Opinion Mining using TRC Techniques

Nirach Romyen, Sureeporn Nualnim, Maleerat Maliyaem, Pudsadee Boonrawd,
Kanchana Viriyapant and Tongpool Heeptaisong
*Information Technology Program, King Mongkut's University of Technology North Bangkok,
1518 Pracharat 1 Road, Bangkok, Thailand*

Keywords: Sentiment Analysis, Opinion Mining, Text Representing Centroids, Co-occurrence Graph.

Abstract: Sentiment analysis is a recent research field in Natural Language Processing (NLP). Text mining and computational techniques determine the sentiment discovered from text. This paper proposes a sentiment analysis using the Text-Representing Centroid (TRC). TRC is a method to determine minimum average distance to all words of the respective document, it also deploys a co-occurrence graph to represent existing relationships among terms in a customer's reviews on particular products and services. A corpus that contains 800 randomly selected hotel reviews from TripAdvisor website is used to evaluate performance by comparison between TRC method and expert's judgment review. The results show 75% accuracy over Thai customer's reviews.

1 INTRODUCTION

For any business, studying customer feedback has become significant to the business because their opinion about the experience they have with a product or service is a resource for improving the customer experience and adjusting business actions to their needs. Moreover, there are also many customer reviews available on the internet, such as websites, blogs, forums, and social media have become a necessary resource for obtaining information related to any topic or domain because of allowing people to provide their opinion about the services and products (Chumwatana, 2015; Juarez, Cervantes Villagómez, & Ayala, 2018). Given the amount of text produced by interactions on websites, blogs, forums, and social media, etc., is important to use advanced techniques to be able to understand and obtain valuable patterns from these large volumes of data (Bouadjenek, Hacid, & Bouzeghoub, 2016).

Sentiment analysis techniques have become an important task in recent years. Sentiment analysis, also called opinion mining is one of the research topic in the area of natural language processing (NLP), text mining and computational techniques. It plays a significant role in our decision-making process to automate the extraction or classification of sentiment from sentiment reviews (Abirami & Gayathri, 2017). The bag-of-words method, which is a traditional method for opinion mining, extracts review into many words, and uses extracted words to identify whether the opinion polarity is positive or negative feedbacks (Chumwatana, 2015). This method is appropriate for capturing word frequency, however structure and information are ignored (Sonawane & Kulkarni, 2014).

In addition, the major problems that should not be ignored in the use of bag-of-words with Thai customer's review are: Thai words are written next to each other, without separate words and special character to identify the end of sentence like in English (Chumwatana, 2015). Graph techniques which mathematically constructs model of relationships and structural information effectively is a growing area of study. This technique is powerful because it can be helpful in the text that yields novel and insightful knowledge from data (Sonawane & Kulkarni, 2014).

Sentiment analysis of Thai reviews is a complex task because of Thai language writing style, a Thai sentence starts from left to right without spaces between words and capital letters, the order of words in a sentence also plays a role in determining the part of speech. In Thai language, some words may have many meanings and derivational suffix used for creating a new word from the base word. Text centroid is an investigated method recently used to categorize and compare documents written in European languages (Mario M Kubek & Herwig Unger, 2016). Moreover, the text-representing centroids method can be applied to Thai documents, too (Nualnim, Romyen, & Sodanil, 2019).

Romyen, N., Nualnim, S., Maliyaem, M., Boonrawd, P., Viriyapant, K. and Heeptaisong, T.
Opinion Mining using TRC Techniques.
DOI: 10.5220/0010315203210326
In *Proceedings of the 10th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2021)*, pages 321-326
ISBN: 978-989-758-486-2

Therefore, this paper proposes a sentiment analysis of Thai customer' reviews using a text-representing centroid method. On the methodology of centroid terms, this method compares the distance between the centroid term of reviews and the reference centroid term in the corpus. This paper applies co-occurrence graphs to determine centroid terms of text documents because it obtains more information about the text documents than word frequency (Nualnim et al., 2019).

The rest of this paper is organized as follows: Section 2, presents the base theories and methods. A methodology is described in section 3. Section 4, covers experiments and results. Finally, section 5, conclusions the paper.

2 RALATED WORK

2.1 Sentiment Analysis

Sentiment analysis (or) opinion mining has gained a lot of interest in text mining and plays a significant role in our daily decision making process (Abirami & Gayathri, 2017; Haruechaiyasak, Kongthon, Palingoon, & Trakultaweekoon, 2018). It is the automated process of understanding an opinion about a given subject from a written language (Krishnakumari & Akshaya, 2019). The task of sentiment analysis consists of determining the polarity associated with a given text document (Castillo, Cervantes, & Vilarino, 2017). Thai sentiment analysis is a complex task because Thai is one of unsegmented languages (Porntrakoon & Moemeng, 2018). For Thai language writing style, a sentence starts from the left to right without capital letters, with no space between words or markers at the end of the sentence. The order of the words in the sentence determines the role and meaning of the words. In Thai language some words may have many meanings and derivational suffix used for creating a new word from the base word. Research on the analysis of sentiment has already been proposed, such as S-Sense (Haruechaiyasak et al., 2018). SenseComp (Porntrakoon & Moemeng, 2018), and SenseTag (Trakultaweekoon & Klaithin, 2016).

The work by Haruechaiyasak et al. (2018) proposed a framework called S-sense for analyzing sentiment on Thai social media. The framework involves two main components, analysis modules, and language resources. The analysis modules, consist of two modules such as intention and sentiment. They are based on a classification algorithm to automatically assign appropriate intention and sentiment class labels for a given text. The language resources are used to train classification models i.e. corpus and lexicon. Porntrakoon and Moemeng (2018) proposed a framework called SenseComp to automatically analyze Thai sentiment of consumer's review in product, price, and shipping dimensions. The methodology of framework use multi-dimensional lexicon and sentiment compensation technique. Sentiment compensation technique is used to automatically compensate the sentiment to a dimension where consumer's review mentioned without a dimension. Additional sentiment analysis for Thai natural language processing is proposed by Chumwatana (2015). This study aims to analyses Thai customers' opinions from the comments on social media. The method combines Thai word extraction technique and sentiment analysis technique to check customers' opinions, in order to classify customers' comments into three categories: neutral, positive and negative. The research process consists of five main steps: (1) collecting Thai customers' comments from social media, (2) Thai word extraction, (3) detecting polarity words, (4) calculating polarity word scores from comments and (5) classifying the comments into groups.

2.2 Text Centroid Representation

Text centroid is a method to identify the representative of the word in the documents that have been inspired by the center of mass in physics. It is the arithmetic mean of all points weighted by the local density as depicted in Figure 1 (Mario M. Kubek & Herwig Unger, 2016). If a physical object has uniform density, its center of mass is the same as the centroid of its shape.

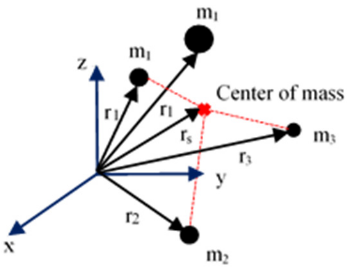

Figure 1: The center of mass in physics.

The distribution of mass is balanced around the center and the average of the weighted coordinates of the distributed mass determines the coordinates and position of the object. To reduce the complexity of calculations, normally this will be replaced with a single mass at the position or the center of mass

(Mario M. Kubek & Herwig Unger, 2016). When compared to text processing, mass point will mean word and distance vector will mean distance in co-occurrence graph.

The centroid of a document is the term with the minimum average distance to all words of the respective document in the co-occurrence graph. This is illustrated in Figure 2.

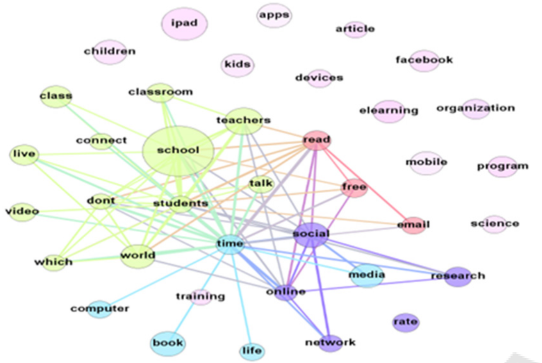

Figure 2: Words of the document in the co-occ. graph.

Consider Figure 2, when you need to find a centroid term of the document consists of the word school, classroom, student, teacher, and computer, the process of finding the centroids is as follows: the word "school" will be the centroid of this document because it is the term with the minimum average distance to another word in the document.

Text-representing centroid is a basic component of the distance-based measure. The co-occurrence graph will be used to determine centroid terms of text documents. The reason for choosing is that the co-occurrence graph obtains more detailed information about the text documents than term frequency vectors. Consequently, it can reflect text structure information and retain important semantic information (Mario M. Kubek & Herwig Unger, 2016).

In co-occurrence graph, terms are represented by vertices and relation between terms is represented by edges. The co-occurrence graphs $G = (W, E)$ could be obtained, if all words of a document or a set of documents W are used to build its set of nodes which are then connected by an edge $(w_a, w_b) \in E$ if $w_a \in W$ and $w_b \in W$ are co-occurrences. Therefore, any two words w_a and w_b are called co-occurrence, which is connected by an edge$(w_a, w_b) \in E$, if they appear together in one sentence. The centroid term can be computed by the following steps:

Step 1: Calculate weight function $g((w_a, w_b))$.

Weight function is the number of occurrences of the two words together in a sentence. If $g((w_a, w_b))$ is high, that means two words are semantically close (Mario M. Kubek & Herwig Unger, 2016). Thus, the following equation is obtained:

$$g((w_a, w_b)) = \frac{freq(w_a, w_b)}{freq(w_a) + freq(w_b) - freq(w_a, w_b)} \quad (1)$$

Where $g((w_a, w_b))$ denotes the weight between node w_a and w_b, $freq(w_a, w_b)$ shows the number of times w_a and w_b occur together in the unit. $freq(w_a)$ and $freq(w_b)$ denotes the frequency of w_a and w_b appearing in the documents respectively.

Step 2: Calculate distance $d(w_a, w_b)$ between two words from graph G which can be defined by:

$$d(w_a, w_b) = \frac{1}{g(w_a, w_b)} \quad (2)$$

In all other cases, there is a shortest path $p = (w_1, w_2), (w_3, w_4), .., (w_k, w_{k+1})$ with $w_1 = w_a, w_{k+1} = w_b$ and $w_i, w_{i+1} \in E$ for all $i = 1(1)k$ such that

$$d(w_a, w_b) = \sum_{i=1}^{k} d((w_i, w_{i+1})) = MIN \quad (3)$$

Whereby in case of a partially connected co-occurrence graph $d(w_a, w_b) = \infty$ must be set.

Step 3: Calculate centroid-term of documents D

This can be calculated by the distance between a given term t and documents D containing N words $Nw_1, Nw_2, \ldots, Nw_n \in D$ that are reachable from t in G which can be defined by

$$d(D, t) = \frac{\sum_{i=1}^{N} d(Nw_{i,} t)}{N} \quad (4)$$

3 RESEARCH METHODOLOGY

In this paper, the proposed method aims to check Thai customer's opinions from the comments using the methodology of text-representing centroids to classify the polarity of text as positive or negative. The proposed techniques are based on a co-occurrence graph that represents existing relationships among terms in a customer's reviews or opinions about the products and services. Its detailed procedure is presented in Figure 3. The overview of the research process of the proposed methodology, which consists of many steps, is described below.

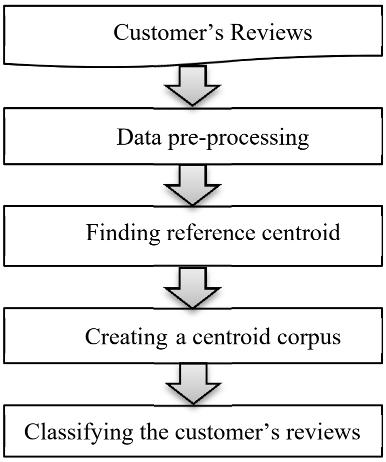
Customer's Reviews
Data pre-processing
Finding reference centroid
Creating a centroid corpus
Classifying the customer's reviews

Figure 3: Process of research methodology.

3.1 Collection of Thai Customer's Reviews

The collection of Thai customer's review is considered the first process for the proposed technique. To perform comment analysis, the collected customers' reviews were classified as positive or negative by Thai language experts to determine the polarity of the review. The Thai customers' reviews collected from Tripadvisor website are then used as an input to extract Thai words in the next step.

3.2 Data Pre-processing

Thai customers' review collected from the first step will be an input for the next process. The pre-processing is a process to ensure the quality of the data before finding the centroid terms, so the pre-processing process consists of several steps. First step is extracting words according to the grammars; second step is considering the meaning of words according to the position of the words in the sentence; the final step is eliminating unwanted words such as conjunctions, and verbs. Therefore, the PyThaiNLP library for python version 3.7, which uses tokenize engine algorithm called newmm (New Maximum Matching algorithm), is applied to divide customer's review into words. After these Thai words are extracted from reviews, the next step is negation checking. This technique uses Parts of speech (POS) to identified negation word, if the word "ดี (good)" follows negation word "ไม่ (not)", then combined negation word with the following word to generate a new word "ไม่ดี (not good)". Subsequently, remove the stop words and choose only adjectives. Therefore, the TLTK library is applied to identify POS tagging.

3.3 Finding Reference Centroid Terms

The reference centroid terms are created to be used as a reference point for the positive sentence and the negative sentence. In order to find the reference centroid terms of a Thai document, centroid terms must be determined with the help of the undirected co-occurrence graph G. Creating reference centroid terms as reference point for positive and negative sentence uses the following steps:

Step1: Create a corpus. This step creates the corpus that contains 400 positive reviews and 400 negative reviews.

Step2: Create a co-occurrence graph. A co-occurrence graph is created by calculating weight of co-occurrence frequency using Formula 1 in this paper.

Step3: Calculate distance between words. The distance of two words in the graph is found using Formula 2 or Formula 3.

Step4: Calculate centroids terms using Formula 4.

3.4 Creating a Centroid Corpus

The corpus used to create the reference co-occurrence graph includes 800 customer reviews selected randomly from Tripadvisor website. The data is divided into 400 positive customer reviews and 400 negative customer reviews. The reviews have been classified as positive or negative by experts.

3.5 Customer Reviews Classification

This process aims to classify the polarity of the text to be positive or negative. The process consists of finding centroid terms of review using Formula 1-4 in this paper. After that, compares the distance between the centroid term of review and the reference centroid term in the corpus. Furthermore, the concept of distance measure is a way of describing what it means for elements of some space to be "close to" or "far away from" each other. If the distance is minimum, it means the meaning is close to each other.

4 EXPERIMENTAL AND RESULTS

This section describes the experiment on analyzation of customers' reviews based on the proposed method.

The research hypothesis is that the methodology of text-representing centroids is a good alternative to classify the polarity of text as positive or negative. This section presents two subsections, classifying the polarity of text and comparing performance.

4.1 Classifying the Polarity of Text

To apply the text-representing centroid to classify the polarity of the text as positive or negative. The centroid terms have been determined with the help of the co-occurrence graph in the corpus. The corpus used to create the reference co-occurrence graph includes 800 customer reviews selected randomly from the TripAdvisor website. In the experiment, 200 customer reviews are used. To be able to evaluate the effectiveness of the text-representing centroids method, this data is chosen equally, divided into 100 positive comments and 100 negative comments from the same data set used to create a co-occurrence graph. For each customer reviews, find the centroids of the customer reviews using Formula 4. Afterward, calculate the distance between the new centroids of the customer reviews, compare it to the reference centroid terms in the corpus. In this experiment, the reference point of the positive comment is a “positive term”, and the reference point of the negative comment is a “negative term”. The example results of finding the polarity of comment are shown in Table 1.

Table 1: Distance from centroid term to positive and negative term.

No	Data	Centroid term	Distance	
			Pos.	Neg.
1	เก่า (old) ไม่ค่อยอร่อย (terrible)	เก่า (old)	0.48	0.35
2	อร่อย(delicious) วุ่นวาย (chaotic) เสียงดัง(loudness)	เสียงดัง (loudness)	0.49	0.43
3	ชอบ(like) เลิศ (excellent) เก่า (old) ดีมาก (good)	เก่า (old)	0.48	0.35
4	ดี (good) อับ (smelly)	ดี (good)	0.28	0.32
5	สะดวก convenient สวย (beautiful) พลุกพล่าน(busily) เก่า (old) โทรม (shabby)	สะดวก convenient	0.35	0.40

Table 1 presents information about centroid terms of customer reviews and the distance between the new centroid terms compared to a positive reference point and a negative reference point in the corpus. In the table, if the distance value is minimum, it means the new centroid term is close to reference point in the corpus. For example, according to Table 1, for the first sentence, when calculating the text-representing centroid, the new centroid term is "old". When calculating the distance between the new centroid term with the positive reference point and the negative reference point in the corpus, it is found that the new centroid term close to a negative reference point. That means the first sentence is a negative opinion.

4.2 Comparing Performance

To evaluate the effectiveness of the text-representing centroids method. The distance measure technique is used to evaluate the accuracy of the sentiment analysis. This technique is a comparison of accuracy between analysis using text-representing centroids methods and customer’s review already categorized by experts. The results are shown in Table 2.

Table 2: The results of comparing performance.

Category	Number of sentence	Result	Accuracy percentage
Positive	100	Pos: 96 sentence Neg: 4 sentence	96
Negative	100	Pos: 46 sentence Neg: 54 sentence	54

The table shows information about the accuracy of the polarity of comments, consisting of 200 comments divided into 100 positive comments and 100 negative comments. In Table 2, it is clear that positive comments have a higher percentage of analytical accuracy than negative comments. Positive comments have an accuracy of 96 percent, while negative comments have an accuracy of 54 percent. The reason that positive comments has higher accuracy than negative counterparts is because the authors give opinions clearly and there are no negative opinions in positive sentences.

5 CONCLUSIONS

This paper presents an approach that uses and applies centroid based techniques on Thai reviews from

Tripadvisor website to analyze their sentiments. The proposed method consists of five main processes: (1) collection of Thai customers' review (2) data preprocessing (3) finding reference centroid terms (4) creating a corpus and (5) classifying the customer reviews. The experiment was created to compare the accuracy between the analysis using the text-representing centroids method and the customer's review already categorized by experts. The experiment results show that the proposed methods correctly analyzes the positive comments better than negative comments. The positive comment classification has an accuracy of 96 percent, while the negative counterpart has an accuracy of 54 percent. Considering the results for comments that are negative, in Thai culture, reviews would begin with positive opinions first and then express negative opinions later. From the above reasons, the use of the methodology of text-centroid to analyze sentiment results yield more errors with negative sentences than positive ones. Future work will involve improving the accuracy of sentiment analysis by considering the importance of previous sentences and focus on the words found in both the positive cluster and the negative cluster.

REFERENCES

Abirami, A. M., & Gayathri, V. (2017, 19-21 Jan. 2017). *A survey on sentiment analysis methods and approach.* Paper presented at the 2016 Eighth International Conference on Advanced Computing (ICoAC).

Bouadjenek, M. R., Hacid, H., & Bouzeghoub, M. (2016). Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms. *Information Systems, 56*, 1-18. doi:https://doi.org/10.1016/j.is. 2015 .07 .008

Castillo, E., Cervantes, O., & Vilarino, D. (2017). Text Analysis Using Different Graph-Based Representations. *Computación y Sistemas, 21*(4), 581-599. doi:10.13053/cys-21-4-2551

Chumwatana, T. (2015). *Using sentiment analysis technique for analyzing Thai customer satisfaction from social media.* Paper presented at the 5th International Conference on Computing and Informatics (ICOCI) 2015, Istanbul, Turkey.

Haruechaiyasak, C., Kongthon, A., Palingoon, P., & Trakultaweekoon, K. (2018). S-Sense: A Sentiment Analysis Framework for Social Media Monitoring Applications. *Information Technology Journal, 14*(1), 11-22.

Juarez, E., Cervantes Villagómez, O., & Ayala, D. (2018). Text Analysis Using Different Graph-Based Representations. *Computación y Sistemas, 21*(4), 581-599. doi:10.13053/cys-21-4-2551

Krishnakumari, K., & Akshaya, P. (2019). A survey on graph based approaches in sentiment analysis. *International Research Journal of Engineering and Technology (IRJET), 6*(7), 1322-1330.

Kubek, M. M., & Unger, H. (2016). *Centroid Terms as Text Representatives*. Paper presented at the Proceedings of the 2016 ACM Symposium on Document Engineering, Vienna, Austria. https://doi.org/10.1145/2960811. 2967150

Nualnim, S., Romyen, N., & Sodanil, M. (2019). *Applicability of Text-representing Centroids for Thai Language Documents.* Paper presented at the Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval.

Porntrakoon, P., & Moemeng, C. (2018). *Thai Sentiment Analysis for Consumer's Review in Multiple Dimensions Using Sentiment Compensation Technique (SenseComp).* Paper presented at the 2018 15th International Conference on Electrical Engineering /Electronics, Computer, Telecommunications and Information Technology (ECTI-CON).

Sonawane, S., & Kulkarni, P. (2014). Graph based Representation and Analysis of Text Document: A Survey of Techniques. *International Journal of Computer Applications, 96*, 1-8. doi:10.5120/16899-6972

Trakultaweekoon, K., & Klaithin, S. (2016, 13-15 July 2016). *SenseTag: A tagging tool for constructing Thai sentiment lexicon.* Paper presented at the 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE).