Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese language

Luu Nguyen, Minh Vo, Son Trinh, Phuc Do

University of Information Technology, Ho chi minh city, Viet Nam
luut.ng@gmail.com, voleminh10t2@gmail.com, sontq@uit.edu.vn, phucdo@uit.edu.vn

Abtract. Social media websites like Twitter, Facebook etc. are a major hub for users to express their opinions online. Sentiment analysis which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, or reviews. Sentiment analysis can be useful in real life. In this paper, we propose a lexicon based method for sentiment analysis with Facebook data for Vietnamese language by focus on two core component in a sentiment system. That is to build Vietnamese emotional dictionary (VED) including 5 sub-dictionaries: noun, verb, adjective, and adverb and propose features which based-on the English emotional analysis method and adaptive with traditional Vietnamese language and then support vector machine classification method to be use to identify the emotional of the user's message. The experimental show that our system has very good performance.

Keywords: Lexicon-based sentiment analysis, Vietnamese, Text analytics, Vietnamese emotional dictionary, Proposing features, Facebook.

1. Introduction

Social media websites like Twitter, Facebook etc. are a major hub for users to express their opinions online. On these social media sites, users post comments and opinions on various topics. Hence these sites become rich sources of information to mine for opinions and analyze user behavior and provide in- sights for user behavior, product feedback, user intentions, lead generation. Businesses spend an enormous amount of time and money to understand their customer opinions about their products and services.

Sentiment analysis is a type of natural language processing for tracking the mood of the public about a particular product or topic. Sentiment analysis, which is also called opinion mining, involves in building a system to collect and examine opinions about the product made in blog posts, comments, or reviews. Sentiment analysis can be useful in several ways. For example, in marketing it helps injudging the success of an ad campaign or new product launch, determine which versions of a product or service are popular and even identify which demo graphics like or dislike particular features. Thus Sentiment Analysis has become a hot research area since 2002. Sentiment Analysis is used to determine sentiments, emotions and attitudes of the user. The text used for analysis can range from big document (e.g. Product reviews from Amazon, blogs) to small status message (e.g. Tweets, Facebook comments).

Lexicon-based approaches to sentiment analysis differ from the more common machine-learning based approaches in that the former rely solely on previously generated lexical resources that store polarity information for lexical items, which are then identified in the texts, assigned a polarity tag, and finally weighed, to come up with an overall score for the text. Such sentiment analysis systems have been proved to perform on par with supervised, statistical systems, with the added benefit of not requiring a training set. In this paper, we implemented a system with lexicon-based approaches for analysis of Facebook message in Vietnamese language.

The rest of the paper is organized as follows. In section 2, we present related work on sentiment analysis and then present sentiment analysis for Vietnamese language system in details in section 3. In Section 4, we experiment with the training and evaluate the final result obtained from the test data. In section 5, we present our conclusions and outline our future work.

2. Related Work

Sentiment Analysis on raw text is a well known problem.The Liu [3] book covers the entire field of sentiment analysis. sentiment Analysis can be done using machine learning, lexicon-based approach or combined.

The lexicon based approach is based on the assumption that the contextual sentiment orientation is the sum of the sentiment orientation of each word or phrase. Turney [11] identifies sentiments based on the semantic orientation of reviews. Taboada [10], Melville [9], Ding [13] use lexicon based approach to extract sentiments.

In fact, machine learning techniques, in any of their flavors, have proven extremely useful, not only in the field of sentiment analysis, but in most text mining and information retrieval applications, as well as a wide range of dataintensive computational tasks. However, their obvious disadvantage in terms of functionality is their limited applicability to subject domains other than the one they were designed for. Although interesting research has been done aimed at extending domain applicability [2], such efforts have shown limited success. An important variable for these approaches is the amount of labeled text available for training the classifier, although they perform well in terms of recall even with relatively small training sets [1]. On the other hand, a growing number of initiatives in the area have explored the possibilities of employing unsupervised lexicon-based approaches.

These rely on dictionaries where lexical items have been assigned either polarity or a valence, which has been extracted either automatically from other dictionaries, or, more uncommonly, manually. The works by Hatzivassiloglou and McKewon [5] and Turney [11] are perhaps classical examples of such an approach. The most salient work in this category is Taboada [10], whose dictionaries were created manually and use an adaptation of Polanyi and Zaenen's [8] concept of Contextual Valence Shifters to produce a system for measuring the semantic orientation of texts, which they call SOCAL(culator).

Combining both methods (machine learning and lexicon-based techniques) has been explored by Kennedy and Inkpen [6], who also employed contextual valence shifters, although they limited their study to one particular subject domain (the traditional movie reviews), using a "traditional" sentiment lexicon (the General Inquirer), which resulted in the "term-counting" (in their own words) approach. The degree of success of knowledge based approaches varies depending on a number of variables, of which the most relevant is no doubt the quality and coverage of the lexical resources employed, since the actual algorithms employed to weigh positive against negative segments are in fact quite simple.

3. Sentiment Analysis for Vietnamese language

Currently, the research of sentiment problem in Vietnamese language is very little, such as Nguyen Ngoc Duy [7], Vo Ngoc Phu [12] have explored different methods of improving the accuracy of sentiment classification. The sentiment orientation of a document can be positive (+), negative (-), or neutral (0). Dictionary has many verbs, adverbs, phrases and idioms. The author based on the combination of Term-counting method and enhanced contextual valence shifters method has improved the accuracy of sentiment classification. The combined method has accuracy 68.984% on the testing dataset, and 69.224% on the training dataset. All of these methods are implemented to classify the reviews based on our new dictionary and the internet movie data set.

Comparing with previous researches related to our topic, our propose method has different points, that are features which were selected adaptation in Vietnamese language, build a Vietnamese emotional language with more words consistent with the Vietnamese grammar based on spelling that people are using on social network. These point helps improve accuracy in sentiment analysis for Vietnamese comments.

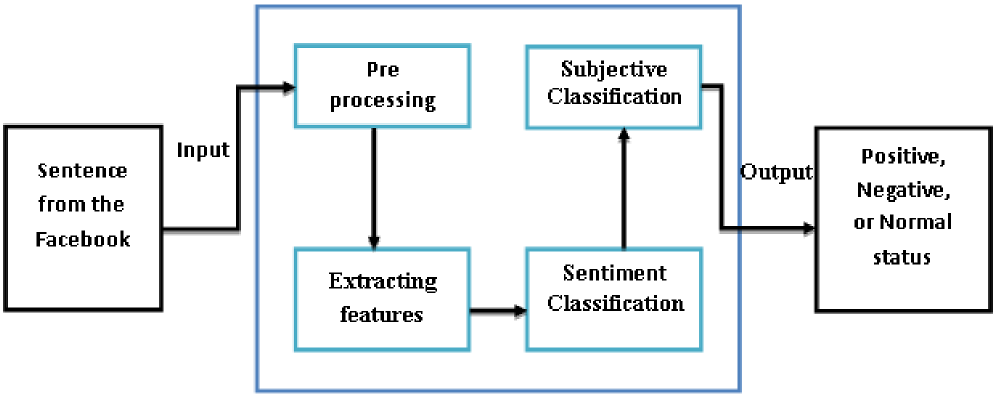

Sentence from the Facebook
Input
Pre processing
Extracting features
Sentiment Classification
Subjective Classification
Output
Positive, Negative, or Normal status

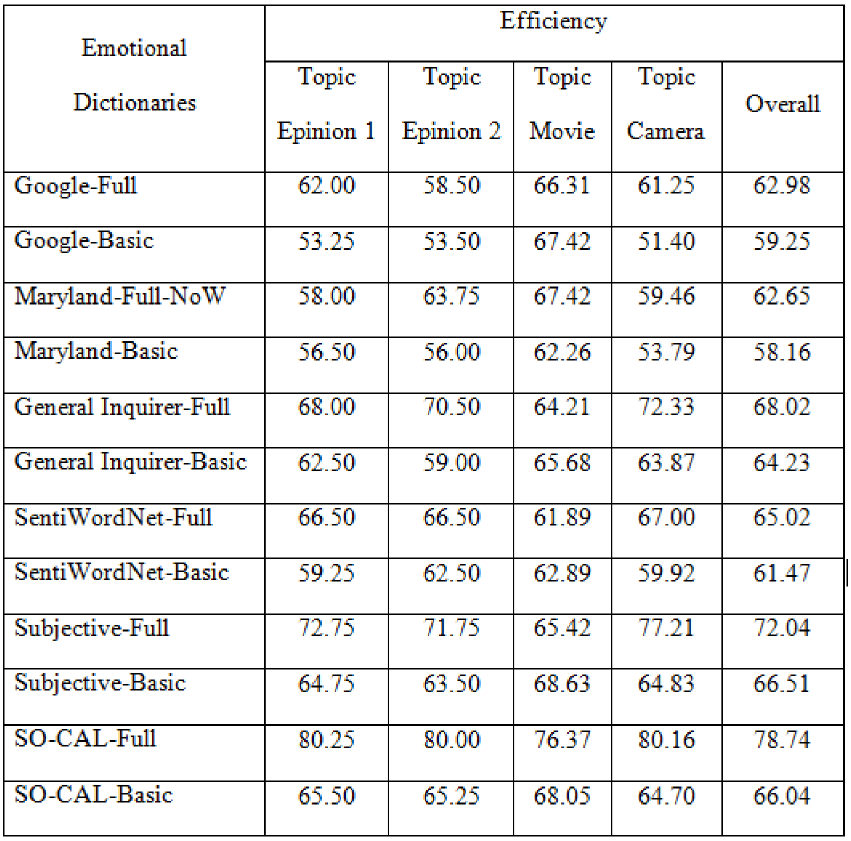
Fig 1. Our proposed system

Our proposed system has 3 components: data collection from the Facebook, preprocessing and extracting features and the third component is analysis emotional for comments from the data. In more details :

- Data collection: Collecting all comments (sentences) from the Facebook.
- Preprocessing and extracting features : Removing stopword, foreign words, icon and pos tagging for all of sentence from the data which has been collected in the previous step and then extract features.
- Analysis emotional: Evaluating the sentence has emotion or non-emotion based on features which were selected adaptation in Vietnamese language and Vietmamese emotional dictionary, and then sentiment classification algorithm has been processed to evaluate the emotion for sentence is in positive, negative or normal status based on support vector machine (SVM) classification method.

3.1 Building emotional dictionary

We created Vietnamese emotional dictionary (VED) which contains 5 sub-dictionaries: noun, verb, adjective, and adverb dictionary. Our dictionary is essentially based on the English SO-CAL (Dictionaries for the Semantic Orientation CALculator) dictionary. We choose SO-CAL, because this dictionary is the best in overall for a lots of topic in experiments. Topic Epinion 1, 2 has 400 documents for each about book, car,computer, cooking, hotel, music and phone. Movie has 1900 documents about films.Camera has 2400 documents about printers and cameras.



Emotional Dictionaries	Efficiency				
	Topic Epinion 1	Topic Epinion 2	Topic Movie	Topic 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

Fig 2. List of emotional dictionaries

In addition, we added some words to our dictionary to make consistent with the Vietnamese grammar and concise spelling that people are using on social network. The number of words in each dictionary of noun, verb, adjective and adverb are 1546 words, 1108 words, 2357 words, 749 words respectively and each word is paired with an integer which describes the corresponding emotional value from the most negative (-5) to the most positive (+5). Notice that no word has SO emotional value at zero value (0).

Table 1. Some words from dictionary of noun

Noun	Emotional value
hoàn hảo (perfection)	5
lộng lẫy (luxury)	4

chiến thắng (victory)	3
phước lành (blessing)	2
độc lập (liberty)	1
tội phạm (crime)	-1
điểm yếu (weakness)	-2

Table 2. Some words from dictionary of verb

Verb	Emotional value
tôn kính (respect)	4
hoan hỉ (delight)	4
thành công (succeed)	3
sáng tạo (create)	2
Tăng (increase)	1
vùi dập (ruin)	-1
xấu hổ (shame)	-2

Table 3. Some words from dictionary of adjective

Adjective	Emotional value
tuyệt vời (perfect)	5
cao cấp (high-grade)	4
bổ ích (helpful)	3
chặt chẽ (close)	2
hợp lý(agreed)	1
cũ (old)	-1
đần độn(silly)	-2

Table 4. Some words from dictionary of adverb

Adverb	Emotional value
thú vị (interestingly)	5
huy hoàng(splendidly)	4
giỏi (well)	3
tươi (freshly)	2
sạch (clean)	1
kỳ quặc(weirdly)	-1
thô (crudely)	-2

Table 5. Some words from intensification dictionary

Intensification	Emotional value
ít (slenderly)	-1.5
chút ít(slightly)	-0.9
hơi (a little)	-0.5
khá (rather)	-0.2
chắc (surely)	0.2
siêu (super)	0.4
hoàn toàn (completely)	0.5

The intensification dictionary has 185 special words in Vietnamese language and each word also has a accompanied decimal to demonstrate the increase or decrease of its emotional value.

Example: If emotional value for word “nhếch nhác” (messy) is (-3) then word “khá nhếch nhác” (rather messy) has emotional value (-3)*(1 – 0,1) = (-2.7). On the same, if emotional value for word “xuất sắc” (excellent) is (5) then word “xuất sắc nhất” (the most excellent) has emotional value 5*(1 + 1) = 10.

3.2 Training

As we know, emotional is extremely complicated. Hence to build a manageable data, we conducted collecting comments and opinions of the social network user and labeled those each sentence in comment to analyze them. Each sentence has subjective nature of every person. The first task is to classify which comment is emotional or non-emotional (also known as subjectivity classification) and the second task is to classify which comment is negative or positive (also known as sentiment classification).

Our data source was chosen from 3 topics: education, movie and sport. Each database contains from 250 to 350 comments of those topics, and then we created a bigger synthetic database from 3 topics which includes 885 comments. In the next step, we divide manually the synthetic database into 2 parts: subjective and objective sentences. After that, the subjective sentences were classified manually into 2 parts: negative and positive sentences.

Table 6. Result of subjective manual classification

Number	Topic	Training data	
		Subjective sentences	Objective sentences
1	Education	173	99
2	Movies	194	95
3	Sports	248	76
4	All	615	270

Table 7. Result of sentiment manual classification

Number	Topic	Training data	
		Positive sentences	Negative sentences
1	Education	133	40
2	Movies	115	79
3	Sports	201	47
4	All	449	166

4. Experimental Model and Result

4.1 Subjective Classification

This method uses 6 features to classify which sentence is emotional or non-emotional:

- Feature 1st: The amount of word in the sentence. It partly displays what the users want to express through the comments. If a lots of number of words are appeared, the user is really interested in this topic.
- Features 2nd, 3rd, 4th and 5th: The total of emotional value of noun, adjective, verb, adverb in the comments. The emotional value of a sentence depends on the type of word which was compared with the VED dictionary.
- Feature 6th: The total of emotional value of a sentence is basically total of 4 attributes that is 2nd, 3rd, 4th, 5th.
- Moreover, the emotional value of a sentence also depend on the type of the sentence. The emotional value of a sentence will be 0 point if this sentence is a condition or a question sentence.

Algorithm

- Input: Sentence has been preprocessing, VED emotional dictionary.
- Output: Feature vectors
- Steps:
  - Count number of words in sentence
  - Find and calculate sum of emotional value of adjective in sentence. (2)
  - Find and calculate sum of emotional value of adverb in sentence. (3)
  - Find and calculate sum of emotional value of noun in sentence. (4)
  - Find and calculate sum of emotional value of verb in sentence. (5)
  - Sum of emotional value of sentence = sum of all values in 2,3,4,5
  - If sentence is question or conditional sentence return 0

    Otherwise, return sum of emotional value of sentence.
  - Return feature vector.

From the feature vector, we use SVM method to classify sentence into subjective (emotional) or objective class (non-emotional).

4.2 Sentiment Classification

After the subjective classification has been processed, we continued to apply the sentiment classification on these sentences. We proposed features which presented in below was based-on the English emotional analysis method and the consistent with traditional Vietnamese language.

- Firstly, emotional value of a sentence depends on the emotional value of each emotional word or phrase. The most basic attributes inherited from subjective analysis. The summary of emotional value of a sentence is total in value of all features above.
- Secondly, emotional value of a sentence which depends on the emotional value of the intensification will be calculated by: Emotional value = value of intensification * value of emotional word
  The total of these values will be the new value of the emotion after review intensification. In the absence of intensification in sentence, this value is the total value of all kinds of emotional words in a sentence.
- Thirdly,emotional value of a sentence also depends on the negative words in the sentence:“không”(no),“không có”(without), … will be calculated by: Emotional value = (-1) * value of emotional word
- Fourthly, emotional value of a sentence which depends on the imperfect words: “nên” (should) , “phải” (must have) , “có thể” (maybe),....will be calculated by: Emotional value = (0.5) * total value of all imperfect words in a sentence
- Fifthly, emotional value of a positive sentence: In fact, traditional vietnamese culture,people avoid using negative words to express their opinions so that the positive words are commonly used. Hence, the emotional value of a positive word will be calculated by: Emotional value = (1 + 0.5) * value of positive word
- Lastly, emotional value of a sentence which has a contrasting-linked word likes: “nhưng”(but),“tuy nhiên” (however),… will be calculted by total of the emotional value of words that subtract the emotional value of the words before the contrasting – linked word by: Emotional value = Emotional value – total of emotional value of the words before the contrasting – linked word

Algorithm

- Input: Sentence has been preprocessing and VED emotional dictionary.
- Output: Feature vectors
- Steps:

For each sentence from the data do :

- Find and calculate sum of emotional value of adjective in sentence. (2)
- Find and calculate sum of emotional value of adverb in sentence. (3)
- Find and calculate sum of emotional value of noun in sentence. (4)
- Find and calculate sum of emotional value of verb in sentence. (5)
- Sum of emotional value of sentence = sum of all values in 2,3,4,5
- Find intensification words in the sentence and update the value of emotional :
  *Emotional value = value of intensification * value of emotional word*
- Find negative words in the sentence and update the value of emotional :
  *Emotional value = (-1) * value of negative word*
- Find imperfect words in the sentence and update the value of emotional:
  *Emotional value = (0.5) * total value of all imperfect words in a sentence*
- Find positive words in the sentence and update the value of emotional:
  *Emotional value = (1 + 0.5) * value of positive word*
- Find linked word in the sentence and update the value of emotional:
  Emotional value = Emotional value – total of emotional value of the words before the contrasting – linked word
- Return feature vector.

4.3 Result

By using our features which had consistent with traditional Vietnamese language and classify based on SVM classification method, we examined the accuracy of subjective classification according to the algorithm in previous step. Results are presented in below:

Table 8. Results of subjective classification

Number	Topic	Result - precision
1	Education	92.6%
2	Movies	89.7%
3	Sports	89.5%
4	ALL	89.8%

We continue to assess the accuracy of the sentiment classification method. Results are presented in below:

Table 9. Results of sentiment classification

Number	Topic	Result - precision
1	Education	90.8%
2	Movies	79.2%
3	Sports	95.0%
4	ALL	89.5%

5. Conclusion

Sentiment detection has a wide variety of applications in information systems, including classifying reviews, summarizing review and other real time applications. In this paper, we proposed a lexicon based method for sentiment analysis with Facebook data for Vietnamese language by focus on two core component in a sentiment system. That is to built Vietnamese emotional dictionary (VED) which contains 5 sub-dictionaries: noun, verb, adjective, and adverb and proposed features which based-on the English emotional analysis method and adaptive with traditional Vietnamese language and then support vector machine classification method has been used to identify the emotional of the user's message. The experimental show that our system has very good performance. In future, we are continuing on improving the performance by building a larger emotional dictionary for Vietnamese language and integrating some technical nature language processing, and solve the big data problem.

References

1. Andreevskaia, A., and Bergler, S. (2007). ClaC and CLaC-NB: knowledge-based and corpus-based approaches to sentiment tagging. Proceedings of the 4th International Workshop on Semantic Evaluations (pp. 117–120). Prague, Czech Republic: Association for Computational Linguistics.
2. Aue, A., and Gamon, M. (2005). Customizing Sentiment Classifiers to New Domains: A Case Study. Presented at the Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria.
3. Bing Liu. 2008 Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 1–167,Morgan & Claypool Publishers.
4. Gamon, M., and Aue, A. (2005). Automatic identification of sentiment vocabulary: exploiting low association with known sentiment terms (pp. 57–64). Ann Arbor, Michigan: Association for Computational Linguistics
5. Hatzivassiloglou, V., and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics (pp. 174–181). Madrid, Spain: Association for Computational Linguistics.
6. Kennedy, A., and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence, 22(2), 110–125.
7. Nguyen Ngoc Duy, Document summarization based on sentiment classification. Master thesis in computer science (Vietnamese), University of Technology Hochiminh city, 2014.

8. Polanyi, L., and Zaenen, A. (2006). Contextual Valence Shifters. Computing Attitude and Affect in Text: Theory and Applications (pp. 1–10). Dordrecht, The Netherlands: Springer.
9. Prem Melville, Wojciech Gryc and Richard D Lawrence. 2011. Sentiment analysis of blogs by combining lex- ical knowledge with text classification. Proceedings
10. Taboada, M., Brooks, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. Computational Linguistics, 37(2), 267–307.
11. Turney, P. D. (2002). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) (pp. 417–424).
12. Vo Ngoc Phu and Phan Thi Tuoi, "Sentiment classification using Enhanced Contextual Valence Shifters", Proceedings of International Conference on Asian Language Processing, Malaysia, 2014.
13. Xiaowen Ding, Bing Liu and Philip S Yu. 2008. A holistic lexicon-based approach to opinion mining. Proceedings of the international conference on Web search and web data mining 231–240, ACM.