Journal of Engineering and Applied Sciences 12 (2): 397-403, 2017
ISSN: 1816-949X

Comparing Techniques for Sentiment Analysis in Cosmetic Industry from Thai Reviews Videos

[1]Preedawon Kadmateekarun, [1]Phayung Meesad and [2]Sumitra Nuanmeesri
[1]Faculty of Information Technology, King Mongkut's
University of Technology North Bangkok, Bangkok, Thailand
[2]Faculty of Science and Technology, Suan Sunandha Rajabhat University, Bangkok, Thailand

Abstract: The cosmetic industry has been in a very high marketing competition by advertising through various media to promote sales and build up images of the products. In addition, consumers can access information through a search engine finding that there are up to 45.3 billion video clips and movies of beauty online in social network such as YouTube. Consumers are able to share messages, voices, pictures and video clips and movies through these media swiftly. There are both content and opinions indicating "like" (Positive) or "dislike" (Negative) on the products. These opinions can be brought to conduct the sentiment analysis on the products. This research focuses on automatic sentiment analysis. Therefore, this study aims at the automatic sentiment analysis which is a part of natural language processing of the cosmetic product, lipstick. The research methodology consists of the following steps. Firstly on the collection of data in Thai language such as criticisms on lipstick from YouTube to separate audio signals; secondly, the audio tracks conversion into texts to cut up into words in transcription. Next, the machine learning technique consisting of Naive Bayes (NB) and Support Vector Machine (SVM) to be used for the analysis of the consumer's sentiment towards the lipstick. Finally, The measurement for the efficiency and the comparative study to find result from those different techniques. As a result, the Support Vector Machine (SVM) Technique is found to offer the best result with the accuracy value at 85.17%.

Key words: Sentiment analysis, Thais reviews videos, naive bayes, support vector machine, natural language

INTRODUCTION

At present the consumers of digital age have adapted their communication behavior towards the beauty industry which becomes a business of highly competitive market. They have changed to use online social networks such as YouTube, Facebook and twitter, etc., in expressing their opinions either positively or negatively by sending messages, emoticons, voices or pictures from online video clips or movies about the products or the services through the stated channels with swiftness that result in abundance of increasing information. This causes more time consuming, greater expertise and increasing expense if the right selection and the relevant data classification to one's interest is needed. Therefore, the sentiment analysis which is a branch of studies on the Natural Language Processing (NLP) is then applied to reduce the already mentioned problems, for it is a process that focuses on the sentiment analysis and the examination of the opinions automatically. In the studies in foreign languages such as English and Spanish, the emotions on messages of comments, emoticons, voices and online video clips or movies are analyzed (Neethu, and Rajasree, 2013; Meesad and Li, 2014; Lima and Castro, 2012; Rosas *et al*., 2013; Kaushik *et al*., 2013). For the sentiment analysis in the research studies in Thai language, it has been found that only texts which are posted are mostly analyzed on the emotion of the consumers on interesting subjects such as the products, the services and politics (Phawattanakul and Luenam, 2013; Sukhum *et al*., 2011; Chumwatana, 2015)

From the stated problems and limitations including the existence of video clips or movies about beauty of about 45.3 billion in number that can be brought to analyze their emotion. Therefore, this study intends to present the comparative study on techniques for the automatic sentiment analysis towards the cosmetic product, lipstick as a subject. Comments or suggestions in Thai language selected from video clips or movies on YouTube through the machine learning technique such as naive bayes NB and Support Vector Machine: SVM, etc. are collected for both positive and negative opinions

Corresponding Author: Preedawon Kadmateekarun, Faculty of Information Technology,
King Mongkut's University of Technology, North Bangkok, Thailand

towards lipstick. Subsequently, the testing is done by the method of data classification of sentiment analysis by the presented technique. A comparative analysis of the result is carried out. Then, the assessment is undertaken by using the prepared video clips and movies for the test and a comparative study is conducted to find the accuracy value of the techniques being presented.

Literature review: The sentiment analysis is applied to the data concerning opinions of the consumers towards the products and services (Product review) whether in the form of messages, voice or video clips. The reason sometimes is that the sentiment analysis using only data rating cannot indicate the real problems or feeling of the consumers. For example, a consumer buys a lipstick and gives it the average rating score of 4 from the total score of 5 but the items in the questionnaires might not cover all cases of the consumer's needs and expectations. Therefore, the consumer might note down or make a video clips or movies recording their opinions or feeling towards that lipstick in order to express it out in some places like his own twitter or YouTube. Accordingly, those points of view might be overlooked or not anticipated by the product's owner. If those information is brought into consideration, it might be able to help the improvement of the products and services. At present the technology of sentiment analysis takes a major role in many organizations all in business concerning products, services, education and medical service provision. By including the technology of sentiment analysis into the system of commercial website or Customer Relationship Management (CRM) of each company or organization in order to facilitate the sentiment analysis of the consumers or customers to reach a fast problem solving.

There are many research studies on Sentiment Analysis reported in foreign languages such as English, Spanish and Chinese, etc. They are categorized by the text messages such as opinions, emoticons, voices and gesture images from online video clips or movies. Lima and Castro has been developing the system of sentiment analysis from messages collected from twitter on a television program of Brazilian Channel by using the training data set consisting of emoticons, words and the combination type. Then, use the technique of naive bayes: NB which offers the most accurate value when using the combination type between emoticons and texts. (Lima and Castro, 2012). Later Kaushik *et al.* have developed the sentiment test system by tracking voice from video clips or movies from YouTube and a model of sentiment and automatic voice perception system is constructed by applying data from comments in many aspects such as product criticisms, movies and social issues, etc. from both male and female reviewers. By using the maximum entropy technique, a model on sentiment is made and the system is tested by the technique of Naive Bayes: NB, Support Vector Machine: SVM and Maximum Entropy. Subsequently, a comparative study is conducted to find accuracy value between techniques used for the sentiment model construction and the standard technique. It is found that the technique used for the model construction provides high value on accuracy (Kaushik *et al.*, 2013). Besides, Rosas, Mihalcea and morency have developed various analysis systems of continuing sentiment in Spanish version by using video clips or movies that make general comments from YouTube. The reviewers are both male and female who use messages, voice and gesture images with one another. It is tested by Support Vector Machine (SVM). The test result shows that the accuracy value is satisfying (Rosas *et al.*, 2013). In the research studies in Thai language it is found that most of the works concern the sentiment analysis of messages. Adithep and rattasit have developed the process of emotion evaluation from 6,000 comments in thai which are collected from web sites on news, variety shows and product criticisms. There are experts who classify opinions into 6 categories which are as follows. Love, joy, surprise, sadness, fear and anger, the number of 1,000 comments for each category, an all are presented. The comparative study on accuracy value are assessed with 3 techniques consisting of Naive Bayes: NB, Support Vector Machine: SVM and decision tree. It is found that the technique of Support Vector Machine: SVM gives the most accuracy value. In addition, Patcharanigarn presents a model on the satisfaction analysis of customers from comments on suggestions by constructing a form of opinion mining. When a comparative study on the result of the techniques between "Decision Tree" and "Naive Bayes: NB" is conducted, it is found that the satisfaction analysis of the customers by using the technique of "decision tree" shows more accurate result. Besides, preedawon and sumitra presents the sentiment analysis process with the technique of naive bayes: NB by using the voice data of video clips or movies from YouTube. The test result shows good level of accuracy value.

Audio extraction and Automatic Speech Recognition (ASR): The audio extraction is a process to extract or rip the audio tracks from video clips or movies. In general, this process separates the video clips into two parts: audio track and movie track or visualized information. We can find a lot of software to extract the audio track from the video clip or YouTube such as YouTube-dl, where the user can extract the audio with some command line.

Automatic speech recognition is a process for deriving the audio tracks into sentences or messages as texted.

Word segmentation: Subsequently, these texts are cut into words with word segmentation steps. In the other words, it can be called "transcription". The dictionary-based approach is used for word segmentation by the technique called the longest matching to find the longest matching of words for the comparison between words input with the words in the dictionary. If such words cannot be found to match the words in the dictionary, the system will cut down the alphabets one by one according to the orthography until the word can be compared word for-word with those in the dictionary. It is found that this technique of word segmentation has high level of correctness but there is a weak point when the compared words are too long at first which might cause errors because some words in Thai language are combined words derived from the combination of words. The transcriber software is used at this point to perform only the audio track.

Feature extraction: The extraction of feature expressing sentiment is a construction of substitute word for features in the study which might be a single word, phrase or clause. The extracted features are categorized into types in vector data which means the vector elements can be substituted by the characteristics of the true value or the word frequency value. For this study, the single word from the word segmentation is the feature and subsequently, the technique of "Bag of Word" is used for the feature extraction.

In feature reduction of the study, the researcher refrains from prepositions and conjunctions. After being cut, they do not cause any changes to the sentiment context.

The construction of the substitute word of the study: For the machine learning technique it is favorable to use the word substitution for the context in place of paying attention to the word meaning. The substitution words are usually represented by the form of vector data of word weight value. Frequently, the word weight value can be in binary or non-binary form depending on the calculation method of weight value. In this research study, the calculation of weight value use the binary type.

Machine learning technique: The machine learning technique with supervisors (Supervise Learning) which is used in this research study are the technique of Naive Bayes (NB) and the technique of Support Vector Machine (SVM).

Naive bayes: NB is a data classification method which performs efficiently on the case with many sample sets and the properties of each sample set are independent. We determine that the probability of data which belongs to v_j group is the data which have the properties of n $X = \{a_1, a_2, \ldots, a_n\}$ or $P(a_1, a_2, \ldots, a_n \mid v_j)$ for its symbol which equals:

$$P(a_1, a_2, \ldots, a_n | v_j) = \prod_{i=1}^{n} P(a_i | v_j) \quad (1)$$

From Eq. 1 Π means multiplied result of $P(a_i \mid v_j)$ total $i = 1, 2, 3, \ldots, n$ and $j = 1, 2, 3, \ldots, n$ the probability value of the word found in each group calculated by multiplying $P(a_1, a_2, \ldots, a_n \mid v_j)$ from Eq. 1 by the probability value of that group. Therefore, $P(v_j)$ equals v_{NB}. Then, compare the two received values. The group that has the highest Probability value is the solution. Therefore, the classification method of Naive Bayes: NB is shown as in Eq. 2:

$$v_{NB} = \arg\arg\max P(v_j) \times \prod_{i=1}^{n} P(a_i | v_j) \quad (2)$$

Support vector machine: SVM the concept of Support Vector Machine: SVM is brought to find the level of decision making. In dividing the data into two parts, we will choose the part with high dimension of data.

We determine that $(x_i, y_i),\ldots,(x_n, y_n)$ is the example used in the instruction. n is the number of sample data. m is the number of the entering data dimension and y is the result that has the value of +1 or-1 as the Equation $(x_i, y_i),\ldots,(x_n, y_n)$; $x \in R^m$, $y \in \{+1,-1\}$. For the linear problem, it is divided into 2 groups by the level of decision making which can be calculated as in Eq. 3.

$$(w \bullet x) + b = 0 \quad (3)$$

When w is the weight value and b is the bias value as in the Eq. 4 for the data classification.

$$(w \bullet x) + b > 0 \text{if} y_i = +1$$

$$(w \bullet x) + b < 0 \text{if} y_i = -1 \quad (4)$$

However, Support Vector Machine: SVM has the kernel function which users can apply in many problem solving methods. In this study the kernel function of polynomial kernel type is being used.

MATERIALS AND METHODS

The technique comparison for the sentiment analysis towards the cosmetic products by using the

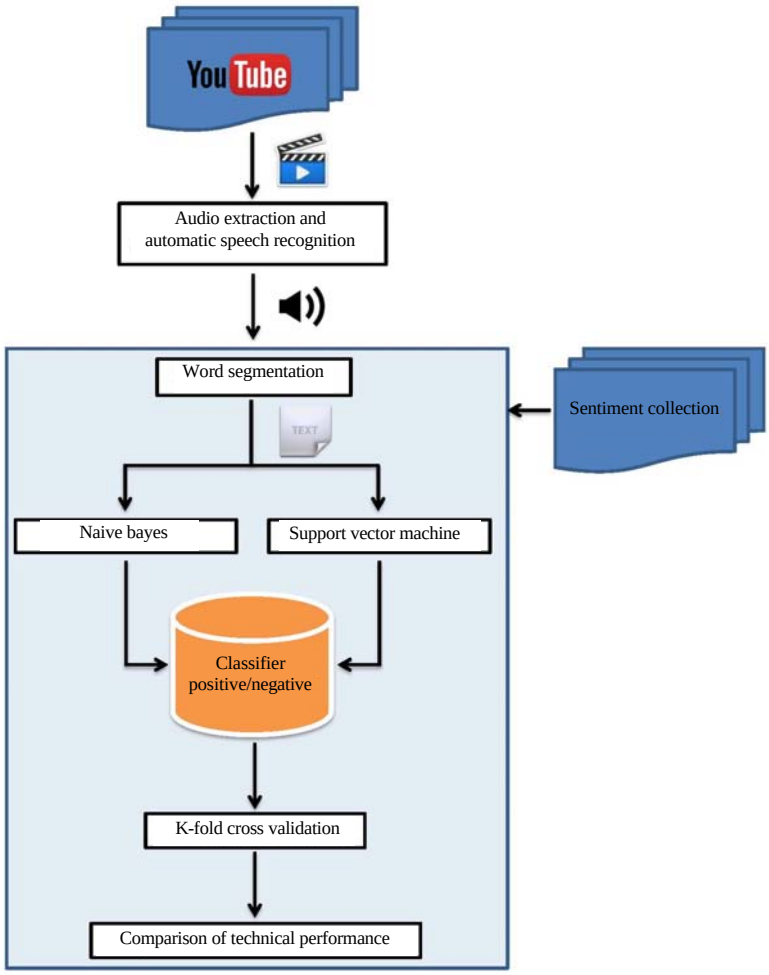

Fig. 1: The comparison process of the sentiment analysis techniques towards the cosmetic products by using comments in Thai language extracted

comments in Thai language extracted from criticism video clips from YouTube is divided into 5 steps as shown in Fig. 1 as follows Sentiment Collection, the step in collecting comments and categorizing them into groups. The step in conducting audio extraction and brought to Automatic Speech Recognition (ASR). The step of preprocessing which is about Word Segmentation and determine the word function. Subsequently, the feature extraction is conducted to construct the vector data with the features of weight, step of Modeling by introducing the data for machine learning with 2 techniques which are: the technique of Naive Bayes: NB and the technique of Support Vector Machine: SVM. The testing method of k-fold cross validation is chosen to be used by determining the value k = 10 and comparison of technical performance, the step in comparing the classification efficiency measurement according to the information retrieving concept.

Sentiment collection: In this study there is a collection of comments in Thai language towards the cosmetic products, lipstick, extracted from the criticism video clips or movies from YouTube. The video clips are between 2-10 min in length of time or the average length of 5 min. From 200 video clips, three listeners conduct the video clip classification into positive clips (100 clips) and negative clips (100 clips).

Audio extraction and Automatic Speech Recognition (ASR): The audio extraction is the process of extracting the sound from 200 video clips or movies from the data collection stage. In this research, Gom player is being used. Subsequently, the audio tracks are transformed into texts by ASR tool which result in texted sentences.

Word segmentation: The form of ARFF file is a text file used for defining the data of the same attributions. They

consist of @HEADER which indicates the relationship between the attributes and the data types and followed by the @DATA the last part of which is the class label that is already been grouped.

Due to the fact that the data-base document that is kept cannot be filed in the data table form; moreover, the comments have been through many processing steps such as word segmentation, feature extraction and substitution paper. Therefore, in this study, the database of text mining in ARFF file using the technique of natural language processing are to be applied with the following steps.

Word segmentation and word function determination: the researcher uses the program of word segmentation called Lexto from the National Electronics and Computer Technology Center (NECTEC) as a tool in performing the word segmentation by which the segmentation is done according to the dictionary by selecting the longest matching and then determining the related vocabulary with those sentiment word groups which cannot be found in the dictionary (Lexitron) but can be added into the vocabulary bank (Corpus) to facilitate the word segmentation to be most accurate as desired. The morphemes from the segmentation are determined for their functions and then entered into the feature extraction process.

The feature extraction is the selection of morphemes for the construction of attributes. This step is to examine the determined functions given in the previous step by selecting all the morphemes except for prepositions and conjunctions which perform as linkers and are categorized in stop word group that are not related to the sentiment features. As they are in the part of feature reduction. On the other hand, the selected morphemes will be kept in "Bag of Word" to construct altogether 7,000 attributes.

The construction of the substitution for the study is the building of metric n×m with attributes from "Bag of Word" of 7,000 attributes which are suggestions of (n) altogether of 200 suggestions. Initially, the researcher determines the Weight value of the word by the True value because it is not a complicated method process. Subsequently, the data are transformed into the ARFF data file type so that it can be used for training data and testing data.

The process of modeling: From the prepared data of all together 200 suggestions, the program of Weka is used for the comparative study of the techniques for sentiment analysis towards the cosmetic products, lipstick, by using the suggestions extracted from Thai reviews video clips or movies from YouTube. The technique of Naive Bayes: NB and the technique of Support Vector Machine: SVM are used by the method of k-fold cross validation determining k = 10. Ninety per cent of the data are for training and ten per cent of the data are for testing in each round of the work. The result from the construction of models shows the Recall value, the Precision value, the F-Measure value, the ROC value, Confusion Matrix and total accuracy.

Table 1: Confusion matrix

	Actual condition	
Variables	Positive	Negative
Test result		
Positive	True Positive (TP_i)	False Positive (FP_i)
Negative	False Negative (FN_i)	True Negative (TN_i)

The comparison of technical performance: This research makes a comparative study on the effectiveness among the techniques by conducting the evaluation according to the concept of information retrieving which are. Precision value, Recall value and F-measure value. The result shows in Table 1and Eq. 5-7:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \quad (6)$$

$$\text{F - Measure}_i = \frac{2 \times (\text{Recall}_i \times \text{Precision}_i)}{\text{Recall}_i \times \text{Precision}_i} \quad (7)$$

RESULTS AND DISCUSSION

The researcher uses the data of 200 sample suggestions and 7,000 attributes to conduct the comparative study of techniques for the Sentiment Analysis towards the cosmetic products, lipstick, by using the suggestions in Thai language extracted from video clips or movies from YouTube in constructing modules in the stated above steps, it is found that the technique of Support Vector Machine offers the accuracy value at 85.17%. On the other hand, the technique of Naïve Bayes offers the accuracy value at only 78.25%. The experiment result of each technique can be concluded as in Fig. 2.

From Fig. 2, it is found that the recall value, the Precision value and the F-measure value from the technique of support vector machine are higher than those from the technique of naive bayes. In this experiment the researcher has controlled on the same number of data in each group, the same number of all

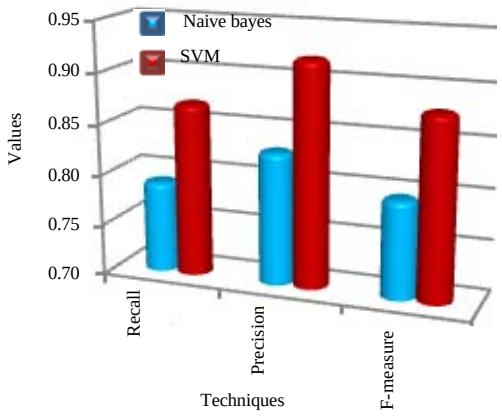

Fig. 2: Graphs comparing the recall value, the precision value and the F-measure value

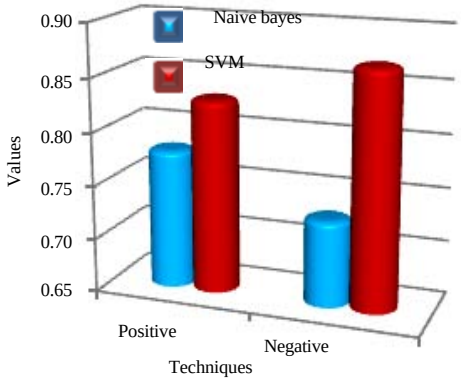

Fig. 3: Graphs comparing F-measure value of each technique

attributes, the same true value used in all data when used in the determination of weight value for each morpheme but the difference is the number of morphemes in the data is not the same. Moreover, in the method of k-fold cross validation the value used for the calculation is 10 of morphemes.

When F-measure value which is used to measure the effectiveness of the sentiment analysis towards the positive and negative features is considered, it is found that the technique of support vector machine offers higher F-measure value as seen in Fig. 3.

CONCLUSION

This research study conducts the comparison between techniques for the Sentiment Analysis towards the cosmetic products, lipstick, by collecting the suggestions in Thai language extracted from video clips or movies from YouTube. Subsequently, two techniques of learning machine: naive bayes and support vector machine are tested with the sample data of 200 suggestions by using the 10-fold cross validation. From the test for effectiveness it shows that the support vector machine can estimate the feeling more correctly than naïve bayes.

The guidelines on the improvement of models on accuracy value from this study suggest improvement on the weight value by considering the frequency of the word appearing in the study in conjunction (TF-IDF) and the analysis of the context around the suggestions for the sarcastic emotion which can be applied for the improvement of the models in order to render more accurate value for each group.

REFERENCES

Chumwatana, T., 2015. Using sentiment analysis technique for analyzing thai customer satisfaction from social media. Proceeding of the 5th International Conference on Computing and Informatics (ICOCI), August 11-13, 2015, Universiti Utara Malaysian, Istanbul, Turkey, pp: 659-664.

Hilao, M.P., 2016. Creative teaching as perceived by english language teachers in private universities. J. Adv. Humanities Soc. Sci., 2: 278-286.

Kaushik, L., A. Sangwan and J.H. Hansen, 2013. Automatic sentiment extraction from YouTube videos. Proceeding of the 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 8-12, 2013, IEEE, Texas, USA., ISBN: 978-1-4799-2756-2, pp: 239-244.

Lima, A.C.E. and D.L.N. Castro, 2012. Automatic sentiment analysis of twitter messages. Proceeding of the 2012 4th International Conference on Computational Aspects of Social Networks (CASoN), November 21-23, 2012, IEEE, Sao Paulo, Brazil, ISBN:978-1-4673-4794-5, pp: 52-57.

Meesad, P. and J. Li, 2014. Stock trend prediction relying on text mining and sentiment analysis with tweets. Proceeding of the 4th World Congress on Information and Communication Technologies (WICT), December 8-11, 2014, IEEE, Bangkok, Thailand, ISBN:978-1-4799-8115-1, pp: 257-262.

Neethu, M.S. and R. Rajasree, 2013. Sentiment analysis in twitter using machine learning techniques. Proceeding of the 4th International Conference on Computing, Communications and Networking Technologies (ICCCNT), July 4-6, 2013, IEEE, Trivandrum, India, ISBN:978-1-4799-3926-8, pp: 1-5.

Phawattanakul, K. and P. Luenam, 2013. Suggestion mining and knowledge construction from thai television program reviews. Proceedings of the International Multi Conference on Engineers and Computer Scientists, March 13-15, 2013, IMECS, Hong Kong, ISBN: 978-988-19251-8-3, pp: 307-312.

Rahman, A., A.A. Zuhair and R.Z. Abid, 2015. Utilization of request mitigators by omani learners of english and native speakers: A comparative study. Int. J. Humanities Arts Soc. Sci., 1: 1-20.

Rosas, V.P., R. Mihalcea and L.P. Morency, 2013. Multimodal sentiment analysis of Spanish online videos. IEEE. Intell. Syst., 28: 38-45.

Sukhum, K., S. Nitsuwat and C. Haruechaiyasak, 2011. Opinion detection in Thai political news columns based on subjectivity analysis. Inf. Technol. J., 14: 27-31.

Taher, M.A., S.P.N. Hrestha, M.M. Rahman and A.K.M.I. Khalid, 2016. Curriculum Linked Video (CLV) as a tool for English Language Teaching (ELT) at secondary school classrooms in Bangladesh. Int. J. Humanities Arts Soc. Sci., 2: 126-13.