International Journal on Advanced Science Engineering Information Technology

Vol.13 (2023) No. 4
ISSN: 2088-5334

Customer Needs Classification from Online Social Media Using Bag-of-Concepts Representation

Kanjana Laosen [a], Adisak Intana [a,*], Phisitchai Chuaynukul [a]

[a] *Andaman Intelligent Tourism and Service Informatics Center, College of Computing, Prince of Songkla University, Phuket, Thailand*
Corresponding author: [*]adisak.i@phuket.psu.ac.th

***Abstract*— Social media platforms are now a very powerful tool for digital marketing strategy because it helps companies to be in direct contact with their customers. A communication problem in digital social media is several customer needs phrases posted on social media, making it difficult for businesses to find relevant posts and respond to customers immediately. Therefore, knowing and understanding the customer requirements for a product can help the product owner to propose the right product to the right customer. This study focuses on understanding customer needs in Thai and classifying them into certain concepts. This study aims to classify customer needs for products in online social media community groups. The model focuses on understanding Thai customer need phrases. We then use a bag of concepts representation, including pattern analysis that applies n-grams together with POS and synonym replacement, conceptual analysis, pattern matching, and class labeling that applies concept sets obtained from the FP-Growth algorithm and represents TD-IDF value in a bag of concepts. The effectiveness of the proposed model is evaluated on five classification algorithms, including Decision Tree, Support Vector Machine, Naïve-Bayes, K-Nearest Neighbor, and RBF Neural Network. The results show that Decision Tree can yield higher accuracy and F-measure values than the others. As this study is an initial step of a personalized product recommendation system in the future, this study will apply the model to the remaining domains for future work.**

***Keywords*— Conceptual level analysis; the bag of concepts; online social media; customer needs classification.**

Manuscript received 4 Mar. 2022; revised 39 Apr. 2023; accepted 13 May 2023. Date of publication 31 Aug. 2023.

I. INTRODUCTION

Generally, many businesses promote their products to their customers through radio, television, and websites. Recently, promoting products to target customers is frequently based on customer information such as historical purchases and interactive behavior data [1], [2]. Although this strategy is well utilized, promoting the right product to the right customer at the right time using the right communication channel is more influential and essential for digital marketing today. Normally, a customer need is terminated within a short period when it is met and responded to by the product owner. Social media platforms [3], [4] are now very powerful communication tools because they allow users to create content quickly. Nowadays, it is also a powerful tool for digital marketing strategy [5] because it helps companies or product owners directly contact their customers. The customer can write a post about the product they need, and the product owner then replies to that post proposing the right product to its customer. Therefore, social media lets the product owners know and understand the customer's needs. However, a communication problem in digital social media is that several customers need phrases posted in social media groups, making it difficult for businesses or product owners to find relevant posts and respond immediately.

For this reason, knowing and understanding customer needs for a product immediately can help the product owner to propose the right product to the right customer at the right time [6], [7]. The Thai language is naturally difficult, causing several problems. The first problem concerns writing with informal language in social media. In the Thai language, the members of social media normally write informal language with short sentences. Therefore, the customer needs posted on social media are usually in phrase format instead of sentence format; the sentence's subject is normally omitted. Although the subject is often omitted, the phrase can still be understood by Thai members but not by a machine. The second problem concerns the order and position of words in phrases or sentences. The word order pattern is normally subject-verb-object. Although sometimes a word is in the middle of a sentence and sometimes at the end, the meaning is still semantically correct. Referring to these two problems above,

this study focuses on understanding customer needs in terms of concept or semantics rather than syntax. As this study is an initial step of a personalized product recommendation system [1], [8] in the future, the objective of this study is to understand customer needs in Thai and classify them into certain concepts. This study deals with two main research questions:

- How to understand the customer needs for products in the Thai language?
- What is the optimal algorithm for classifying customer needs?

This study proposes a classification model for customer needs using a bag of concepts representation to answer these two questions. We collected customer need phrases from two online social media platforms, Facebook and Web board, where people communicate about buying and selling products in Thai. Facebook group is where the members can post their needs about products and comment on any posts. Phuketall.com is a web board allowing members to post their product needs. As this study is an initial step of our research, there are several domains of products to be achieved; we first select only one product domain. In summary, our contributions are as follows: 1) we present a bag of concepts representation for customer need classification by including pattern analysis that applies n-grams together with POS and synonym replacement, conceptual analysis, pattern matching, and class labeling that applies concept sets obtained from FP-Growth algorithm and represents TD-IDF value in a bag of concepts, 2) we evaluated our model on five classification algorithms including Decision Tree, Support Vector Machine, Naïve-Bayes, K-Nearest Neighbor, and RBF Neural Network for finding an optimal algorithm.

A. Word Segmentation

The Thai language is naturally written without spaces between the words in sentences. Each paragraph has no word and sentence boundaries [9]. Japanese and Chinese have no word boundaries [10], [11]. Therefore, several NLP tasks must segment sentences into words [12], [13]. Many techniques develop the existing Thai word segmentation programs. Each program gives different results, such as LexTo [14], which uses the longest matching and backtracking technique, and TLex [15], [16], which uses a conditional random field. Currently, the most popular Thai word segmentation program is LexTo. LexTo is a dictionary-based word segmentation using the longest matching and backtracking technique and uses a LEXiTRON dictionary [17] that contains 42,222 words. It is developed by The National Electronics and Computer Technology Center (NECTEC). This study applies LexTo for Thai word segmentation because it outputs the expected tokenized results.

B. Bag of Concepts (BOC) and Term Frequency Inverse Document Frequency (TF-IDF)

Bag-of-Concepts (BOC) is a model proposed by Sahlgren and Cöster [18]. BOC, in which the semantics of a document is represented as a set of concepts or meanings. BOC represents a document by the frequency of concepts and normally is concerned with the semantics of the document. This model introduced increases in the performance of support vector machines in text categorization. The document classification tasks are usually performed by BOC, such as using bag-of-concepts representation for document classification [19] and learning concept embedding for document classification [20].

Term frequency-inverse document frequency (TF-IDF) [21] is a term weighting technique that determines the relevance of a word to a document. It uses term frequency and document frequency to create a weighted term for document representation. TF concerns how frequently a term occurs in a document, and IDF concerns how important a term is. TF-IDF is performed well in Word and document representation tasks such as sentiment analysis and topic modeling [22]. The BOC and TF-IDF are also important for semantic analysis [23]–[26]. TF-IDF is calculated as in equation (1).

$$tfidf_{ik} = tf_{ik} \times \log\left(\frac{N}{n_i}\right) \quad (1)$$

Where, tf_{ik} Denotes the number of occurrences of term i in the document number k, n_i Denotes the number of documents containing i, and N is the total number of documents. The bag of concept representation in this study differs from several other studies in which this study uses an n-gram to structure the pattern of words and calculate the TF-IDF value for a bag of concept representation. Moreover, structuring a bag of concepts is based on natural language processing in each step.

C. FP-Growth

FP-Growth [27] is an efficient algorithm for calculating and finding frequent transaction item sets. FP-Growth creates FP-Tree from the database and finds the frequent item sets from FP-Tree created. The study of [28] took advantage of FP-Growth for classification. Specifically, our study applied the FP-Growth algorithm for the labelling class.

D. Customer Need Phrases Classification

Customer needs classification in this study concerns classifying customer needs into certain classes; therefore, we evaluate our model on several classification algorithms to find an optimal algorithm. This study classifies customer needs on five algorithms for evaluating the effectiveness of our model as follows:

The first algorithm is Decision Tree (DT) [29]. It is a simple and popular algorithm for classification [30]. It is a tree structure that consists of branches and nodes. Each branch is a decision path or possible alternative. The root node is the topmost node, the decision node is a decision to be made, and the leaf node is a decision class in the decision path. C4.5 is an algorithm improved from ID3, which uses attribute selection measures to select attributes with the largest Information Gain. C4.5 is performed by several domains, such as customer purchase behavior prediction [31], landslide susceptibility analyses [32], education [33], and sentiment analysis [34]. The second algorithm is support vector machine (SVM) [35]. It is a supervised machine-learning algorithm. It finds an optimal separating hyperplane that maximizes the margin between two classes. Many studies related to SVM algorithms, such as sentiment analysis for Twitter [36]–[38], sentiment analysis for COVID-19 [39], and learning and teaching [40]. The third is Naive-Bayes (NB) [41], a

classification algorithm using the Bayes probability theorem to classify unlabeled observations. Several studies apply NB for several tasks, such as sentiment analysis [42], [43], chatbot [44] and opinion mining [45], [46]. The fourth is K-Nearest Neighbor K-NN [47], a simple machine-learning algorithm. This algorithm uses similarity measures such as Euclidean distance, Correlation distance, and Cosine similarity. Several studies classify using different datasets, such as document classification [48] and recommendation systems [49]. The last algorithm is RBF Neural Network (RBFN). It is a feedforward network composed of three layers: the input, hidden, and output layers [50].

II. Material and Method

This section proposes a customer need phrases classification model for the Thai language in online social media. Our model consists of seven steps as follows:

A. Data Collection and Preprocessing

Firstly, we collect the customer need phrases on products from social media and store them in the database created. As the phrases collected concern several domains, we filter only phrases concerning customer needs on products by using keywords about finding products such as "ต้องการ (Want)", "ต้องการหา (Want to find)", and "กำลังมองหา (Looking for)" as shown in Table 1. Since house and rental houses are essential for human life and are mostly posted by members, buying and selling properties is our model's first domain to be an initial case study. A total of 10,098 customer need phrases, including 5,419 phrases from Facebook and 4,679 phrases from web boards, are obtained from this step. The examples of customer need phrases in Thai language output from this step are shown in Table 2. From Table 1., we translate Thai words into English for easier understanding. In Thai, the members normally post short sentences without a subject. As some phrases contain symbols, HTML tags, or special characters that affect the output of the word segmentation process in the next step, such as "
", "+_+", "^^", "!" and "*-*". Moreover, these symbols do not detract from the meaning of the sentence. We then remove them from all phrases. We also replace the repeated characters, such as "ๆๆๆๆ" with only one character.

TABLE I
THE EXAMPLES OF KEYWORDS

Thai keywords	Translate to English	Thai keywords	Translate to English
ต้องการ	Want	มองหา	Looking for
ต้องการหา	Want to find	กำลังหา	Looking for
ต้องการซื้อ	Want to buy	กำลังมองหา	Looking for
ต้องการหาซื้อ	Want to buy	ตามหา	Find
อยากได้	Want	อยากจะซื้อ	Want to buy
หาซื้อ	Buy	อยากซื้อ	Want to buy
ใครมี	Who has..?	ใครขาย	Who does sell..?
ที่ไหนมี	Where is there..?	ที่ไหนขาย	Where to sell..?

TABLE II
THE EXAMPLES OF CUSTOMER NEED PHRASES IN THE THAI LANGUAGE

ID	Customer need phrases
1	อยากซื้อบ้านราคาไม่เกิน 1.5 ล้าน มีตรงไหนบ้างค่ะรบกวนแนะนำให้หน่อย (want to buy a house, not more than 1,500,000 Baht, where is it? please recommend me)
2	ขออนุญาตครับหาบ้านแถวสามกองเดือนละไม่เกิน 8,500 เทศบาลนครภูเก็ต (Execute me, find a house around SamKong in Phuket, not more than 8,500 Baht per month)
3	กำลังหาบ้านดับราคาไม่เกิน 2 ล้านบาท โซนสวนหลวง นาคา ดาวรุ่ง (Looking for a house, not more than 2,000,000 Baht around SuanLoung, Naka, Dawrung)

B. Word Segmentation

The customer needs texts obtained from step *A*. are segmented by LexTo. The output of this step is the texts segmented, as shown in Table 3. Afterword segmentation processing, spaces are still encountered from texts segmented, and we also remove spaces from the segmented texts, as shown in Table 3.

TABLE III
THE EXAMPLE OF WORD SEGMENTATION

Original text	Segmented text	After removing space
หาบ้านราคาไม่เกิน1,500,000 บาท มี 2 ห้องนอน มีที่จอดรถยนต์ครับ	หา | บ้าน | ราคา | ไม่ | เกิน | 1,500,000 | บาท | | มี | 2 | ห้องนอน | | มี | ที่จอดรถ | ยนต์ | ครับ|	หา | บ้าน | ราคา | ไม่ | เกิน | 1,500,000 | บาท | มี | 2 | ห้องนอน | มี | ที่จอดรถ | ยนต์ | ครับ|

C. Pattern Analysis

This step concerns pattern analysis that consists of three processes as follows:

1) Pattern Selection: In this step, the unique 6,812 words are obtained from the 10,098 needed phrases. We then select only words that concern product or product properties such as "ราคา price", "ที่จอดรถ parking", and "ห้องนอน bedroom". We obtained a total of 281 words that concern products and used them as keywords. From 10,098 phrases, we select only phrases that contain at least one of the 281 keywords and split those phrases into 2, 3 and 4-gram results. We then select unique n-grams results and sort them by frequency, and each unique n-grams result represents a unique pattern. All patterns will be created as pattern trees. We use 281 keywords as branches of pattern trees; each keyword can be a branch of a tree. The sub-branches of the tree represent the possible patterns according to n-grams. Each pattern in the same branch concerns the same product property. For example, in Fig. 1, the branch name "ที่จอดรถ parking" represents the property "parking". The "f" represents the frequency of the pattern. For example, "ที่จอดรถ (parking) | (f=433)" means that the frequency of pattern concerning "ที่จอดรถ (parking)" is 433. The branch "ที่จอดรถ parking | ยนต์ (motor) | (f= 171)" in the sub-branches level 2 of the tree represents the pattern based on 2 grams, and the branch "ที่จอดรถ parking | ยนต์ (motor) | ได้" in the sub-branches level 3 of the tree represents the pattern based on 3-grams, respectively. Therefore, a tree represents a pattern tree consisting of more than one branch; each branch represents a possible pattern.

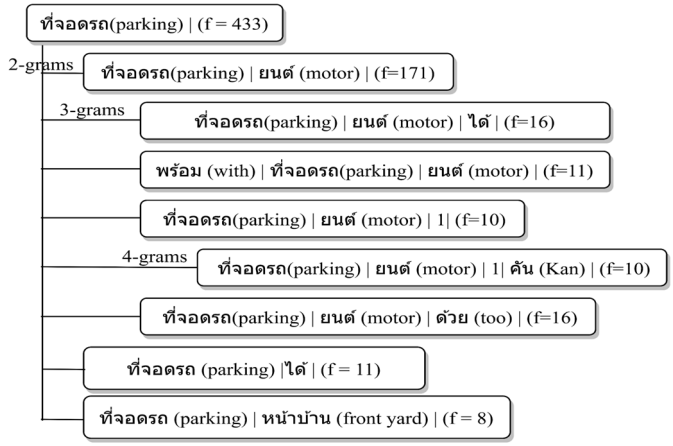


Fig. 1 The example of a pattern

2) Part-of-Speech Tagging: In this step, Thai - English Electronic Dictionary LEXiTRON identifies part of speech in the pattern trees. The part of speech of each word obtained from LEXiTRON is identified in the parentheses "()"; for example, "(N)" represents a noun and "(V)" represents a verb, as shown in the examples *E*1 – *E*3 below.

- *E*1:ครัว (N)
- *E*2:ที่จอดรถ (N)
- *E*3:NUMERIC 1 | นอน (N) | NUMERIC 1| นั่งเล่น (V)

3) Synonym Replacement: After identifying part of speech for each pattern, synonym replacement is considered because some words in the pattern have synonym words that mean exactly or nearly the same and can be used instead. We use the synonyms of each word from AsianWordNet [51]. The synonym words obtained are represented in the brackets "{}" as shown in example *E*4 below.

- *E*4:NUMERIC | ห้องครัว (N) | ห้องนั่งเล่น (living room)(N) {ห้องสันทนาการ (N)}

D. Conceptual Analysis

The previous step concerns pattern analysis and this step concerns semantic analysis. Each pattern obtained from the previous step will be manually defined into 21 concepts such as "Toilet", "Price", "Location", "Bedroom", "Parking", "and "Room". Each pattern can be defined into one or more concepts, as shown in the examples *E*5 – *E*6 below.

- *E*5:ครัว (kitchen) (N) => [Kitchen]
- *E*6:NUMERIC | ห้องครัว (N) | ห้องนั่งเล่น(living room)(N) {ห้องสันทนาการ(N)} => [Kitchen] [Living room]

Referencing the examples above, the example *E*5: "ครัว (kitchen)(N)" can be defined into only one concept; that is the concept "Kitchen", the example *E*6: "NUMERIC | ห้องครัว (N) | ห้องนั่งเล่น (living room) (N) {ห้องสันทนาการ(N)} => [Kitchen] [Living room]" can be defined into two concepts; "Kitchen" and "Living room". After defining concepts for all patterns, these patterns defined are validated by three experts in NLP and ontology. This evaluation aims to evaluate if the patterns and concepts we manually defined are semantically aligned. In cases of disagreement, the experts can comment on improvement. The results from three experts are considered to improve our patterns. The output of this step is named conceptual pattern trees. Finally, we have verified 1,308 conceptual patterns.

TABLE IV
THE EXAMPLES OF BAG OF CONCEPTS

ID	Type	Toilet	Price	Location	Bedroom	…
1	T	F	F	F	F	…
2	F	F	T	T	T	…
3	F	F	T	F	F	…
…	…	…	…	…	…	…
10,098	T	T	T	F	F	…

E. Pattern Matching

We create a bag of concepts (BOC) by mapping between 21 conceptual pattern trees, and the segmented customer need phrases. We create a mapping table. The attributes (columns of the table) represent the conceptual pattern trees. The 10,098 segmented texts obtained from step *B.* are mapped to the conceptual pattern trees, as shown in Table 4. The column "ID" represents the number of segmented texts. The next 21 columns represent the concepts of the product. For example, "Type" represents the "Type of house" concept. "Toilet" represents the concept of "Toilet", and "Price" represents the concept of "Price". The value in the table can be only two values: "T" and "F". T (True) means that we found patterns of a conceptual pattern in given segmented text, and F (False) means that we found no patterns of a conceptual pattern in given segmented text. The bag of concepts finally consists of 713 unique conceptual patterns mapped from 10,098 segmented texts.

F. Class Labelling

This step consists of two processes as follows.

1) Frequent Concept Set: All 713 unique conceptual patterns obtained from the previous step are largely to be classes for classification because too many classes could lead to poor effectiveness in the output of classification. We, therefore, need to reduce the number of classes by finding the frequent concept set (FCS) from the FP-Growth algorithm. Each set of co-occurred concepts obtained from FP-Growth will be a class and will be added to the frequent concept trees (FCT). This process sets the maximum size of items as 5. An example of a concept set with different sizes is shown in Table 5. We obtain 12 FCTs with 95 sets of occurrence concepts; each consisting of at least two. For example, in the first row of Table 5., "Price", "Bedroom", and "Kitchen" are three occurrence concepts. The examples of the FCTs are shown in Fig. 2. For example, the (Bedroom) represents the concept that started with "bedroom", the (Location) represents the concept that began with "location", and the (Price) represents the concept started with "price", respectively.

TABLE V
THE EXAMPLES OF CONCEPT SETS WITH DIFFERENT SIZES

Size	Item1	Item2	Item3	Item4	Item5
3	Price	Bedroom	Kitchen		
3	Location	Bedroom	Parking		
3	Location	Bedroom	Toilet		
3	Bedroom	Parking	Toilet		
4	Rent/Buy	Price	Location	Bedroom	
4	Rent/Buy	Price	Location	Parking	
5	Rent/Buy	Price	Location	Bedroom	Parking
5	Rent/Buy	Price	Location	Bedroom	Toilet

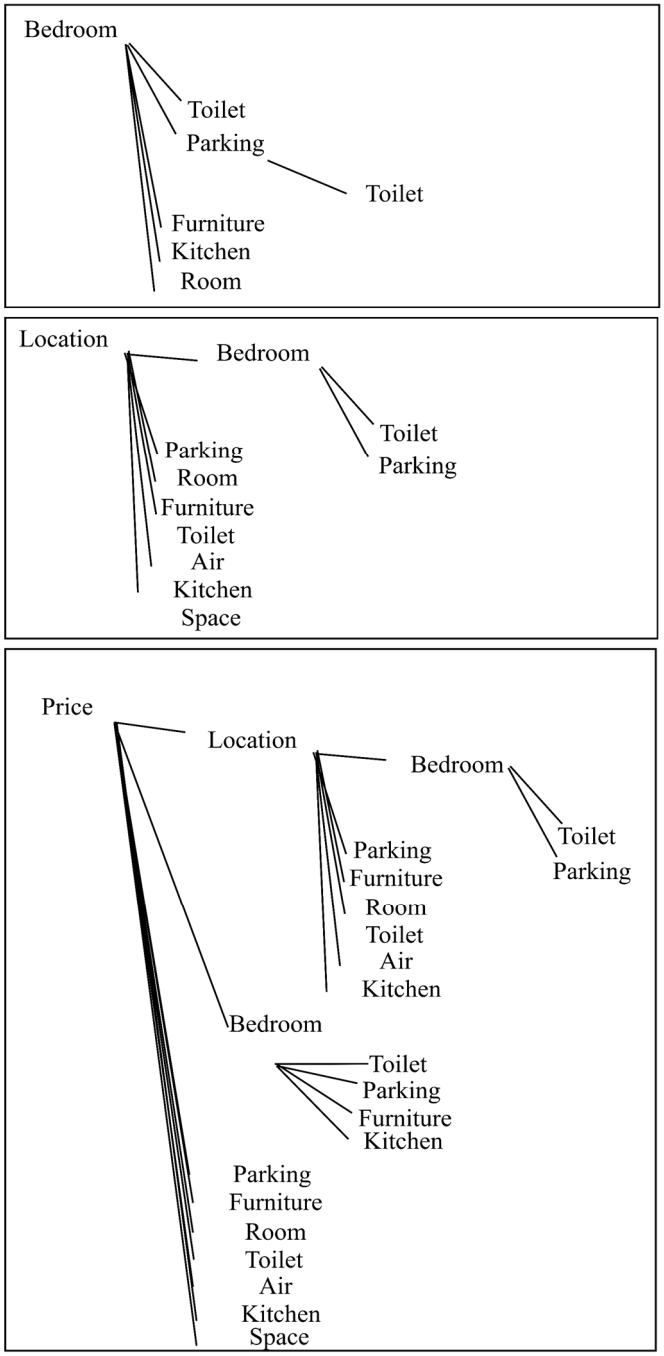
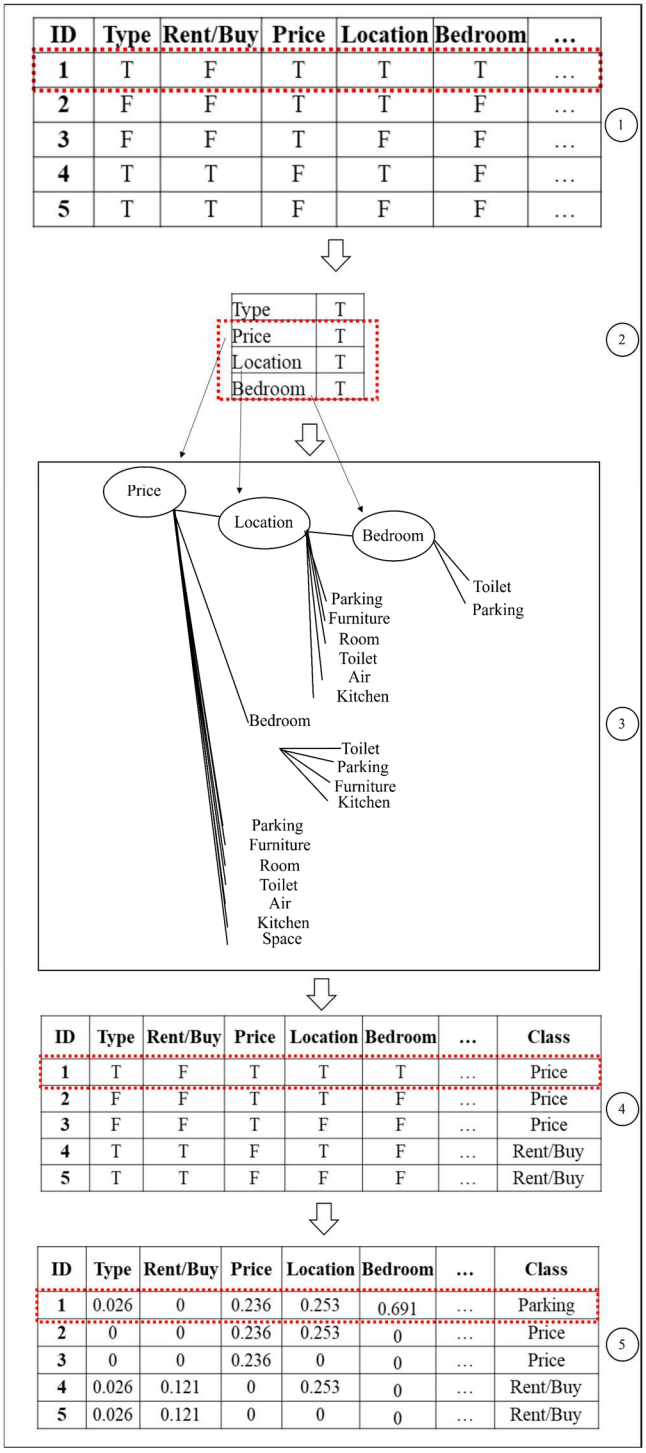


Fig. 2 The examples of frequent concept trees

2) Class Labelling and Term Weighting: In this step, referencing Fig. 3, we use BOC as a train set. We then define the class for every record, as shown in Fig. 3(step 1). We first select only the concepts that represent value "T" in each record and match those concepts with the concept patterns in the FCT as shown in Fig. 3(step 2), in cases where no concepts are matching those in FCT, for example, there are no concepts that start with the concept "Type", we then consider the next idea in order; that is "Price" as shown in Fig. 3(step 3). Those concepts are finally matched with only three concepts; "Price", "Location", and "Bedroom", based on the fact that there is no concept of "Type". We then use the first concept matched as the class of this record, as shown in Fig. 3(step 4). Some records cannot be matched with any concepts from the matching process.



ID	Type	Rent/Buy	Price	Location	Bedroom	…
1	T	F	T	T	T	…
2	F	F	T	T	F	…
3	F	F	T	F	F	…
4	T	T	F	T	F	…
5	T	T	F	F	F	…

1

Type	T
Price	T
Location	T
Bedroom	T

2

3

ID	Type	Rent/Buy	Price	Location	Bedroom	…	Class
1	T	F	T	T	T	…	Price
2	F	F	T	T	F	…	Price
3	F	F	T	F	F	…	Price
4	T	T	F	T	F	…	Rent/Buy
5	T	T	F	F	F	…	Rent/Buy

4

ID	Type	Rent/Buy	Price	Location	Bedroom	…	Class
1	0.026	0	0.236	0.253	0.691	…	Parking
2	0	0	0.236	0.253	0	…	Price
3	0	0	0.236	0	0	…	Price
4	0.026	0.121	0	0.253	0	…	Rent/Buy
5	0.026	0.121	0	0	0	…	Rent/Buy

5

Fig. 3 The example of class labeling and term weighting process

3) We then put them in the class named "Undefined". Therefore, there are 13 classes, including "Undefined", for classification. For term weighting, we calculate the TF-IDF and replace "T" with the TF-IDF value in BOC, as shown in Fig. 3(step 5). The column "ID" represents the customer's need ID. The column "Class" represents the class. The 21 remaining columns are 21 concepts. For example, the number 0.026 between the row "ID#1" and the column "Type" means that the concept "Type" occurs in the customer need ID#1 with the TF-IDF value 0.026. The number 0 between the row "ID#5" and the column "Price" means that the concept "Price" does not occur in the customer need ID#5.

G. Classification and Evaluation

In this step, the effectiveness of this model is evaluated on five classification algorithms including DT, SVM, NB, K-NN and RBFN. We perform the classification algorithms by using 10-fold cross validation. To assess the effectiveness of this model, we use accuracy, precision, recall and F-Measure, which are defined as follows:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

$$F-measure = \frac{2\times Precision\times Recall}{Precesion+Recall} \quad (5)$$

Where TP is truly positive, which refers to the number of predicted positives that are correct, FP is false positive, which refers to the number of predicted positives that are incorrect, TN is truly negative, which refers to the number of predicted negatives that are correct, and FN is a false negative, which refers to the number of predicted negatives that are incorrect.

III. Result and Discussion

We first evaluate the effectiveness of K-NN using a K value from 1 to 10. The result shows that classifying with K=4 gives higher accuracy than the others, as shown in Table 6.

TABLE VI
THE ACCURACY RESULTS OF CLASSIFICATION USING K-NN

K value	Accuracy (%)	K value	Accuracy (%)
1	99.69	6	99.75
2	99.62	7	99.67
3	99.7	8	99.68
4	99.76	9	99.7
5	99.73	10	99.74

We then use K=4 for comparison with the other algorithms. The experimental results of all algorithms are shown in Table 7.

TABLE VII
EXPERIMENTAL RESULTS

Algorithm	Accuracy	Precision	Recall	F-measure
DT	99.97	78.22	78.08	78.15
SVM	99.95	77.52	77.5	77.51
RBFN	99.94	76.25	76.15	76.2
Naïve-Bayes	99.93	77.24	77.29	77.26
K-NN, k = 4	99.76	73.7	73.86	73.78

The accuracy values of DT, SVM, RBFN, NB, and K-NN are 99.97%, 99.95%, 99.94%, 99.93%, and 99.76%, respectively. The results show that all algorithms give a high accuracy value (more than 99.00%); the highest accuracy value is obtained by using DT. The precision values of DT, SVM, RBFN, NB, and K-NN are 78.22%, 77.52%, 76.25%, 77.24% and 73.70%, respectively. The DT gives a higher value than the others. The recall values of DT, SVM, RBFN, NB, and K-NN are 78.08%, 77.50%, 76.15%, 77.29%, and 73.86%, respectively. The DT gives a higher value than the others. The F-measure values of DT, SVM, RBFN, NB, and K-NN are 78.15%, 77.51%, 76.20%, 77.26% and 73.78%, respectively. The DT gives a higher value than the others.

Although the DT gives a higher accuracy value than the others, all algorithms still give an accuracy higher than 99.00%. In summary, the results show that the DT algorithm provides higher accuracy, precision, recall, and F-Measure values than the others, leading to an optimal classification algorithm. These accurate results are due to this model using a bag of concepts together with TF-IDF instead of a bag of words and using a frequent concept tree obtained from the FP-Growth algorithm.

IV. Conclusion

This research proposes a model for classifying customer need phrases in the need to buy and sell properties from Facebook and web boards. This model presented a method for representing BOC in semantic analysis and applies TF-IDF for term weighting, representing TF-IDF value in BOC. As this study focuses on understanding Thai customer needs by classifying phrases posted on social media into a particular concept, creating the customer need ideas for semantic classification is the main task of this study. Therefore, BOC is a selected technique instead of BoW (Bag of Words). This is because BOC is a collection of words that are semantically related.

Furthermore, this research has presented a novel approach to creating BOC from patterns of words that are related in terms of semantics and order of word occurrence. As many classes of concepts obtained according to customer needs were encountered, the FP-Growth algorithm is also used in our proposed model. This affects the number of classes to be reduced from 713 to 13, which makes the classification more accurate and effective. The effectiveness of the proposed model is evaluated and compared between five classification algorithms: DT, SVM, NB, K-NN, and RBFN. The result has shown that accurate results from the DT algorithm were achieved and met the objective of this study. Thus, we can conclude that the proposed model is more applicable to knowing and understanding customer needs in Thai from online social media as we seek.

Although this model has performed well, it is suitable to be applied with the need of buying and selling property domain. We need to improve some points for future work, such as using word relatedness and word similarity, enhancing the effectiveness of synonyms for this model, extracting name entities in customer needs phrases, and expanding this model to other exciting domains. Understanding the syntax and semantics of phrases written in social media and automatically classifying customer needs terms are also necessarily required for future work. Furthermore, we have found some interesting points from our study: applying this model to other languages, such as English. This can be suggested by studying the nature and syntax of writing such language in social media.

References

[1] S.-H. Liao, R. Widowati, and Y.-C. Hsieh, "Investigating online social media users' behaviors for social commerce recommendations," *Technol. Soc.*, vol. 66, p. 101655, 2021, doi: 10.1016/j.techsoc.2021.101655.

[2] T. Hou, B. Yannou, Y. Leroy, and E. Poirson, "Mining customer product reviews for product development: a summarization process," *Expert Syst. Appl.*, vol. 132, pp. 141–150, 2019, doi: 10.1016/j.eswa.2019.04.069.

[3] B. Jeong, J. Yoon, and J.-M. Lee, "Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis," *Int. J. Inf. Manage.*, vol. 48, pp. 280–290, 2019, doi: 10.1016/j.ijinfomgt.2017.09.009.

[4] Y. K. Dwivedi *et al.*, "Setting the future of digital and social media marketing research: perspectives and research propositions," *Int. J. Inf. Manage.*, vol. 59, p. 102168, 2021, doi: 10.1016/j.ijinfomgt.2020.102168.

[5] F. Li, J. Larimo, and L. C. Leonidou, "Social media marketing strategy: definition, conceptualization, taxonomy, validation, and future agenda," *J. Acad. Mark. Sci.*, vol. 49, no. 1, pp. 51–70, 2021, doi: 10.1007/s11747-020-00733-3.

[6] S. Qiu, L. Wu, Y. Yang, and G. Zeng, "Offering the right incentive at the right time: leveraging customer mental accounting to promote prepaid service," *Ann. Tour. Res.*, vol. 93, p. 103367, 2022, doi: 10.1016/j.annals.2022.103367.

[7] O. S. Itani, A. Kalra, and J. Riley, "Complementary effects of CRM and social media on customer co-creation and sales performance in B2B firms: The role of salesperson self-determination needs," *Inf. Manag.*, vol. 59, no. 3, p. 103621, 2022, doi: 10.1016/j.im.2022.103621.

[8] S. Basu, "Personalized product recommendations and firm performance," *Electron. Commer. Res. Appl.*, vol. 48, p. 101074, 2021, doi: 10.1016/j.elerap.2021.101074.

[9] L. Lowphansirikul, C. Polpanumas, A. T. Rutherford, and S. Nutanong, "A large English–Thai parallel corpus from the web and machine-generated text," *Lang. Resour. Eval.*, 2021, doi: 10.1007/s10579-021-09536-6.

[10] J. Pan, M. Yan, E. M. Richter, H. Shu, and R. Kliegl, "The Beijing sentence corpus: a Chinese sentence corpus with eye movement data and predictability norms," *Behav. Res. Methods*, 2021, doi: 10.3758/s13428-021-01730-2.

[11] S. Li, Y. Wang, Z. Lan, X. Yuan, L. Zhang, and G. Yan, "Effects of word spacing on children's reading: evidence from eye movements," *Read. Writ.*, vol. 35, no. 4, pp. 1019–1033, 2022, doi: 10.1007/s11145-021-10215-9.

[12] K. Paripremkul and O. Sornil, "Segmenting words in Thai language using minimum text units and conditional random field," *J. Adv. Inf. Technol.*, vol. 12, no. 2, pp. 135–141, 2021.

[13] C. Saetia, E. Chuangsuwanich, T. Chalothorn, and P. Vateekul, "Semi-supervised Thai sentence segmentation using local and distant word representations," *arXiv Prepr. arXiv1908.01294*, 2019.

[14] National Electronics and Computer Technology Center, "Thai lexeme tokenizer:LexTo." .

[15] C. Haruechaiyasak and S. Kongyoung, "TLex: Thai lexeme analyser based on the conditional random fields," *Proc. Int. Symp. Nat. Lang. Process.*, 2009.

[16] National Electronics and Computer Technology Center, "TLex." .

[17] National Electronics and Computer Technology Center, "Thai lexeme tokenizer: lexitron dictionary." .

[18] M. Sahlgren and R. Cöster, "Using bag-of-concepts to improve the performance of support vector machines in text categorization," in *Proceedings of the 20th International Conference on Computational Linguistics*, 2004, p. 487, doi: 10.3115/1220355.1220425.

[19] P. Li, K. Mao, Y. Xu, Q. Li, and J. Zhang, "Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base," *Know.-Based Syst.*, vol. 193, no. C, Apr. 2020, doi: 10.1016/j.knosys.2019.105436.

[20] W. Shalaby and W. Zadrozny, "Learning concept embeddings for dataless classification via efficient bag-of-concepts densification," *Knowl. Inf. Syst.*, vol. 61, 2019, doi: 10.1007/s10115-018-1321-8.

[21] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.

[22] M. Mujahid *et al.*, "Sentiment analysis and topic modeling on Tweets about online education during covid-19," *Appl. Sci.*, vol. 11, no. 18, p. 8438, 2021, doi: http://dx.doi.org/10.3390/app11188438.

[23] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, "A comparative study of language transformers for video question answering," *Neurocomputing*, vol. 445, pp. 121–133, 2021, doi: https://doi.org/10.1016/j.neucom.2021.02.092.

[24] P. K. Jain, W. Quamer, V. Saravanan, and R. Pamula, "Employing BERT-DCNN with sentic knowledge base for social media sentiment analysis," *J. Ambient Intell. Humaniz. Comput.*, 2022, doi: 10.1007/s12652-022-03698-z.

[25] A. Mahmoud and M. Zrigui, "Semantic similarity analysis for corpus development and paraphrase detection in Arabic," *Int. Arab J. Inf. Technol.*, vol. 18, pp. 1–7, 2020, doi: 10.34028/iajit/18/1/1.

[26] A. Jalilifard, V. F. Carid'a, A. F. Mansano, and R. Cristo, "Semantic sensitive TF-IDF to determine word relevance in documents," *ArXiv*, vol. abs/2001.0, 2021.

[27] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," *SIGMOD Rec.*, vol. 29, no. 2, pp. 1–12, May 2000, doi: 10.1145/335191.335372.

[28] S. Barbon Junior *et al.*, "Sport action mining: dribbling recognition in soccer," *Multimed. Tools Appl.*, vol. 81, no. 3, pp. 4341–4364, 2022, doi: 10.1007/s11042-021-11784-1.

[29] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986, doi: 10.1007/BF00116251.

[30] F. M. Shamrat, R. Ranjan, K. Hasib, A. Yadav, and A. Siddique, "Performance evaluation among ID3, C4.5, and CART decision tree algorithm," 2022, pp. 127–142.

[31] O. Abualghanam, S. Al-Khatib, and M. Hiari, "Data mining model for predicting customer purchase behavior in e-commerce context," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, p. 421, 2022, doi: 10.14569/IJACSA.2022.0130249.

[32] B. F. Tanyu, A. Abbaspour, Y. Alimohammadlou, and G. Tecuci, "Landslide susceptibility analyses using random rorest, C4.5, and C5.0 with balanced and unbalanced datasets," *CATENA*, vol. 203, p. 105355, 2021, doi: 10.1016/j.catena.2021.105355.

[33] J. Santoso, N. Ginantra, M. Arifin, R. Riinawati, D. Sudrajat, and R. Rahim, "Comparison of classification data mining C4.5 and naïve bayes algorithms of EDM dataset," *TEM J.*, vol. 10, pp. 1738–1744, 2021, doi: 10.18421/TEM104-34.

[34] F. Es-sabery, K. Es-sabery, H. Garmani, and A. Hair, "Sentiment analysis of covid19 tweets using A mapReduce fuzzified hybrid classifier based on C4.5 decision tree and convolutional neural network," *E3S Web of Conferences*, vol. 297. EDP Sciences, Les Ulis, 2021, doi: 10.1051/e3sconf/202129701052.

[35] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[36] B. AlBadani, R. Shi, and J. Dong, "A novel machine learning approach for sentiment analysis on twitter incorporating the universal language model fine-tuning and SVM," *Appl. Syst. Innov.*, vol. 5, no. 1, 2022, doi: 10.3390/asi5010013.

[37] T. H. Jaya Hidayat, Y. Ruldeviyani, A. R. Aditama, G. R. Madya, A. W. Nugraha, and M. W. Adisaputra, "Sentiment analysis of twitter data related to Rinca Island development using Doc2Vec and SVM and logistic regression as classifier," *Procedia Comput. Sci.*, vol. 197, pp. 660–667, 2022, doi: 10.1016/j.procs.2021.12.187.

[38] B. Paul, S. Guchhait, T. Dey, D. Das Adhikary, and S. Bera, "A Comparative study on sentiment analysis influencing word embedding using SVM and KNN," in *Cyber Intelligence and Information Retrieval*, Springer, 2022, pp. 199–211.

[39] H. Kaur, S. U. Ahsaan, B. Alankar, and V. Chang, "A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets," *Inf. Syst. Front.*, vol. 23, no. 6, pp. 1417–1429, 2021, doi: 10.1007/s10796-021-10135-7.

[40] R. Vidhya and G. Vadivu, "Towards developing an ensemble based two-level student classification model (ESCM) using advanced learning patterns and analytics," *J. Ambient Intell. Humaniz. Comput.*, vol. 12, no. 7, pp. 7095–7105, Jul. 2021, doi: 10.1007/s12652-020-02375-3.

[41] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Mach. Learn.*, vol. 29, no. 2, pp. 131–163, 1997, doi: 10.1023/A:1007465528199.

[42] J. Ji, H. Wang, S. Song, and J. Pi, "Sentiment analysis of comments of wooden furniture based on naive Bayesian model," *Prog. Artif. Intell.*, vol. 10, no. 1, pp. 23–35, 2021, doi: 10.1007/s13748-020-00221-3.

[43] J. Gautam, M. Atrey, N. Malsa, A. Balyan, R. Shaw, and A. Ghosh, "Twitter data sentiment analysis using naive bayes classifier and generation of heat map for analyzing intensity geographically," 2021, pp. 129–139.

[44] V. V., J. B. Cooper, and R. L. J., "Algorithm Inspection for Chatbot Performance Evaluation," *Procedia Comput. Sci.*, vol. 171, pp. 2267–2274, 2020, doi: 10.1016/j.procs.2020.04.245.

[45] R. R. Sethuraman and J. S. K. Athisayam, "An Improved Feature Selection Based on Naive Bayes with Kernel Density Estimator for Opinion Mining," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 4059–4071, 2021, doi: 10.1007/s13369-021-05381-5.

[46] R. S. Kumar, A. F. Saviour Devaraj, M. Rajeswari, E. G. Julie, Y. H. Robinson, and V. Shanmuganathan, "Exploration of sentiment

analysis and legitimate artistry for opinion mining," *Multimed. Tools Appl.*, vol. 81, no. 9, pp. 11989–12004, 2022, doi: 10.1007/s11042-020-10480-w.

[47] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967, doi: 10.1109/TIT.1967.1053964.

[48] M. Gayathri and R. J. Kannan, "Ontology based concept extraction and classification of ayurvedic documents," *Procedia Comput. Sci.*, vol. 172, pp. 511–516, 2020, doi: 10.1016/j.procs.2020.05.061.

[49] T. Anwar and V. Uma, "Comparative study of recommender system approaches and movie recommendation using collaborative filtering," *Int. J. Syst. Assur. Eng. Manag.*, vol. 12, no. 3, pp. 426–436, 2021, doi: 10.1007/s13198-021-01087-x.

[50] Z.-R. He, Y.-T. Lin, C.-Y. Wu, Y.-J. You, and S.-J. Lee, "Pattern classification based on RBF Networks with self-constructing clustering and hybrid learning," *Appl. Sci.*, vol. 10, no. 17, 2020, doi: 10.3390/app10175886.

[51] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara, "Language resource management system for asian wordnet collaboration and its web service application.," 2010.