International Journal of GEOMATE, Sept., 2019 Vol.17, Issue 61, pp. 74 - 79
ISSN: 2186-2982 (P), 2186-2990 (O), Japan, DOI: https://doi.org/10.21660/2019.61.4700
Special Issue on Science, Engineering & Environment

THE AWARENESS OF ENVIRONMENT CONSERVATION BASED ON OPINION DATA MINING FROM SOCIAL MEDIA

*Kunyanuth Kularbphettong[1]

[1]Faculty of Science and Technology, Suan Sunandha Rajabhat University, Thailand

*Corresponding Author, Received: 20 Dec. 2018, Revised: 23 Jan. 2019, Accepted: 06 Feb. 2019

ABSTRACT: Social media has become an important role in society and people widely use social media platforms to express and criticize their opinion. With the previous studies, there is much research to develop and harvest information and knowledge from social media data for decision making and prediction. The awareness of environmental conservation has become a significant issue nowadays and building and raising environmental awareness is to conserve and protect nature for the benefits of humans. Opinion mining is one of text mining approaches to assess the attitude in a given subject and the attitude may be positive or negative opinion. The paper proposes to conduct social media opinion mining in case of the awareness of environment conservation of Thai people. This study shows that social media effects to build and raise the awareness of environment protection and furthermore this research can apply to other fields and industries aspects as well.

Keywords: Sentiment analysis, Opinion mining, Awareness, Environment conservation, Social media

1. INTRODUCTION

Urban growth in all regions of the world is undeniable, especially in capital cities or economy cities, and the growth of the city means the movement of a large group of people who come together to pursue a career and live. This situation causes enormous consumption and environmental degradation. The environmental crisis is one of the major problems in the world that affects the viability of the most biological systems in the future. Therefore, the requirement to quickly resolve environmental degradation problem is essential to cultivate good moral traits in environmental awareness. Environment awareness is the foundation of the psychology to acknowledge the human actions that cause to deteriorate environment and the result will get back to destroy human life.

Nowadays, social media are increasingly used to disseminate information and express everyone's opinion to an online society where users can communicate and convey their attitude through the social network. Social media has become a new digital medium that has played a significant role in the rapidly changing society in terms of being part of the current news media process. When the internet technology is growing by leaps and bounds, social media takes an action to be a channel for news media and effects the digital economy of the country. There are many kinds of social media platforms according to the purpose of using as following [1]: social networking (Facebook, LinkedIn, Google+), microblogging (Twitter, Tumblr), photo sharing (Instagram, Snapchat, Pinterest), and video sharing (YouTube, Facebook Live, Periscope, Vimeo). Social media becomes a vital part of everyday life and the number of social users has rapidly increased to reach 3.196 billion, up 13 percent year-on-year [2]. Therefore, the use of social media to disseminate information is a new phenomenon of Thai social life and it has an influence on people's behavior and decision making in various fields.

Also, the trend of using online media is likely at a rapid pace and the influence of user-generated content plays an increasingly important role to impact on people's behavior because user generates content is more reliable and generates faster and stronger streams. The power of social movement through online media acts as the communication mechanisms to interact and integrate online community for driving the purpose of social issues in Thai society.

Opinion mining is one of text mining approaches to assess the attitude in a given subject and the attitude may be positive or negative opinion and sentiment analysis is the analysis of emotions and feelings from the text to express the feelings of people such as positive feelings or negative feeling. The information of people think has always been a crucial part to influence decision making. Social media sentiment helps to understand the perceived positive or negative mood what people post on social media This research proposes to conduct social media opinion mining in case of the awareness of environment conservation of Thai people to understand and perceive the feeling of the awareness of environment conservation of Thai people.

2. RELATED WORKS

A literature review of relevant researches for exploration and adaptation information shows that sentiment analysis is the instrument to understand what people think and what feeling they are. It amalgamates with many other fields like natural language processing, statistics, and text analysis to extract the emotional feeling from the text. According to Pak and Paroubek [3], a corpus from the research was collected from Twitter to analyze linguistic sentiment analysis in the English language. Tweets from twitter were analyzed to classify users' categorization in news, politics, and culture [4]. Bollen et al. [5] investigated the tweets posted on Dow Jones Industrial Average (DJIA) to measure positive vs. negative mood and Google-Profile of Mood States (GPOMS) by using Self-Organizing and Fuzzy Neural Network and the result shown that the mood states will directly affect investment decisions and the stock market. Facebook posts were performed on sentiment analysis to measure data available to public domain [6]. The comments of fan page Facebook and tweet of Twitter were classified into some categories, positive, negative, and neutral sentiment by using TF-IDF in Indonesia; Gojek, Grab, and Uber and the results indicate that the comments on social media have evaluated the performance of these business transports online [7]. Barbosa et al. [8] propose a method to identify sentiments from tweets and the data sources was provided by labels to solve different bias. The social media analytics engine, by employed fuzzy similarity-based classification method, was proposed to automatically classify text into sentiment categories (positive, negative, neutral and mixed) and it is able to collect, filter, classify, and analyze social media text data and describes and predicts analytic information on dashboard [9]. Also, explained extends six emotions on new smart services over mobile devices, the approach is used emotional dictionaries and considers linguistic parameters to identify results [10].

There are several approaches proposed to extract information from social media and it can be classified to 2 main techniques as follows: statistical technique and machine learning technique. Linear discriminant analysis (LDA), one of the statistical techniques, was applied to explore the relationship between Facebook fan pages and visitor engagements of the exhibitions [11] and logistic regression was used to analyze textual data from social [12]. Machine Learning was applied to extract information from text data and for example, Gurkhe et al. [13] implemented the machine extracted the polarity (positive, negative or neutral) of social media data set by using naive Bayesian technique. With the advanced technologies, the enhancement of this research is continually moving forward to widespread in various other fields.

3. RESEARCH METHODOLOGIES

This section describes how to conduct this research and the methodologies applied in this project as follows: Data preparation; Feature extraction; Model Building and Testing and Evaluation.

3.1 Data Preparation

In the first stage, data were collected from social media and internet websites and they were posted to express the opinion about environment conservation.

Fig.1 Example of selected Thai hashtags

The example of the hashtag data	Meaning
#รักโลก #รักษ์โลก	Love global
#โลกร้อน	Global Warming
#ประหยัดไฟ	Power saving
#พกถุงผ้า	Carry Your Cloth Bags
#ใช้จักรยาน	Use a bicycle
#รักษาสิ่งแวดล้อม	save the environment
#ขยะรีไซเคิล	recyclable waste
#ปลูกต้นไม้	to Plant a Tree
#คัดแยกขยะ	Waste sorting
#อนุรักษ์พลังงาน	Energy conservation
#ลดเมืองร้อนด้วยมือเรา	Reduce Urban Heat
#อนุรักษ์ป่าไม้	Forest Conservation
#ประหยัดพลังงานไฟฟ้ารักษาสิ่งแวดล้อม	Save energy, save the environment

The researcher collected tweets via the Twitter API and Tweepy API [14] is used to retrieve the tweets and data, like Twitter, hashtag, created tweet time, tweet text, and retweet count was stored in the database. Some selected Thai hashtags were presented in Table 1.

Table 1 The example of data stored as a CSV file

Attributes	Description
user_ID	identifier of the user
screen_Name	Name of user
favorite_count	the number of favorites
retweet_count:	the number of retweets
location	location of the user
created_text	the date of creation
hashtag	type of metadata tag
message	the text of the tweet

The information collected from the tweet was grouped to be 2 parts: the details of the users and the details of the message and data were stored as a CSV file as shown in Table 1.

The second type of data source is news and related websites related to the concept of environmental conservation. Data is the text format that requires to be preprocessing to clean and manage data before analyzing text mining. Preprocessing step of the data set is following: removing URLs, hashtags, username, and symbols; replacing the emotion icons and correcting the spelling words and checking the profanity in tweets. However, the Thai language is different from English because it's no marks or symbols to indicate the scope of each word or sentence. Also, there are many different and specific forms of word patterns. Therefore, it is difficult to check and manage the collected information and the next section will explain to handle the collected input data.

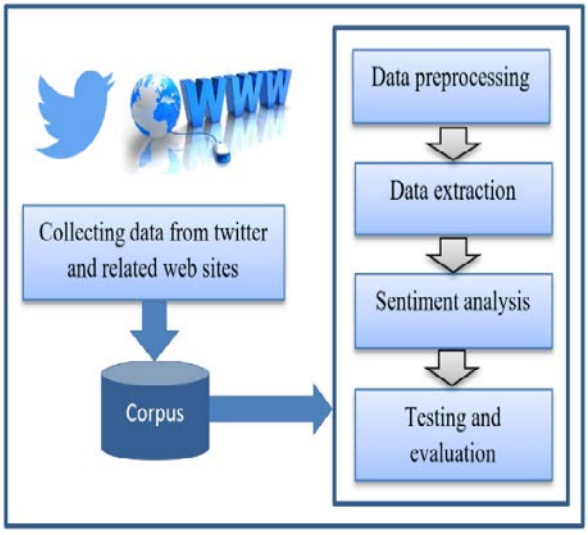

Fig.2 the System Overview

Figure 1 was shown the system overview and data was collected from the tweeter and related websites. LEXiTRON Corpus was used to analyze words and meaning of words [15]. This research was specified domain in the awareness of environment conservation of Thai people collected data from similar related information.

3.2 Feature Selection

Feature Selection is the reduction of data size by reducing the original data size and losing key features using the selection technique. This process of extracting comments is to pull out the feature of the comment to determine what features use this project. A Thai stop word is removed and the insignificant and unmeaning words were eliminated without changing the meaning of the text. The result was presented in figure 3. For instance, the unmeaning words were a non-significant word and in the feature selection process, the feature will be cut off.

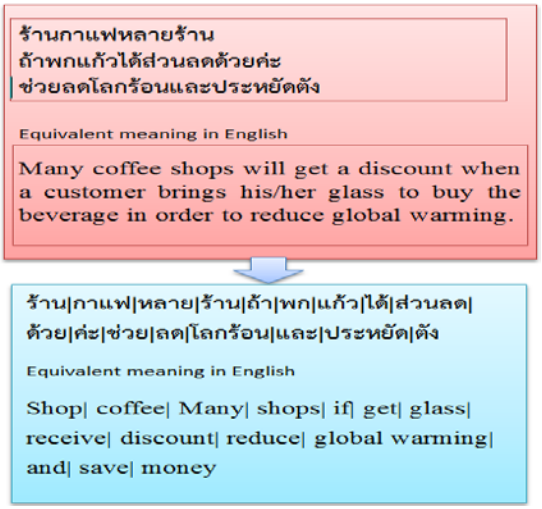

Fig.3 Example Sentence

Text segmentation is one of the important approaches in natural language processing (NLP) because the input text is needed to be tokenized into individual terms or words before being processed further [16]. Thai segmentation technique is applied to extract data and data extraction stage is the process to extract data and labels the training set, annotated positive, negative or neutral, to identify the category of text by hand. The data consists of 2,964 sets and assigned weight each set as a positive, negative and neutral review. Labels were used as 1 for a positive review, 0 for a neutral review and -1 for a negative review. TF-IDF (Term Frequency-Inverse Document Frequency) was applied to determine to index for using for the training process.

$$W_{(f,d)} = TF_{(f,d)} \times IDF_{(f)} \quad (1)$$

$$IDF_{(f)} = log \frac{|D|}{|DF_{(f)}|} \quad (2)$$

Where TF (f, d) presents the frequency of the feature (f) in documents (d) and W (f, d) shows the weight of a feature (f) in (d). |D| explains the number of documents in the training data set (Training Set) and |DF (f)| is the number of documents that feature (f) appears. IDF $_{(f)}$ is inverse document frequency used to identify positive, negative or neutral.

Unlike TFIDF mentioned above, this study divides data into two parts, one for positive data, and one for negative. Furthermore, according to K. Ghag, K. Shah [17], SentiTFIDF was applied to this

project to classify the positive, negative and neutral emotional mood. If the term positive is larger than the term of negative, the term is classified as positive. On the other hand, if the term of negative is larger than the term of positive, the term is classified as negative and then the set of emotional words were stored in the database to be further next process.

3.3 Model Building

This stage uses two different classification models to build the model, Naïve Bayes and Support Vector Machine (SVM). The data were divided to be 2 sets: training and testing sets with a ratio of 70:30.

Naive Bayes classifier, based on Bayes' theorem, is one of the crucial techniques in machine learning used to many fields like customer segmentation or sentiment analysis. The Naïve Bayes estimates a probability to assume the statistical independence of each feature.

$$P(c|x) = \frac{P(X|C)P(c)}{P(x)} \quad (3)$$

P (c|x) is the posterior probability of the attribute X that has set the label of class C. P (x|c) is the likelihood that the data in class C contains attribute X. P(c) is prior probability of class C. P (x) is the predictor prior probability.

Support Vector Machine is the significant supervised algorithm that can solve a classification problem and the concept of SVM is to define data in the feature space and create hyperplane to separate different class labels. SVM classifier was applied to categorize text and given the excellent result [18]. According to Seyyed M. H. Dadgar et al. (2016), TF-IDF and SVM classifier was used to classify two BBC datasets and five groups of 20Newsgroup datasets and the results were likable with 97.84% and 94.93% in precision measurement [19].

3.4 Testing and Evaluation

Accuracy, recall, precision, and F-measure were used to evaluate the performance of text models [20].

$$precision = \frac{TP}{TP+FP} \quad (4)$$

$$recall = \frac{TP}{TP+FN} \quad (5)$$

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

$$F-measure = \frac{2\times precision\times Recall}{precision+Recall} \quad (7)$$

Where TP represents the number of correctly classified documents (True Positive), FP is all documents retrieved (False Positive) and accuracy is the percentage of documents correctly classified, recall is the percentage of relevant documents correctly retrieved (TP) with respect to all relevant documents (TP + FN) and F-Measure has consisted of a single measure Precision (P) and Recall (R).

4. EXPERIMENT RESULTS

After preprocessing data, to label data with polarity words from the comments were applied to set train data of extracted words. Then, sentiment analysis was used to label each comment to three groups of similar comments: positive, negative and neutral based on their polarity scores. The result gained from the experiment by using Naïve Bayes classification and SVM algorithms as shown in table 2. The data from this project was collected from social network and accuracy, precision and F-measure was used to evaluate the effectiveness of classification models.

Table 2 Classification performance

	Accuracy	Precision	Recall	F-score
SVM	80.5	82.5	85.7	83
Naïve bay	72.6	79.7	76.8	74.2

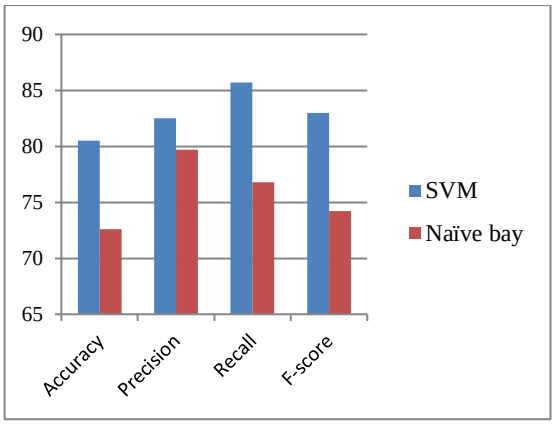

Fig.4 The Results of Classification Performance

The SVM classification model is better than Naïve bay with the accuracy of 80.5, the precision of

82.5, recall of 85.7 and f-measure of 83 as displayed in fig 4.
The results show that the SVM Model is better than the Naïve bay Model. In all classification performance values of the SVM Model indicate higher than these of the Naïve bay. Also, the result when using the model shows that the model can use to detect the concept of environmental conservation and the positive awareness is more the other awareness in the number of words and the average length of Tweet's on data set.

5. CONCLUSION

The awareness of environment conservation has currently become an important topic and to educate and implant environment awareness social media has influenced in society and people. Therefore, the sentiment analysis is the powerful approach to identify opinion, to extract opinion's feature, to classify sentiment, and to display the results in visualization and summarization. This research describes the methods to conduct social media opinion mining in case of the awareness of environment conservation of Thai people. This approach collected data from Twitter and related websites and then it shows how to preprocess and extraction the results. The results show that social media effects to build and raise the awareness of environmental protection and furthermore this research can apply to other fields and industries aspects as well. However, there are some errors with word wrapping and comment extraction because of the typing error and occurring of new social words.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial subsidy provided by Suan Sunandha Rajabhat University.

7. REFERENCES

[1] Warren Jolly., the 6 Most Effective Types of Social Media Advertising in 2018. E-COMMERCE MARKETING. Available: http://www.bigcommerce.com/blog/social-media-advertising/

[2] Dave Chaffey., Global social media research summary 2018. Available: https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/

[3] Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

[4] Muhammet Baykara and Uğur Gürtürk., Classification of social media shares using sentiment analysis. proceeding of International Conference on Computer Science and Engineering (UBMK, 2017) DOI: 10.1109/UBMK.2017.8093536

[5] Bollen, J., Mao, H., and Zeng, X.-J., Twitter mood predicts the stock market. Journal of Computational Science vol. 2(1):1–8, 2010.

[6] Federico Neri, Carlo Aliprandi, Federico Capeci, Monstserrat Cuadros, Tomas., Sentiment Analysis on social media. IEEE/ACM International conference on Advances in Social Networks analysis and mining, pp. 919-926, 2012.

[7] M. H. Saragih and A. S. Girsang,., Sentiment Analysis of Customer Engagement on Social Media in Transport Online. International Conference on Sustainable Information Engineering and Technology (SIET), pp. 24–29, 2017.

[8] L. Barbosa, J. Feng., Robust Sentiment Detection on Twitter from Biased and Noisy Data. COLING 2010 Poster vol, pp. 36-44.2010.

[9] Wang, Z., Chong, C. S., Lan, L., Yang, Y., Ho, S. B., & Tong, J. C., Fine-grained sentiment analysis of social media with emotion sensing, the Future Technologies Conference, San Francisco, CA, USA. Dec. 2016, DOI:10.1109/FTC.2016.7821783

[10] Athena Vakali, Despoina Chatzakou, Vassiliki Koutsonikola, and Georgios Andreadis, Social data sentiment analysis in smart environments extending dual polarities for crowd pulse capturing. 2013, Available: https://pdfs.semanticscholar.org/1b2c/374cae2ee5c57965211bb5b715b4b45d31c3.pdf

[11] Lee, T., Shia, B., & Huh, C., Social Media Sentimental Analysis in Exhibition's Visitor Engagement Prediction. American Journal of Industrial and Business Management, 06(03), pp. 392-400, 2016.

[12] Virgile Landeiro, Aron Culotta, Robust text classification in the presence of confounding bias. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, Feb 2016.

[13] Dhiraj Gurkha, Niraj Pal and Rishit Bhatia, Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification, International Journal of Computer Applications (0975 8887) vol. 99 - No. 13, August 2014

[14] Tweety API, Available: http://www.tweepy.org/
[15] LEXITRON, Available: https://www.nstda.or.th/th/nstda-knowledge/11180-lexitron
[16] Choochart Haruechaiyasak, Sarawoot Kongyoung, and Matthew N. Dailey. 2008. A Comparative Study on Thai Word Segmentation Approaches. In Proceedings of ECTICON 2008.
[17] K. Ghag, K. Shah, SentiTFIDF-Sentiment classification using relative Term Frequency-Inverse Document Frequency. International Journal of Advanced Computer Science and Applications, vol. 05, no. 02, pp. 36-43, 2014.
[18] Z. Liu, X. Lv, K. Liu, and S. Shi, "Study on SVM Compared with the other Text Classification Methods," 2010 Second International Conference on Education Technology and Computer Science, pp. 219–222, 2010.
[19] Seyyed Mohammad Hossein Dadgar et al "A Novel Text Mining Approach Based on TF-IDF and Support Vector Machine for News Classification" 2nd IEEE International Conference on Engineering and Technology (ICETECH), 17th& 18th March 2016
[20] Kunyanuth Kularbphettong, Tiwa Sreekram., The Effectiveness of Thai Spoonerism Application. Part of the Advances in Intelligent Systems and Computing book series (AISC, volume 575)2017, pp 379-384.