Feature-Based Opinion Mining on Smart-Phone Reviews

Mongkol Saensuk, Panida Songram, Phatthanaphong Chomphuwiset

Department of Computer Science, Mahasarakham University,
Kantarawichai District, Mahasarakham, Thailand

*Corresponding Author: Mongkolsaensuk@gmail.com

Abstract

Feature-based opinion mining is a technique that identifies positive and negative polarity based on object features. The mining technique is different from traditional opinion mining as the feature-based technique does not only examine and summarizes the overall opinion of each review. In smart-phone market, the overall review may not be a practical recommendation for customers to choose their phones. It is essential to summarize opinions of users based on each features of smart-phones. This will be advantage for customers who want to buy smart-phones and smart-phone companies that can use this information to improve features of their smart-phone. In this paper, we propose a method for mining opinions on smart-phone reviews written in Thai. The method summarizes positive and negative polarity of each feature of smart-phones. In this paper, first, smart-phone reviews are collected from smart-phone pages on Facebook using Facebook Graph API. Second, the review dataset are clean and then perform word segmentation using a Thai segmentation technique. Then finding similarity word of each feature is performed as the same feature may write the different words. Finally, each feature is decided to be either negative or positive by considering polarity of words which are after the feature. From the experimental result, it was shown than the proposed method gives 70.17% of accuracy.

Keywords: Feature-based opinion mining, Opinion mining, Opinion classification, Smart-phone mining

1. Introduction

Currently, smart-phone is a part of daily life. Thai people spend the most time day with their smart-phone because people can use it for many purposes, such as

photography, communication, travel, information searching etc. The InsightExpress company, the market research in the US, reports that 98% of Thai people habitually use smart-phones. The number of thai people using smart-phones has increasing 29%. Based on increasing demand of smart-phones, many smart-phone companies tried to produce new smart-phones with user demands. If any new smart-phone releases, they will advertise it. The most popular advertisement channel is social networks such as Facebook, Twister, Linkedin etc. The reviewers are able to write their opinion on the social network websites. The most opinion is about the products and services. Usually, the reviews on smart-phones are about features of smart-phones. For examples, “beautiful camera”, “run out of battery so fast”, “hot machine and it will explode” etc. These opinions express the attitudes, sentiments, and emotions which is very useful for smart-phone companies and customers. The companies use the reviews to improve their products and the people who want to buy a smart-phone usually read some existing reviews on phone pages to help them make a decision. Unfortunately, the number of the existing reviews is very large and difficult to summarize each feature of smart-phone by human. Therefore, this paper proposed feature-based opinion mining on smart-phone reviews written in Thai. The reviews are summarized to identify either positive or negative opinion of each feature of smart-phone.

The remainder of the paper is organized as follows. The related works are mentioned in section 2. The proposed methodology is explained in section 3. The experimental results are shown in section 4. Finally, conclusion is given in section 5.

2. Related Works

Opinion mining or sentiment analysis ⁽⁷⁻⁹⁾ was proposed to analysis sentiments, emotions, and attitudes of people to their interesting topic. Opinions are normally mined for text documents which have to pass text pre-processing steps. First, cleaning text is the modification of text from wrong text to correct text or words. English text can use the spelling correction algorithm ⁽¹⁰⁾, but it cannot be used in Thai text. However, Thai text cleaning can be performed by using a dictionary. Word segmentation ⁽⁹⁾ is the process to separate words from a document. Words in English are able to be separated by space, point (.), and comma (,). Writing style of Thai text is the long consecutive sentence without space or point or comma, so space, point, and comma cannot be used to separate Thai sentences. Therefore, a well-known algorithm, called Longest Matching algorithm ⁽⁶⁾, was proposed for word segmentation of Thai text. Next, feature selection is used to select significant feature and reduce a large number of features. Feature selection methods were used in opinion mining i.e. Information Gain, Chi-square, Gain ratio etc.

After pre-processing data, pre-processed data is normally classified to specify positive or negative opinion of topics by using machine learning techniques i.e. Decision Tree (DT), Naïve Bayes (NB), Support vector machine (SVM), Maximum Entropy (ME) etc. Haddi et al., presented the opinion classification for movie review data set using SVM to classify the data into two classes, i.e. positive and negative. Bakliwal et al., (2011)⁽²⁾ presented the sentiment classification with machine learning approach for movie and product review data set with NB, Multi-layer perception (MLP) and SVM. The output of the opinion classification is positive and negative. The experimental result showed that MLP is the best approach and gives 78.32% of accuracy. Diayan et al.,⁽³⁾ presented the machine learning techniques for cameras (comprising with 41,000 reviews) and compared with two algorithms: NB and SVM. The experimental results were shown that SVM gave 83.12% of accuracy and NB gave 81.12% of accuracy. Khainar and Kinikar (2013)⁽⁴⁾ presented the simple data and algorithm of classification using NB, Maximum Entropy (ME) and SVM.

From the previous researches, they applied machine learning approaches to summarize overall opinions which unable to represent the result of consumer needs, because the result of machine learning classification reports only positive and negative opinion without considering opinions of each feature of interesting object. Thus, feature-based opinion mining was proposed to summarize opinion of each

feature. The output of feature-based opinion mining is more general and useful in many real applications. For example, a sentence, “*The actor is not good, but story of this movie is very good*” will be classified by a machine learning technique that the movie is positive or negative, but it is not summarize that which feature is positive or negative. Actually, it should be summarized the actor is negative and the story is positive that will get more information for making decision. Therefore, feature-based opinion mining should be used for summarize the opinion.

Wu et al.,⁽⁵⁾ presented the interaction of customer opinion from dataset in www.tripadvisor.com, and select interesting features from property of hotels such as rooms, cleanness, service, location etc. They cab summarize opinion of each feature. Prombrut.,⁽⁶⁾ presented opinion classification on mobile phones reviews written in Thai. The reviews are collected from websites and blogs that contain comments about a mobile phone. Path-of-speech (POS tags) is used to specify feature of mobile phone (property of mobile phone) and polarity words. If a word is noun, it will be extracted to be a feature. If a word is verb or adjective or adverb, if will be extracted to be polarity word. Lui.,⁽¹¹⁾ proposed feature-based opinion classification of property of a phone. The features of phone in this research consist of voice, screen, battery, size, weight.

3. The Research Methodology

Feature-based opinion mining in this paper consists of the process as shown in fig. 1. The details of the process is explained in the following sections.

Fig. 1. Process of feature-based opinion mining on smart-phone

3.1 Data collection

The 133 reviews of smart-phone are collected as a dataset from 10th of December 2014 to 22nd of February 2015 – obtaining from smart-phone pages on Facebook by using graph API version 2.3. All reviews are written in Thai. In addition, this paper collects words to create three dictionaries as follows:

(a) Dictionary A

Dictionary A collects the most commonly misspelled words written on social networks and their spelled words. This dictionary composes of 69 misspelled words (together with their corresponding spelled words) are collected from a survey conducting with Thai teenagers who commonly use social networks^{(12) (13)}.

(b) Dictionary B

Dictionary B defines polarity words that are collected from “Mining Opinion in Product Reviews: A case Study of Mobile Phone Reviews”⁽⁶⁾ and social networking websites. Polarity words consist of 40 positive and 27 negative words.

(c) Dictionary C

Dictionary C imposes 11 feature words and their similarity feature words, which are collected from the website detailing smart-phone features in the market¹². We collect the similarity word as some of feature words may be written in the different ways, such as screens and monitors etc.

3.2 Data pre-processing

(a) Text cleaning: in this process, misspelled words are corrected (as to obtain consistent data) before processing to the further steps. To achieve this, dictionary A is utilized. The processes of text cleaning are as follows:

Step 1: Read the data text file to find misspelled words.

Step 2: Replace misspelled words with the correct words by using dictionary A.

(b) Word segmentation: in this process, sentences are separated in words using a software called LexTo provided by NECTEC³. LexTo was developed based on the Longest Matching algorithm with the standard dictionary of NECTEC containing 42,222 Thai words. The example output of this process is demonstrated in table 1.

Table 1. Word segmentation using longest matching algorithm.

Opinion Sentence	Word segmentation
ระบบเร็วกว่าเดิมมาก (System is much faster)	ระบบ เร็ว กว่า เดิม มาก
แบตเตอรี่หมดเร็วแต่เครื่องสวย (Battery dies insanely fast but the body is beautiful)	แบต หมด เร็ว แต่ เครื่อง สวย
เครื่องไม่ค้าง (Machine is not stuck)	เครื่อง ไม่ ค้าง

3.3 Feature and polarity identification

Feature identification is performed by using the third dictionary (dictionary C). Segmented words of each sentence are compared to feature dictionary C. This process is to identify the certain features matching with the words in sentences. The matched features are represented by the numbers, which are used to examine the polarity. The features are represented by using number as table 2. For polarity identification, segmented words of each sentence are compared to polarity words in dictionary B to identify positive or negative words. If the word is positive, it will be represented by (+), otherwise, it will be represented by (-). Then the remaining words in the sentence will be removed. The example of feature and polarity identification is shown in table 3.

Table 2. Feature and feature representation.

Features	Feature Representation
Camera	1
Connectivity and port	2
Size and screen	3
Battery	4
OS and application	5
CPU and ram	6
Memory	7
Voice	8
Color	9
Price	10
Machine and overall	11

¹ <http://www.share4yotha.org/>

² <http://lgmobilelove.com/featurephone/>

³ <http://www.thailang.nectec.or.th/best/?q=node/4>

Table 3. Feature and polarity identification.

Word segmentation	Feature and polarity identification
ระบบ เร็ว กว้าง ดีม มาก	5 +
แบตเตอรี่ หกมด เร็ว แต่ เครื่อง สวย	4 - + 11 +
เครื่อง ไม่ ค้าง	11 - -

3.4 Opinion sentence identification

Opinion sentence identification is performed by considering (+) and (-) after feature number as follows. Opinion sentence identification is show in table 4.

(a) If (-) and (-) are found together, it will be changed to (+).

(b) If (+) and (+) are found together, it will be changed to (+).

(c) If (-) and (+) are found together, it will be changed to (-).

Table 4. Opinion sentence identification.

Feature and polarity identification	Opinion sentence identification
5 +	5 +
4 - + 11 +	4 - + 11 +
1 - -	11 +

3.5 Summarize generation.

This method summarizes opinions of each feature on smart-phone by counting positive and negative opinions of each feature. From experimental, the summarization is shown in table 5.

Table 5. Summarization of the smart-phone.

Features	Positive	Negative
Camera	31	3
Connectivity and port	0	1
Size and screen	2	3
Battery	0	0
OS and application	1	0
CPU and ram	0	0
Memory	0	0
Voice	0	1
Color	1	1
Price	0	4
Machine and overall	58	15

4. Experimental Evaluation

The experiments are performed to analyze the accuracy of the proposed method. 133 smart-phone reviews comprising with 171 relating smart-phone features are experimented. The proposed method gives accuracy rate in equation (1) and the summarization is shown in figure. 2.

Fig. 2. Summarization of the smart-phone.

$$\text{Accuracy rate} = \frac{120}{171} \times 100 = 70.17 \quad (1)$$

The accuracy rate of this experiment is 70.17%. The users can use the result from figure 2 for decision. In this experiment found that some features cannot be extracted because they are not in the dictionary and the segmentation algorithm cannot separate correct features.

5. Conclusions

This paper proposes feature-based opinion mining on the reviews of smart-phones, which are written in Thai. The results of this mining are demonstrated as the degree of positive and negative polarity of each feature. This result is able to benefit for both consumers and companies. In this study, the opinion mining process consists of the data pre-processing and the opinion classification with feature-based. The accuracy rate from the experimental result is 70.17%. The Future works will be focusing on the automatic generation of polarity dictionary. In addition, the application of on-line opinion classification is one of the future works.

References

- (1) Emma Haddi, Xiaohui Liu and Yong Shi : “The Role of Text Pre-processing in Sentiment Analysis”, *Information Technology and Quantitative Management (ITQM2013)*, Vol.17, pp.26-32, 2013
- (2) Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, Vasudeva Varma : “Mining Sentiments from Tweets”, *Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pp. 11-18, 2012
- (3) Diayan, Tiwari and Alam : “To Classify Opinion of Different Domain Using Machine Learning Techniques”, *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, No. 5, pp. 184-187, 2013
- (4) Jayashri Khairnar and Mayura Kinikar : “Machine Learning Algorithms for Opinion Mining and Sentiment Classification”, *International Journal of Scientific and Research Publications*, Vol. 3, No. 6, pp. 1-6, 2013
- (5) Yingcia Wu, Furu Wei, Shixia Liu, Norman Au, Weiwei Cui, Hong Zhou, Huamin Qu : “OpinionSeer: Interactive Visualization of Hotel Customer Feedback”, *IEEE Transactions on Visualization and Computer Graphics*, Vol. 16, No. 6, pp. 1109-1118, 2010
- (6) Naris Prombut : “Mining Opinion in Product Reviews : A case Study of Mobile Phone Reviews”, *Thesis of Master degree of Computer Science King Mongkut’s University of Thechnology North Bangkok (Thai text)* ,pp. 1-68, 2007
- (7) Vinodhini, Chandrasekaran : “Sentiment Analysis and Opinion Mining: A Survey”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 2, No. 6, pp. 282-292, 2012
- (8) Isidro Peñalver-Martinez, Francisco Garcia-Sanchez, Rafael Valencia-Garcia, Miguel Angel Rodríguez-García, Valentin Moreno, Anabel Fraga and Joes Luis Sánchez-Cervantes : “Feature-based Opinion Mining through Ontologies”, *Expert Systems with Applications*, Vol. 41, No. 13, pp. 5995-6008, 2014
- (9) Blessy Selvam and Abirami : “A Survey on Opinion Mining Framework”, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 2, No. 9, pp. 3544-3549, 2013
- (10) Poongodi and Radha : “Classification of user Opinions from tweets using Machine Learning Techniques”, *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. 3, No. 9, pp. 1061-1065, 2013
- (11) Bing Lui : “Sentiment Analysis and Opinion Mining”, 2012
- (12) Karnrawee Chomchoei : “”Internet Thai Language” The Specific Language of Thailand People, the new Generation of Communication on the Internet (In Thai text)”, *Research Report*, 2556
- (13) Suwanna Tungteekaruk : “Using the Internet Language as a Teaching Medium (Thai text)”,