International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Improving Responsiveness Conversation of Thai Chatbot through Sentiment Analysis Classification Techniques

Sumitra Nuanmeesri, Lap Poomhiran

Abstract**:** ***Nowadays, internet and social media are play and important role for the business and marketing. Especially, the social media marketing drives the businesses with fierce competition. if there is communication between a large number of customers, it is necessary to have the staff to coordinate thoroughly Resulting in higher expenses as well. Chatbot can be solve this problem by action like a human to deliver a suitable message for their customers. This paper proposes the techniques for analyzing the sentiments that coexist with chat messages or the conversations. Naïve Bayes, K-Nearest Neighbor, and Support Vector Machine techniques were used to classify the sentiments based on Cross-Industry Standard Process for Data Mining. As a result, the highest accuracy is produced by Support Vector Machine with value at 94.60% for improving the chatbot able to communicate effectively with sticker messages.***

Keywords: chatbot, classification, conversation, sentiment analysis, social media marketing.

I. INTRODUCTION

At present, internet technology was used in variety of business and marketing. It plays and important role to the marketing combined with the use of social media to enhance the efficiency of the work even more. Causing marketers or business owners to pay more attention to social media as well. Continuous communication between customers and the marketing department is a process that makes a business buy or helps customers to become confident and able to make more purchases. In the case of large customer communication, business needs more officers to provide any customer questions or inquiries. This is a direct impact on the cost of hiring more personnel. Even though the business has enough officer to communicate with their customers, the conversation may be unsuitable for the customers.

The chatbot was built to solve these problems. Chatbots are one software that is frequently communicated to managers in their online help experience [1]. It has abilities for communication and delivers a suitable message for customer inquiries. Based on the principles of chatbot algorithm, it will compare the keywords in the sentences from the customers to find the most suitable answer back. There are many forms of content to respond to customers such as text, image, hyperlink, sticker, and some actions. A chatbot will random the message with probability from the group of the suitable message that is related to the keywords of customer inquiry.

Today, most people like to communicate continuously or synchronous communication such as chat or instant messaging, audio chart, video call, etc. By popular messages in this day and age, which prefer to use words or texts or short content that is concise, but can get the main topic, idea, or express the feelings clearly. With the limitation of typing a message on social media, it affects the development of reading skills to the user. On the other hand, a long message would be ignored or skipped by the user. For this reason, a short message in the form of non-text is famous for the conversation. This kind of messages are in the form of images or stickers. According to Tech News [2], in July 2019, LINE user in Thailand has up to 65 sets of stickers per user. The sticker can communicate from images that make the user easily. It causes the sticker business to engage in many social media communications. Communicating by sending stickers to each other helps to make a good impression on the conversation. It makes it better to access emotions than messages. It also helps to recover the feeling of conversation better. Moreover, it saves a lot of time when typing text.

Therefore, this research has the idea to analyze the sentiment from Thai messages including the type of sticker answer in the dialogue between buyers and sellers of online products on social media with the classification techniques.

II. RELATED WORKS

Sentiment analysis uses Natural Language Processing (NLP) to investigate and analyze opinion. Presently, there are a number of researchers that analyzed sentiments in communications. There are several sentiment analysis techniques such as:

- *Naïve Bayes (NB):* is an algorithm provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c) [3]. Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence, as formulated in (1) [4].

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
Sumitra Nuameesri*, Assistant Professor, Department of Information Technology Suan Sunandha Rajabhat University, Thailand. Email: Sumitra.nu@ssru.ac.th
Lap Poomhiran, Ph.D. Department of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand. Email: lap_p@windowslive.com

$$P(c|b)=\frac{P(b|c)P(c)}{P(b)} \quad (1)$$

Where:

P(c|b) is the posterior probability of class (target) given predictor (attribute).

P(c) is the prior probability of class.

P(b|c) is the likelihood which is the probability of predictor given class.

P(b) is the prior probability of predictor.

- *K-Nearest Neighbor (KNN):* is a supervised machine learning algorithm useful for classification problems. It measures the distance between the test data and the input and gives the prediction [3]. The most common distance measures are the Euclidean. Euclidean distance is measuring the straight-line distance between two samples, as formulated in (2) [5].

$$dist(X_a, X_b) = \sqrt{\sum_{i=1}^{n}(x_{ai} - x_{bj})^2} \quad (2)$$

Where:

X_a is the posterior probability of class (target) given predictor (attribute).

X_b is the prior probability of class.

- *Support Vector Machine (SVM):* is used for finding a decision plane to separate data into two groups, by building a midline between two groups for widening the scopes so they are as distant as possible [3]. It is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. In two-dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side. The SVM is an artificial neural network that copies neuron characteristics to classify an input space in the form of a high-dimensional dataset in the feature space by using the Kernel Function to adjust the form of the data [6].

To illustrate, (xi, yi), ... , (xn, yn) is set to be to be a sample in the teaching demonstration. n is the number of samples. m is the number of input space dimensions. y is the result which is set to be +1 or -1, as shown in formulae in (3) [7].

$$(x_i,y_i),\ldots,(x_n,y_n) \text{ when } x \in R^m,\ y \in \{+1, -1\} \quad (3)$$

In terms of the linear problem, the high-dimensional data is separated into two groups. The decision plane is calculated according to formulae (4) [7].

$$(w*x)+b=0 \quad (4)$$

When *w* is weight value, and *b* is bias value, it will be used for the calculation according to formulae (5) and (6) in order to classify data [7].

$$(w*x)+b>0 \text{ if } y_i = +1 \quad (5)$$

$$(w*x)+b<0 \text{ if } y_i = -1 \quad (6)$$

Data mining technique in order to create a model predicting the chances of sentiment analysis. The effectiveness of the model was compared by three techniques, including Naïve Bayes, K-Nearest Neighbors and Support Vector Machine. The findings showed that the model developed by the SVM technique had prediction effectiveness with more that 80% accuracy [8], meaning it could be used for utilizing all features and when selecting the most representative features by SVM. Developing the sentiment estimation process based on Thai comments, where 6,000 comments from news, entertainment, and product review websites. The specialists classified the comments into 6 groups of feelings (love, joy, surprise, sadness, fear, and anger; each group consisting of 1,000 opinions). Then, they compared the technical performances between the Naïve Bayes, SVM, and Decision tree. The findings revealed that the SVM technique had the highest accuracy rate with 69.15% [9]. A model for classified emotions towards the Hurricane Sandy hashtag on Twitter into 4 types consisting of positive, anger, fear, and other. The SVM technique was 75.9% accurate, while the Naïve Bayes technique was determined to be 69.1% accurate [10].

III. RESEARCH METHODOLOGY

The research methods based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) [11], the research process in this study consists of 6 stages show as Figure 1.

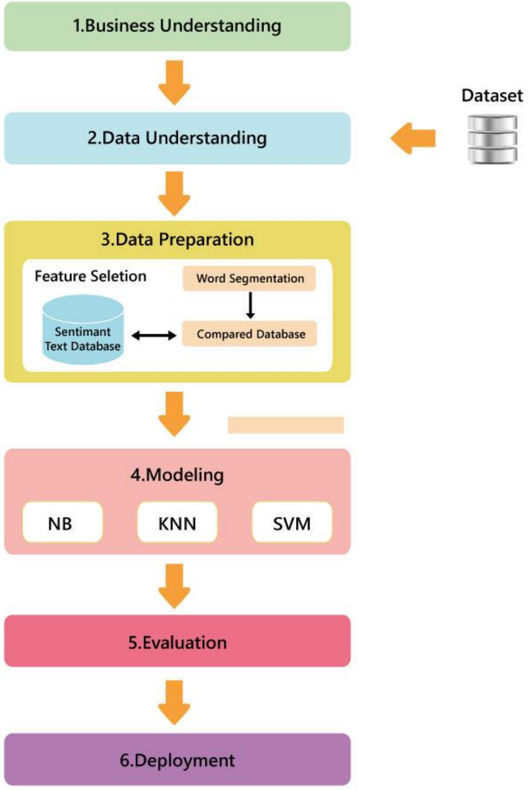


Fig. 1. CRISP-DM of this research.

A. Business understanding

Computerized correspondence advancements have become a necessary part for associations to communicate with their clients.

Numerous organizations offer online administrations by means of live talk interfaces, which empower clients to legitimately cooperate with client support representatives. This type of content-based help experience is a savvy administration arrangement and frequently the favored method for correspondence for youngsters. One innovation which is regularly conveyed to help administration representatives in online assistance experiences are chatbots. Chatbots are programming based frameworks intended to interface with people through content based common language and can be found crosswise over businesses. Chatbot is a computer program that has the ability to hold a conversation with human beings using Natural Language Speech that interacts with human users through automated chat. Chatbots could communicate effectively, sentimental analysis so valuable is its ability to conceptualize social interactions. Some ways sentimental analysis can enhance user experience with chatbots such as [12].

- *Adaptable Customer Assistance*: can alter their responses so that they're aligned with the customer's emotions. These appropriated reactions make for amazing, connecting with encounters with clients.
- *Routing Frustrated or Angry Customers:* Customers who are clearly upset at the beginning of a conversation are quickly recognized and routed to a live rep. That way, the customer will get customized support quickly and efficiently.
- *Customer Categorization:* can identify your happiest and unhappiest users inside your customer base. By segmenting your audience based on customer satisfaction, you can organize support for clients in danger of agitate and prize clients who have exhibited long haul dependability.
- *Record Overall Customer Satisfaction*: can perceive your clients' general view of administration, brand, and results of organization. This furnishes the chatbot with understanding into how clients are feeling before they associate with them.

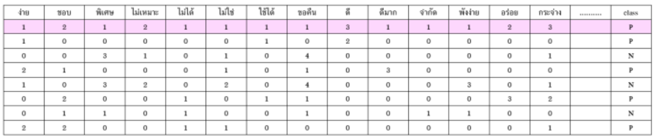
If chatbots are able to communicate effectively, can understand, and can edit their answers to respond in accordance with the mood of the customer, it will result in the decision to buy products and retain customers. This research has collected information about to analyze the features of positive and negative Thai conversations of customers and merchants in Facebook Pages and Facebook Messenger conversation and developed a classification model to improve the chatbot able to communicate effectively with sticker.

B. Data understanding

- *Data Collection:* data were collected to evaluate the sentiment of Thai dialog between customers and traders on Facebook Pages and Facebook Messenger to collect 1,500 sentences. All messages are grouped into the several kinds of words such as noun, pronoun, verb, preposition, conjunction, adjective, and multimedia items (images or stickers).
- *Word Segmentation*: ThSplitLib library was applied to seperate Thai words from sentences. It is the word segmentation algorithm with dictionary-based [13]. ThSplitLib library was applied to separate Thai words from sentences. It is the word segmentation algorithm with dictionary-based [13]. The longest matching method is a core function to predict the vocabulary. Normally, this library supports the sentences from social media.
- *Database Comparison:* Vector Space Model (VSM) method [14] was conducted in this work. Each word was compared with emotional words in database then classified into positive, negative, and neutral of sentimental.

C. Data preparation

After collecting the data on sentences, this work analyzed and cross-checked the data. This section built the matrix (n x m) which consisted of the Attribute (m) from the sentiment text database, there are (n) sentences from the 1,500 sentences. A word was valued based on the accuracy value using the term frequency (TF) method. The data was then converted into an ARFF file. which the last attribute is the class label which has been classified into the negative or positive group to be used for both the trial session and then testing in order to be loaded in Weka program version 3.8.3.



ง่าย	ชอบ	พิเศษ	ไม่เหมาะ	ไม่ได้	ไม่ใช่	ใช้ได้	ขอคืน	ดี	ดีมาก	จำกัด	ฟังง่าย	อร่อย	กระจ่าง	....	class
1	2	1	2	1	1	1	1	3	1	1	1	2	3		P
1	0	0	0	0	0	1	0	2	0	0	0	0	0		P
0	0	3	1	0	1	0	4	0	0	0	0	0	1		N
2	1	0	0	0	1	0	1	0	3	0	0	0	0		P
1	0	3	2	0	2	0	4	0	0	0	3	0	1		N
0	2	0	0	1	0	1	1	0	0	0	0	3	2		P
0	1	1	0	1	0	0	1	0	0	1	1	0	0		N
2	2	0	0	1	1	0	0	0	0	0	0	0	1		P

Fig. 2. Example of data used for modeling.

D. Modeling

This section is the development of the model by using Weka Program with three techniques 1) Naïve Bayes, 2) K-Nearest Neighbor, and 3) Support Vector Machine, by using RBFKernel-typed Sequential Minimal Optimization (SMO), PolyKernel and Puk, LibSVM with Polynomial Kernel and Radial Basis function, and Linear alongside with parameter adjustment. Data separation training set and test set using K-fold Cross Validation and Percentage Split techniques were used to build the sentiment analysis model for the two documents. K-fold Cross Validation separates data into k parts and each k part is equal. In this research, k is set as 10; meaning each of the datasets included 1,500 sentences which were separated into 10 equal parts. Subsequently, a part of the data will be used as the subject for testing model efficacy. This process was repeated to completion. The Percentage Split technique separated data into two groups with random percentages. In this research, the split value was arbitrarily 65%, with 35% of data randomly selected for the trial set, and 65% of data for the real experiment.

E. Evaluation

The model's effectiveness was evaluated by 10-fold Cross validation approach and Percentage Split technique, the split value was arbitrarily 65%, with 35% of data randomly selected for the trial set, and 65% of data for the real experiment. Three model were created by the NB, KNN, and SVM with the accuracy measurement, can be calculated according to formulae (5) [3].

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{5}$$

Where:
TP = Positive tend to have True Positive.
FP = Positive tend to have False Positive.
TN = Negative tend to have True Negative.
FN = Negative tend to have False Negative.

IV. EXPERIMENTAL RESULTS

In this section, we provide the results of accuracy of the sentiment analysis classification model by using the three techniques were employed, including NB, KNN, and SVM. Comparison between using K-fold Cross Validation and Percentage Split techniques, it was determined that the average accuracy rate of K-fold Cross Validation, where K =10, was higher than the percentage split at 65%. According to Table I, when considering the 13 models derived from the parameter adjustments for each technique, the experiment results can be concluded as follows:

Table- I: Comparison of Effectiveness in terms of Accuracy in Each Technique and the Parameters

Technique and Parameter value	K fold=10	split=65%
NB	85.24%	81.28%
KNN	88.20%	87.32%
SVM(SMO), kernel=poly, c=1	84.20%	79.50%
SVM(SMO), kernel=poly, c=50000	90.89%	89.23%
SVM(SMO), kernel=rbf, c=1, gamma=0.01	47.30%	44.12%
SVM(SMO), kernel=rbf, c=50000, gamma=0.01	91.00%	89.65%
SVM(SMO), kernel=rbf, c=50000, gamma=0.1	94.60%	93.79%
SVM(SMO), kernel=puk, c=1, omega=1, sigma=1	92.20%	92.25%
SVM(SMO), kernel=puk, c=50000, omega=1, sigma=1	92.60%	92.94%
LibSVM, kernel=linear, cost=1000, gamma=10	91.40%	90.00%
LibSVM, kernel=polynomial, cost=1000, gamma=10	89.70%	88.25%
LibSVM, kernel=radial, cost=1000, gamma=10	93.87%	91.65%
LibSVM, kernel=sigmoid, cost=1000, gamma=10	91.00%	87.05%

From Table 1 data indicated that data separation using the K-fold Cross Validation, where K=10, affected the accuracy of the SVM(SMO), kernel=puk, c=50000, omega=1, sigma=1 with K-fold =10 achieved the highest accuracy is 94.6%.

The results of the accuracy comparison between Percentage Split at 65% By K-fold Cross, where K=10 of the sentiment analysis classification model by using the three techniques were employed, including NB, KNN, and SVM show in Figure 3.

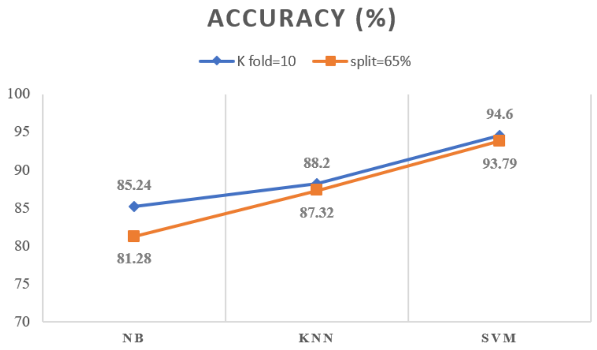


Fig. 3. Results of the Accuracy Rate Comparison Between Percentage Split at 65% By K-fold Cross, where K =10

The research methodology was conducted based on CRISP-DM. The research findings revealed that data separation using the K-fold Cross Validation, where K=10, affected the accuracy of the SMV function, in which the accuracy rate was 94.6% show in Figure 4.

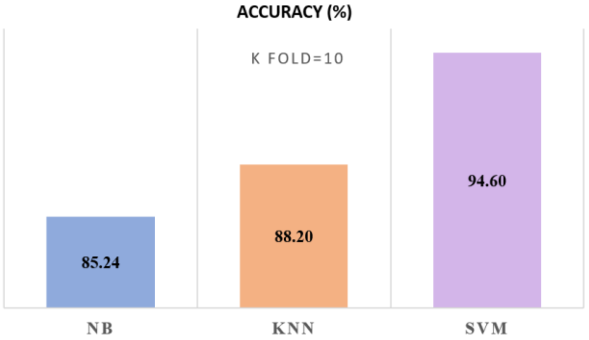
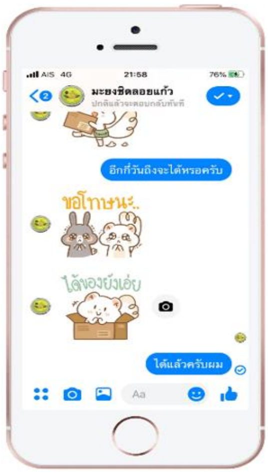


Fig. 4. Results of the evaluation of the model's effectiveness.

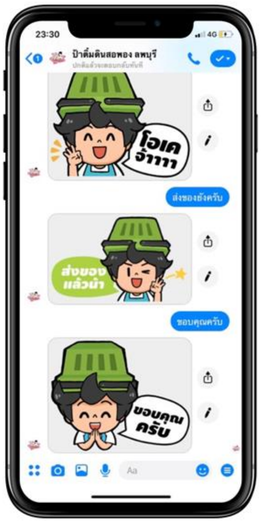
After obtaining the most effective sentiment analysis model, which was created by the SVM method the research team applied the SVM to develop the prototype model of sentiment analysis and response conversation in sticker on Facebook Chatbot for community products for Thai user, as illustrated in Figure 5 and Figure 6.



อีกกี่วันถึงจะได้หรอครับ

ได้แล้วครับผม

Fig. 5. The screen displaying the chatbot conversation by sticker for community product (caring).



ส่งของยังครับ

ขอบคุณครับ

Fig. 6. The screen displaying the chatbot conversation by sticker for community product (happiness).

V. CONCLUSION

This research presents the development of the improving responsiveness conversation of Thai chatbot on social media marketing through sentiment analysis classification techniques. Three sentiment analysis techniques were employed, including Naïve Bayes (NB), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). The research methodology was conducted based on CRISP-DM. The research findings revealed that the model developed by SVM separation using the K-fold Cross Validation, where K=10 achieved the highest accuracy (94.6%). After applying the model to the development of an edible and poisonous mushrooms classification model, it was found out that the model worked effectively. This conforms to the study conducted by Brahimi et al. [8], which introduced the application of the SVM technique to compared the effectiveness of several sentiment analysis techniques such as Naïve Bayes, K-Nearest Neighbor and Support Vector Machine, the research results showed that SVM technique could provide better prediction than Naïve Bayes (NB), K-Nearest Neighbor (KNN).

ACKNOWLEDGMENT

We would like to express our gratitude to the Institute for Research and Development, Suan Sunandha Rajabhat University, who offered us the opportunities to conduct this research.

REFERENCES

1. Google & Temasek, *e-Conomy SEA 2019 report*, Think with Google, 2019.
2. Tech News. (2019, July 23). LINE stepped into the seventh year, revealing that Thai people have LINE stickers up to 65 sets per person. [Online]. Available: https://www.techoffside.com/2019/07/line-stickers-awards-2019
3. H. W. Ian, F. Eibe, and A. H. Mark, “Data Mining: Practical Machine Learning Tools and Techniques,” 3th Edition, Burlington, 2011.
4. Naive Bayesian. (2019, June 25). [Online]. Available: https://www.saedsayad.com/naive_bayesian.htm
5. K-Nearest Neighbors. (2019, June 29). [Online]. Available: https://bradleyboehmke.github.io/HOML/knn.html
6. SVM (Support Vector Machine) - Theory. (2019, June 29). [Online]. Available: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72
7. S. Nuanmeesri, “Sentiment Analysis of Thai Sounds in Social Media Videos by using Support Vector Machine,” *Indian Journal of Science and Technology*, vol. 12, no. 1), 2019, pp. 1–8.
8. B. Brahimi, M. Touahria, and A. A. K. Tari, “Data and text mining techniques for classifying Arabic tweet polarity,” *Journal of Digital Information Management*, vol. 4, no.14, 2016, pp.15–25.
9. A. Chaisal, and R. Sukhahut, “Emotion Prediction from Thai Comments Using Machine Learning Technique,” *The 9th National Conference on Computing and Information Technology*, 2013, pp. 260–266.
10. J. Brynielsson, F. Johansson, C. Jonsson, and A. Westling, “Emotion classification of social media posts for estimating people’s reactions to communicated alert messages during crises,” *Security Informatics*, vol. 3, no. 1, 2014, pp. 1–11.
11. C. Shearer, “The CRISP-DM model: The new blueprint for data mining,” *Journal of Data Warehousing*, vol. 5, no. 4, 2000, pp. 13–22.
12. How Chatbots Use Sentiment Analysis to Improve Customer Satisfaction. (2019, June 25). [Online]. Available: https://blog.hubspot.com/service/chabot-sentiment-analysis
13. Thai Split Library: THsplitlib 3.0 [Computer software]. (2019, June 29). [Online]. Available: http://www.alogik.com/thsplitlib/
14. M. Melucci, “Vector-Space Model,” *Encyclopedia of Database Systems*, Springerlink, 2009.

AUTHORS PROFILE

Sumitra Nuanmeesri, was born in Ratchaburi, Thailand. She has obtained a Ph.D. in Information Technology at King Mongkut’s University of Technology North Bangkok (KMUTNB), Thailand. She is Assistant Professor and lecturer in Information Technology Department, Faculty of Science and Technology, at Suan Sunandha Rajabhat University (SSRU) in Bangkok. She has taught web design and programming, object-oriented programming, and internet marketing. Her research interests include mobile application programming, information and image retrieval, big data and mining, machine learning, recognition of text and speech, supply chain management system, augmented and virtual reality development.

Lap Poomhiran, is currently a Ph.D. student in Information Technology, Faculty of Information Technology, King Mongkut’s University of Technology North Bangkok (KMUTNB), Thailand. His research interests include web and mobile programming, augmented reality (AR) and virtual reality (VR).