in this dictionary are taken from a variety of sources and the 3 biggest are:

Epinions 1: a collection of 400 texts on 8 different topics: books, cars, computers, cookware, hotels, movies, music, and phones, and equally divided into negative half and half positive. A subset of 100 texts containing 2000 movie comments in the Polarity dataset (Pang, Lee, and Vaithyanathan 2002; Pang and Lee 2004, 2005).

Positive and negative words from the General Inquirer dictionary (Stone et al. 1966; Stone 1997). The enhanced word dictionary includes more than 200 words divided into 2 categories: words that increase the level of semantics (amplifiers) and those that reduce the level of semantics (downtoners).

Table 1. Some Enhanced Words

Enhanced Words		**Level**
English	Vietnamese	
Slightly	Hơi	-0.5
somewhat	một chút	-0.3
Pretty	Khá	-0.1
Really	thật sự	+0.15
Very	rất	+0.25
extraordinarily	cực kỳ	+0.5
(the) most	nhất	+1

For example: The word “sleazy” has an SO value of -3, then “pretty sleazy” (quite sleazy) has an SO value of -3*(1 – 0.1) = -2.7. The word "excellent" has an SO value of 5, then "most excellent" has an SO value of 5 * (1 + 1) = 10.

In with negative words are divided into 2 types: Switch negation words such as not (no), never (never), nobody (nobody),… simply reverses the polarity of a word or, more simply, changes the sign of the value. SO of the word.

For example: “Good” has an SO value of +3, then “not good” has an SO value of -3.

Shift negation: If using Switch negation, “excellent” will have an SO value of 5, “not excellent” will have an SO value of -5. Similarly "not good" will have SO value of -3. In fact, “not excellent” will have a more positive emotion than “not good”. To avoid that, Shift negation will change the SO value of the negation to match reality.

For example: Cruise is not good (emotional value: 4 - 4 = 0), but I must admit he is not mean (emotional value: -3 + 4 = 1).

Compare the performance of different dictionaries with SO-CAL dictionaries, we have the Table 2 performance comparison table of different dictionaries with SO-CAL dictionary.

Table 2. Performance comparison table of different dictionaries with SO-CAL dictionary

Dictionary	Performance on test suite				
	Epinion 1	Epinion 2	Movie	Camera	Overall
Google-Full	62.00	58.50	66.31	61.25	62.98
Google-Basic	53.25	53.50	67.42	51.40	59.25
Maryland-Full-NoW	58.00	63.75	67.42	59.46	62.65
Maryland-Basic	56.50	56.00	62.26	53.79	58.16
General Inquirer-Full	68.00	70.50	64.21	72.33	68.02
General Inquirer-Basic	62.50	59.00	65.68	63.87	64.23
SentiWordNet-Full	66.50	66.50	61.89	67.00	65.02
SentiWordNet-Basic	59.25	62.50	62.89	59.92	61.47
Subjective-Full	72.75	71.75	65.42	77.21	72.04
Subjective-Basic	64.75	63.50	68.63	64.83	66.51
SO-CAL-Full	80.25	80.00	76.37	80.16	78.74
SO-CAL-Basic	65.50	65.25	68.05	64.70	66.04

To translate the SO-CAL English dictionary, we used a combination of Viettien Dictionary and Google Translate.

Viettien Dictionary was first published by Nguyen Viet Khoa - Institute of Foreign Languages, Hanoi University of Science and Technology in August 2010 and the latest update as of March 2015 is version v4.0b published by Viettien Dictionary in July 2014 on the Mac OS platform is also the version that I use. As of July 2014, Viettien's English-Vietnamese dictionary has more than 390,000 words and Google Translate: This service has supported most languages in the world, including Vietnamese. The initial translation quality of Google Translate is not good. But because it is equipped with an interactive feature that helps people to change the meaning of the word to the best fit, the quality has improved. Google's translation speed is very good compared to other similar online services for Vietnamese and especially in the ability to translate long texts.

The dictionary translation process is performed sequentially from the beginning to the end of each dictionary in the SO-CAL dictionary. Cases occurring during translation:

An English word has only one Vietnamese meaning: we will add the Vietnamese meaning and emotional value of the word to the SO-CAL Vietnamese dictionary.

For example: The phrase “mega-star” (+3) is translated into Vietnamese as “super star” (+3). The phrase “queen-sized” (+3) is translated into Vietnamese as “large size” (+3).

An English word with many Vietnamese meanings: we will choose the most commonly used and concise Vietnamese meaning to add to the SO-CAL Vietnamese dictionary. If in the remaining meanings there is a short meaning and is synonymous with the previously selected meaning, we also add that meaning to the SO-CAL Vietnamese dictionary. The selected meanings are usually short from one to three words and after being added to the SO-CAL Vietnamese emotional dictionary, they will retain the emotional value of the English words translated in the SO dictionary CAL English.

For example: The word “exceptional” (+5) is translated into Vietnamese as “excellent” (+5) and “outstanding” (+5). The word “glorious” is translated into Vietnamese as “vẻ vang” (+5) and “vinh quang” (+5).

An English word or phrase that is not in the English-Vietnamese dictionary or is a combination of many words leading to a Vietnamese meaning that is too long: When there is an English word that is not in both of the above English-Vietnamese dictionaries then we will skip that English word. And when we encounter an English word or phrase

that is combined by many words leading to a Vietnamese meaning that is too long, we will try to shorten their meaning as short as possible. If that doesn't work, I'll omit the word or phrase.

For example: The word “ritz-carlton” is not in both English-Vietnamese dictionaries above, so we omitted this phrase. The word “all-too-rare” translated as “tất cả quá hiếm” was too long, so we omitted this phrase. The word “well-fitting” (+4) was translated into Vietnamese as “fit” (+4).)

Added Vietnamese word or phrase already in SO-CAL Vietnamese dictionary: When there is an added Vietnamese word or phrase already in SO-CAL Vietnamese dictionary, we will remove that word.

For example: The word “perfect” means “perfect” but it has the word “impeccable” which is translated as “perfect ” first. So the phrase "perfect" of the word "perfect" is not added to the SO-CAL Vietnamese dictionary anymore.

In addition, to match the Vietnamese grammar and the concise writing of the comments on social networks, we added some words and phrases with fewer words but still synonymous with the words or phrases in the list. SO-CAL Vietnamese dictionary.

For example: we see that the word “may” has fewer words but is still synonymous with the word “luck” (+2), so we add the word “may” (+2) to the SO-CAL Vietnamese dictionary.

After translating the SO-CAL English dictionary into Vietnamese, we obtained the SO-CAL Vietnamese dictionary including 5 small dictionaries: Noun dictionary (1544 words), verb dictionary (1105 words), adjective dictionary (2357 words), adverb dictionary (749 words) and intensifier dictionary (185 words).

Table 3. Some words in SO-CAL Vietnamese dictionary

Noun	**Emotional value**
hoàn hảo	5
lộng lẫy	4
chiến thắng	3
phước lành	2
độc lập	1
tội phạm	-1
điểm yếu	-2
tai ương	-3
thảm họa	-4
kỳ quái	-5
Verb	**Emotional value**
tôn kính	4
hoan hỉ	4
thành công	3
sáng tạo	2
Tăng	1
vùi dập	-1
xấu hổ	-2
nguyền rủa	-3
Ghét	-4
ghê tởm	-5
Adjective	**Emotional value**

tuyệt vời	5
cao cấp	4
bổ ích	3
chặt chẽ	2
hợp lý	1
cũ	-1
đần độn	-2
bẩn	-3
tai hại	-4
thảm khốc	-5
Adverb	**Emotional value**
thú vị	5
huy hoàng	4
giỏi	3
tươi	2
sạch	1
kỳ quặc	-1
thô	-2
kém cỏi	-3
tàn bạo	-4
khiếp	-5
Intensifier	**Emotional value**
ít	-1.5
chút ít	-0.9
hơi	-0.5
khá	-0.2
chắc	0.2
siêu	0.4
hoàn toàn	0.5
nhất	1

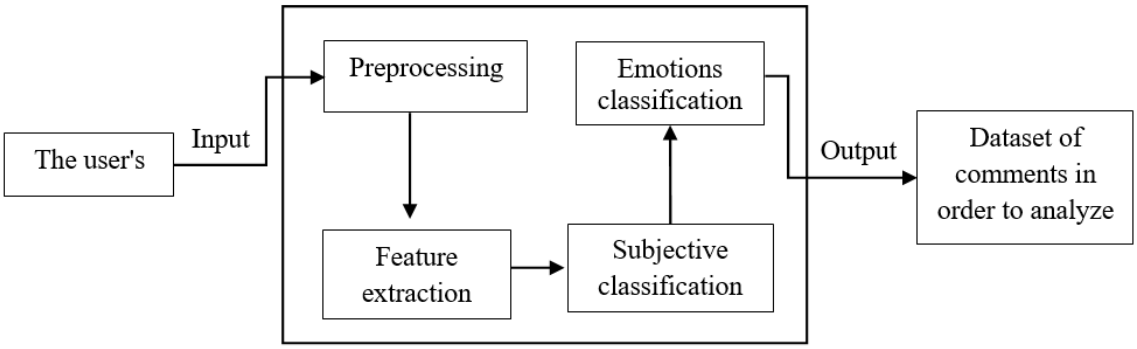
3 Machine Learning Model

3.1 The overall model

- Input: The raw Vietnamese user comments.

- Emotion analysis system: Preprocessing, feature extraction, subjective classification and Emotions classification.

- Output: The Vietnamese comments after being analyzed by the emotion analysis system are categorized into three types: No emotion, positive emotion, and negative emotion.



First, the input data is a set of raw Vietnamese comments. These comments are considered "raw" because before they can be used, we need to address several issues such as handling Vietnamese diacritics, processing emoticons, dealing with stop words, etc., collectively known as preprocessing. After the preprocessing stage, we obtain a set of standardized comments. The system then starts extracting features from each sentence based on the emotion dictionary and factors influencing emotions in the sentence. Using these extracted features, subjective classification and emotion classification of the comment set are performed, ultimately resulting in a set of comments categorized into three types: no emotion, positive emotion, and negative emotion.

3.2 Subjective Classification Model

Subjective classification is a necessary first step to emotional analysis. In this part, the work to be done is to evaluate and classify the data after preprocessing into 2 classes: subjective class and objective class.

The subjective classification is mainly based on the matching method with the emotional dictionary. Therefore, we choose the word matching method with SO-CAL emotion dictionary. A subjective (emotional) sentence usually has an emotional word.

For example: “The blue house” is an objective sentence because it has no emotional words in it. “Beautiful house” is a subjective sentence because it has the emotional word “beautiful”.

This is the most basic and simplest method for classifying a sentence as subjective or objective [7, 8]. Accordingly, the selection of the best features to evaluate subjective sentences is studied to get optimal results.

Beside, we have there exceptions such as:

Sentence classification method based on emotional words is the main method to classify subjective sentences. However, the level of accuracy is not high because there are exceptions, which are cases where sentences contain words that contain emotions but do not express emotions. Specifically, they are interrogative and conditional sentences.

Interrogative Sentences: The basic feature of interrogative sentences is that they often contain the words "what", "how", “why". These sentences even have words containing emotions, but they are still sentences without emotions.

For example: “Why are you wearing such unrefined clothes?” It is a question and has no emotion. Although the sentence contains the word "subtle" emotion, in fact this sentence has no emotion at all. It is just a question that the speaker asks the listener to answer.

Conditional sentences: The characteristic of conditional sentences is that there are often words: “if…then…”, “if only…then…”… In both cases, the sentences contain no emotion even though they contain emotional words.

For example: “If it rains tomorrow, I will be very sad.”. In the sentence with the word "very sad" the SO value is (-2)*(1+0.2) = (-2.4) but the above sentence is unlikely to take place in reality, but only the speaker's speculation. It may rain tomorrow, but the speaker is not necessarily sad. So the sentence will have no emotion.

"If you study well, I will let you play." In the sentence with the word "good", the SO value is (+3) but the above event did not happen. So the sentence above will have no emotion.

In addition to the cases above, we have noticed that a sentence contains emotions if it is a long sentence. Usually, short sentences are just nouns (people, things, places, etc.), verbs or adverbs, and these sentences usually don't contain emotions. When the speaker has intended to express a long sentence, most will put an emotional element in it. However, assessing how long a sentence is long enough and emotional requires more experimentation and separate studies on this issue. Within the scope of the thesis, we choose the value of 5 word units as a landmark for a long and short sentence.

From the tagger file and the Vietnamese SO-CAL dictionary, we extract features to evaluate a sentence with or without emotion, we choose the following features:

Feature number 1: The number of words in the sentence is the number of words in a sentence also shows the feelings that the speaker and writer want to express to the listeners and readers. If the word count is normally large that would be an emotional sentence because the speaker, the writer has invested a considerable amount of effort and is clearly interested in the topic in question. On the contrary, if the number of words is too small, it may be a noun for people, things, etc.

Features No. 2, 3, 4 and 5: the total emotional value of the words: adjectives, adverbs, nouns and verbs in the sentence. The emotional value in the sentence depends on the word type and the emotional value of that word is matched with the SO-CAL Vietnamese dictionary. We found that the emotional value in a sentence mainly depends on the following types of words: adverbs, adjectives, nouns and verbs. Accordingly, corresponding to the total emotional value of each type of word we choose into a feature.

The total emotional value of the adverb in the sentence. After being labeled, the adverb tags are checked and matched with the adverb dictionary in the SO-CAL Vietnamese dictionary. If they are the same, this value is added to the total adverb emo value. If there is no adverb or dictionary matching in the sentence, this value defaults to 0.

Exactly the same for adjectives, nouns and verbs. These word tags match the corresponding dictionary in the Vietnamese SO-CAL dictionary. If there are no matches or the sentence does not contain these word types, the default value will be 0.

Feature number 6: the total emotional value of the sentence. This feature represents the total emotional value of the sentence. The value of this feature is basically the sum of the four features above that we built. Although they are related and this value seems redundant, in fact this summation is necessary because if the sum of the above values is zero, the subjective assessment will not be certain. correct. In addition, the emotional value in a sentence depends not only on the word containing the emotion, a subjective sentence also depends on its type of sentence. If it is an interrogative sentence or a commanding sentence, it has absolutely no emotional value. Therefore, the total

emotional value of a sentence can also be zero if the sentence belongs to one of the two types of sentences above.

Subjective classification algorithm:

Input: tagger file and Vietnamese SO-CAL dictionary.

Output: The file has a vector structure, with each line being a feature vector.

Application methods:

For each sentence in the dataset, extract the values

1) Total words

2) The total emotional value of the adjectives.

3) The total emotional value of the adverbs.

4) The total emotional value of the nouns.

5) The total emotional value of the verbs.

6) Emotional value of the whole sentence:

If the sentence belongs to the question or the current sentence, return 0. Otherwise, return the sum of the numeric features 2, 3, 4 and 5. Returns the feature vector.

From the result file of the feature extraction process above. We use the classification method using the SVM algorithm with the collected training data to conduct the classification. The return result of this process is the result of classifying the sentence into two classes: subjective and objective.

The following examples detail the program feature extraction process for a particular sentence.

For example: The sentence "Cô ấy vừa đẹp mà vừa học giỏi nữa." feature will be extracted and return the following values: “1:9.0 2:7.0 3:0.0 4:0.0 5:0.0 6:7.0”. These values have the following meanings:

Feature number 1 is the number of words in the sentence. Here the value is 9.0. Feature number 2, 3, 4 and 5 are the sum of emotional values of adjectives, adverbs, nouns and verbs in the sentence, respectively. The total emotional value of the adjectives in the sentence is 7.0, including: "beautiful" with the value (+4) and "good" (+3). The total emotional value of words: adverbs, nouns and verbs in this sentence is zero because the sentence has no adverbs, nouns and verbs. Feature 6 is the sum of emotional values of all words in features 2, 3, 4 and 5. This value is 7.0 including: adjective (+7.0), adverb (0), noun word (0) and verb (0).

From the above characteristics, the sentence " Cô ấy vừa đẹp mà vừa học giỏi nữa." is a subjective sentence that contains emotion.

For example: The sentence “Nếu học tốt hơn thì tôi sẽ đăng ký kỳ thi tới.” after feature extraction, the result will be as follows: “1:10.0 2:3.0 3:0.0 4:0.0 5:0.0 6:0.0”. Although the above sentence has the emotional value of the adjective (+3), the total emotional value is (0) because this is a conditional sentence. Therefore, this is an objective sentence that does not contain emotions.

With the manual classification test dataset, the SVM classification method and the training data set, we test the accuracy of the subjective classification method. The results of the accuracy assessment are given in the Table 4.

Table 4. Accuracy assessment results

Number	Subject	Results (accuracy: %)
1	Education	86.7%
2	Movie	87.1%
3	Sport	67.7%
4	General	89.8%

3.3 Emotions Classification Method

After identifying emotional sentences, we rely on SO-CAL Vietnamese emotion dictionary and features extracted from Vietnamese sentence characteristics to calculate the emotional value of sentences. The calculated value helps to classify emotional sentences into sentences with positive emotions or sentences with negative emotions.

Sentiment word is the element that has the greatest influence on the emotional value of a sentence and is often used to express negative or positive emotions. For example: the words “tốt”, “tuyệt vời”, “đẹp” are words containing positive emotions and “xấu xí”, “kinh khủng”, “tệ hại” are words containing negative emotions. Beside the individual words, there are also emotional phrases such as “không thể tin được”, “như một giấc mơ”, etc called an emotional dictionary.

The simplest way to calculate the emotional value of a sentence is to sum the emotional value of the emotional words in that sentence.

For example: “Anh ấy thông minh và đẹp trai”. The word "thông minh" has an SO value of (+4) and "đẹp trai" has an SO value of (+4) so the total SO value of the sentence is (+8). “Chiếc áo này hợp thời trang” The sentence “Chiếc áo này hợp thời trang” has only one phrase that carries the feeling of "hợp thời trang", so the total SO value of the sentence is also equal to the SO value of this word (+2).

In addition, words with emotional value are influenced by reinforcement words such as: "đẹp", "hơi đẹp", "rất đẹp" and "đẹp nhất". If we only based on the emotional dictionary, the above words and phrases will have the same SO value. But in reality they carry positive emotions and ranked in ascending emotional value as "hơi đẹp", "đẹp", "rất đẹp", “đẹp nhất”.

Morever, classification emotion we are easy to confuse positive and negative. Some words have the ability to change the polarity of emotional words or phrases such as “không”, “không được”, “không phải”, “không bao giờ”, etc. For example: the word “tốt” carries positive emotions. “không tốt” carries a negative emotion.

To solve the above problems, it is necessary to further analyze other features of the sentence. Each feature will gradually solve each specific problem.

The emotional value of a sentence depends on the reinforcement word, the reinforcement word is divided into two types: increasing the level of semantics (amplifiers) and reducing the level of semantics (downtoners) [9]. In SO -CAL also adds enhanced word dictionary. Words affected by reinforcement words will have an emotional value that changes depending on whether the value increases or decreases the semantic level of that reinforcement.

For example: The word "mệt mỏi" carries the value SO (-3). If it is preceded by the strengthening word “hơi” (-0.5) then the SO value of “hơi mệt mỏi” is: (-3)*(1-0.5) = (-

1.5). The word "đẹp" has SO value of (+4), then "rất đẹp" has SO value of: (+4)*(1+0.2) = (+4.8); The word "giỏi" has SO value of (+3), then "giỏi nhất" has SO value: (+3)*(1+1) = (+6).

Besides, the emotional value of the sentence depends on the negative word, the emotional value is influenced by the negative word, changing the emotional value of the emotional word. When speaking or writing, we often use negative words including: "không", "không được", "không phải", etc. to express a level of emotion opposite to that of the emotional word that follows. from that negation.

Therefore, for emotional words preceded by a negative word, the emotional value of that word will be reversed or easier to understand than changing the sign of the emotional value of the word.

For example: The word “tốt” has SO value of (+3), then “không tốt” has SO value of (-3); The word "bịa đặt" has SO value of (-2), then "không bịa đặt" has SO value of (+2).

The emotional value of the sentence depends on the word defect: The defective words include: “nên”, “phải” and “có thể”. Sentences containing the word defect often show a lower degree of emotional distress than similar sentences that do not contain the defective word. Obviously, we can easily see that the sentence: "Bạn có thể làm tốt hơn", the object mentioned here really has not done the best of his ability, and the emotional meaning will be lower than the sentence: "Bạn làm tốt". Therefore, the selection of a level of emotional mitigation in sentences with defective words is a fact of concern, however, how much of that mitigating value is appropriate requires time for further investigation and research. In this paper the mitigating value that we have chosen is 50%. Accordingly, sentences containing defective words, the emotional value of the sentence is reduced by 50% compared to the emotional value of all words with emotional meaning in the sentence.

For example: The sentence “Bạn có thể làm tốt hơn.”. The phrase “tốt hơn” has an SO value of (+2) but in the sentence there is a defective word “có thể” so the SO value of “tốt hơn” is reduced to (+1). Or the saying "Chúng ta phải thật mạnh mẽ.". The phrase “thật mạnh mẽ” has SO value of (+2)*(1 + 0.3) = (+2.6) but in the sentence with the defective word “phải” so the SO value of “thật mạnh mẽ” will remain (+1.3).

Sentiment value of sentences tends to be positive: Classification of emotions based on emotion dictionaries often shows a positive bias (Kennedy and Inkpen, 2006) [10]. In fact, people tend to use more positive words. There are many ways to balance the positive and the negative. In particular, increasing the emotional value of words with negative connotations is said to be the most effective. We have tested many levels of increase in emotional value of negative words, and the results returned when increasing the emotional value of negative words by 50% is the best.

For example: Sentence “Hôm nay giá vàng tăng và giá đô la giảm”. The word “giảm” with an SO value of (-2) will be increased by 50% to (-2) *(1+0.5) = (-3).

From subjective sentences containing emotions, we proceed to label these sentences again to form a new tagger file. Then from the new tagger file above and the Vietnamese SO-CAL dictionary to classify emotions. The emotional classification of a sentence is actually the selection of a good set of features to achieve high accuracy. The feature set we have edited to suit the Vietnamese language features.

Subjective classification algorithm:

Input: tagger file and SO-CAL Vietnamese dictionary.

Output: The file has a vector structure, with each line being a feature vector.

Application methods:

For each sentence in the dataset, extract the values

1) The total emotional value of the adjectives.
2) The total emotional value of the adverbs.
3) The total emotional value of the nouns.
4) The total emotional value of the verbs.
5) Emotional value of the whole sentence: sum of features number 2, 3, 4 and 5.
6) Emotional value depends on reinforcement word.

Emotional value = value of reinforcement * value of emotional word

7) The emotional value depends on the linking word with opposite meanings.

Emotional value = Emotional value − total of emotional value of the words before the contrasting − linked word

8) The emotional value depends on the word defect.

Emotional value = (0.5) * total value of all word defect in a sentence

9) The emotional value of the sentence tends to be positive.

Emotional value = (1 + 0.5) * value of positive word

10) The emotional value depends on the negative word changing.

Emotional value = (−1) * value of negative word

Returns the feature vector

The system based on the extracted feature will use the machine learning method with the above training dataset to classify each sentence: positive and negative class. The end result is that the data is classified into two categories: positive and negative.

For example of the emotional classification process for a comment. To be able to classify emotions, we must first subjectively classify whether sentences contain emotions or not. Therefore, in this example, we present both subjective and emotional classifications to have the best overview of the entire program's execution.

Analyze emotions for comments: “Chúc mừng em một nhân tài trong tương lai. Hãy cố gắng học tốt nhất, để trở thành nhân tài cho đất nước Việt Nam nhé.". After preprocessing and labeling the returned data is as follows:

```
<doc>
        <s>
                <w pos="V">Chúc mừng</w>
                <w pos="N">em</w>
                <w pos="M">một</w>
                <w pos="N">nhân tài</w>
                <w pos="E">trong</w>
                <w pos="N">tương lai</w>
        </s>
        <s>
                <w pos="R">Hãy</w>
                <w pos="V">cố gắng</w>
                <w pos="V">học</w>
                <w pos="A">tốt</w>
                <w pos="R">nhất</w>
                <w pos=",">,</w>
                <w pos="E">để</w>
                <w pos="V">trở thành</w>
                <w pos="N">nhân tài</w>
```

```
                <w pos="E">cho</w>
                <w pos="N">đất nước</w>
                <w pos="Np">Việt Nam</w>
                <w pos="I">nhé</w>
        </s>
</doc>
```

After preprocessing, the system proceeds to withdraw the concentration. The returned data is typical of each sentence in the above comment as follows: For the sentence: " Chúc mừng em một nhân tài trong tương lai.".

Characteristic withdrawal results for the subjective analysis process are: 1:6.0 2:0.0 3:0.0 4:3.0 5:1.0 6:4.0

In which: Feature number 1 (characteristic of the number of words in a sentence) has a value of 6.0 because the sentence has 6 words. Features 2, 3, 4 and 5 are the total emotional value of the words in the sentence, respectively, in the following order:

Feature number 2 and 3 both have the value 0.0 because there are no adjectives (A tag) and adverb (R tag) in the sentence.

Feature number 4 has a value of 3.0. The noun (card N) "nhân tài" in the sentence has an emotional value of 3.0.

Feature number 5 has a value of 1.0. The verb (V tag) "chúc mừng" has an emotional value of 1.0.

Feature number 6 (characteristics of the total emotional value of the whole sentence) has a value of 4.0 (0.0 + 0.0 + 3.0 + 1.0). We see, this is a normal sentence and does not belong to exceptional cases. Therefore, the total emotional value in the sentence is equal to the total emotional value of the words in the sentence. That is, in this case the value of feature 6 is equal to the sum of the values of features 2, 3, 4 and 5 combined. The subjective classification result returns this as a subjective sentence containing emotions.

After classifying the above sentence as subjective wooden sentences containing emotions, the program continues to extract features for the emotional classification process as:

1:0.0 2:0.0 3:3.0 4:1.0 5:4.0 6:4.0 7:4.0 8:4.0 9:4.0 10:4.0

In which: features No. 1, 2, 3, 4 and 5 are inherited from features No. 2, 3, 4, 5 and 6 in the subjective analysis. After analysis, the sentence above does not have special elements such as: reinforcing words, linking words with opposite meanings, defective words, negative words and changing negative words. Therefore, the features number 6, 7, 8, 9 and 10 all have the value of 4.0 and are equal to the feature value of 5. That is, the emotional value of this sentence depends only on the word containing the emotion. rather than dependent on other factors.

The emotional classification result returned for this sentence is a positive sentence because the features are all positive. To evaluate the accuracy of the emotion classification method when using the SVM machine learning algorithm on the collected data set in combination with the Vietnamese emotion dictionary. The results are presented in the following Table 5.

Table 5. Accuracy assessment results using SVM

No.	Subject	Results (Accuracy: %)
1	Education	89.5%

2	Movie	89.3%
3	Sport	88.2%
4	General	89.5%

4 THE EXPERIMENTAL RESULTS

To value our model, we focus to collect three sets of data from 3 topics: education, movies and sports following: Education: includes 405 comments; Movie: includes 379 comments; Sports: 500 comments and combined to build a larger dataset consisting of 885 sentences.

We use 885 sentences to train our model to classify subjective and emotional classification. The results are presented in the Table 6.

To test our model, we collected 443 comments from visitors to aggregate for analysis. The model has classified these 443 comments into 314 objective comments and 129 subjective comments and after that analyzing these subjective comments, 269 positive comments and 45 negative comments.

Table 6. The results of training our model

No.	Subject	Training dataset			
		Subjective Sentence	Objective Sentence	Positive Sentence	Negative Sentence
1	Education	173	99	133	40
2	Movie	194	95	115	79
3	Sport	248	76	201	47
4	General	615	270	449	166

Evaluate the accuracy of the our method are presented in the following Table 4.2.

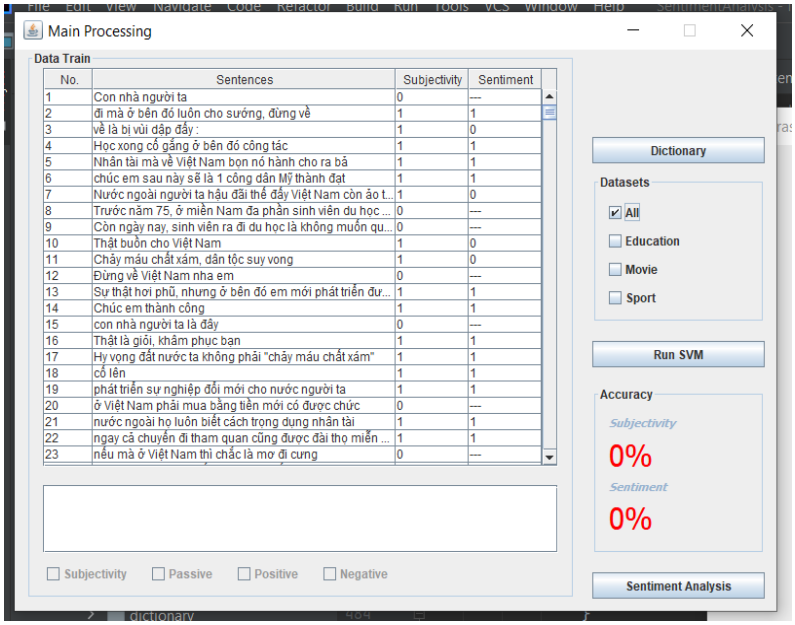
Table 7. The results of training our mode

No.	Classified	Acccuracy (%)
1	Subjective classification	89,8%
2	Emotional classification	89.5%

Moreover, we have built the system to test our method with graphics user interface to help everyone can easy to use our method. Some emotion classification interfaces, rating level, user interest level:

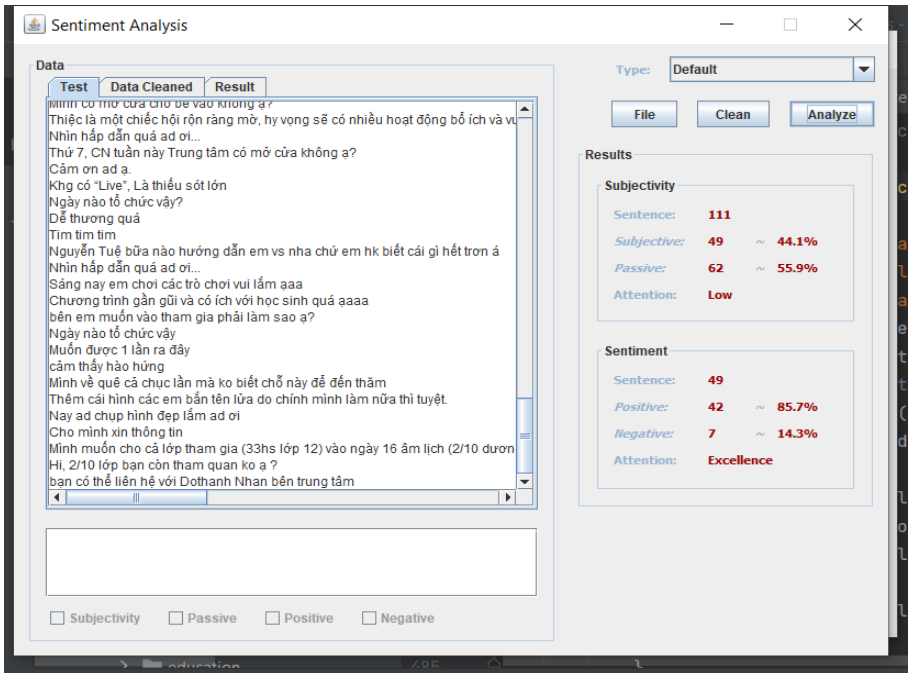
Figure 1. Start screen interface.

In addition to the interfaces for the above main functions, the program also has a number of other information display interfaces in the Figure 2.



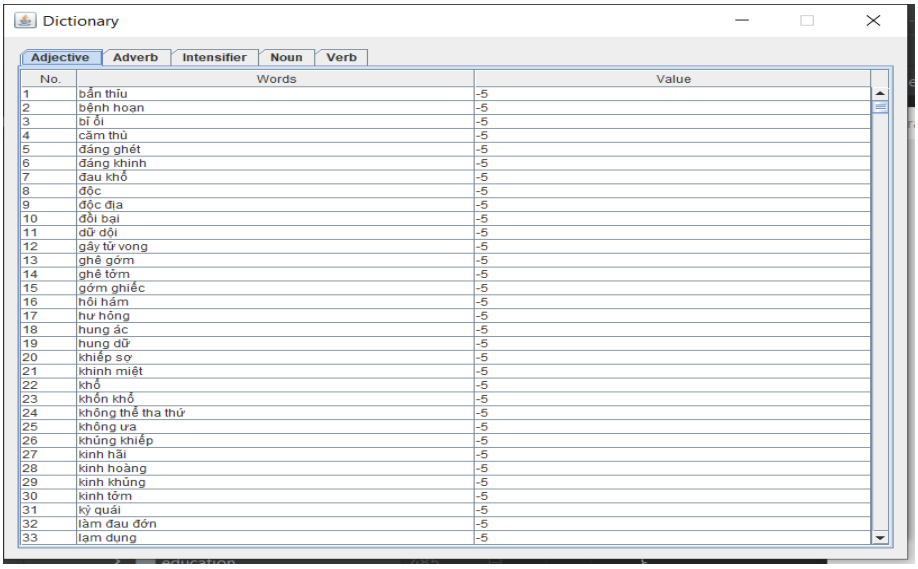
Figue 2. Main function screen interface.

Data list interface after emotional analysis is represented in the Figure 3.



Figue 3. Interface after emotional analysis.

Beside, we have Dictionary display interface is showed in the Figure 4.



Figue 4. Dictionary display interface.

5 CONCLUSIONS

Our method sentiment analysis users evaluate products using lexical analysis combined with machine learning has more accuracy and usefull for Vietnames users. The method has analyzed the emotions of users through analyzing their comments and evaluations for information posted or shared about services and products at the Explorascience QuyNhon. During the process, we were exposed to many related studies, so they helped us to have a broader, deeper and more thorough understanding of the problem. We have developed a feasible method for emotion analysis in Vietnamese language based on the linguistic features of Internet users and at the Explorascience QuyNhon. The results of

the method are built an emotion classification model based on machine learning method combined with word analysis using emotion dictionary.

In additional, the translation of the emotion dictionary from the English dictionary will not only be accurate, but also will be enough to meet the reliability of use. To be able to develop a Vietnamese emotion analysis method, it is necessary to build a large enough and high-accuracy Vietnamese emotional dictionary.

Moreover, we need to train and test our model in big data framework to prove our method also obtain good accuracy for large data volume in the real world.

References

[1] Pang, B., and Lee, L. (2004) A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *arXiv preprint cs/0409058*.

[2] Go, A., Huang, L., and Bhayani, R. (2009) Twitter sentiment analysis, *Entropy 17*, 252.

[3] Liu, B. (2012) Sentiment analysis and opinion mining, *Synthesis lectures on human language technologies 5*, 1-167.

[4] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011) Lexicon-based methods for sentiment analysis, *Computational linguistics 37*, 267-307.

[5] Trinh, S., Nguyen, L., Vo, M., and Do, P. (2016) Lexicon-based sentiment analysis of Facebook comments in Vietnamese language, *Recent developments in intelligent information and database systems*, 263-276.

[6] Mahmood, A., Kamaruddin, S., Naser, R., and Nadzir, M. (2020) A combination of lexicon and machine learning approaches for sentiment analysis on Facebook, *J. Syst. Manag. Sci 10*, 140-150.

[7] Duy, N. (2014) Document summarization based on sentiment classification, Master thesis in computer science (Vietnamese), University of Technology ….

[8] Phu, V. N., and Tuoi, P. T. (2014) Sentiment classification using enhanced contextual valence shifters, In *2014 International Conference on Asian Language Processing (IALP)*, pp 224-229, IEEE.

[9] Kouloumpis, E., Wilson, T., and Moore, J. (2011) Twitter sentiment analysis: The good the bad and the omg!, In *Proceedings of the international AAAI conference on web and social media*, pp 538-541.

[10] Kennedy, A., and Inkpen, D. (2006) Sentiment classification of movie reviews using contextual valence shifters, *Computational intelligence 22*, 110-125.

Notes on contributors

Van Lam Ho received his Ph.D. degree in computer science and engineering at Yuan Ze University, Taiwan in 2016. He is Professor at the Department of Information Technology, Quy Nhon University of Vietnam. His main teaching and research interests include Algorithm, Machine Learning and Data Science. He has published several research articles in international journals of Artificial Intelligence, Data Science, Computer Science .

Minh Vo Le, Master of computer science, Quy Nhon University. His main teaching and research interests include Artificial Intelligence, Internet of Things and STEM at ExploraScience Quy Nhon.

Tran Xuan Viet, Master of Applied Data Science, Quy Nhon University. He is currently an information technology officer at Quy Hoa Central Leprosy - Dermatology Hospital.

Nguyen Ngoc Dung is Lecture at the Department of Information Technology, Quy Nhon University of Vietnam. His main teaching and research interests include Communication Networks, Cybersecurity.

Trang-Thi Ho received her Ph.D. degree in computer science from the National Taiwan University of Science and Technology in 2020. From January 2021 to August 2022, she worked as a Postdoctoral Researcher at the Research Center for Information Technology Innovation, Academia Sinica, Taiwan. She is currently an Assistant Professor at the Department of Computer Science and Information Engineering, TamKang University, Taiwan. Her main teaching and research interests include Machine Learning, Artificial Intelligence, Data Science, and Computer Vision.