


A Survey of Automatic Text Classification Based on Thai Social Media Data

Tanatorn Tanantong, Thammasat Research Unit in Data Innovation and Artificial Intelligence, Department of Computer Science, Faculty of Science and Technology, Thammasat University, Thailand*

 <https://orcid.org/0000-0002-8927-2365>

Monchai Parnkow, Thammasat Research Unit in Data Innovation and Artificial Intelligence, Department of Computer Science, Faculty of Science and Technology, Thammasat University, Thailand

ABSTRACT

In the digital age, the information on social media, such as Facebook, Twitter, and Instagram, is increasing rapidly. Therefore, it has led to studies and researches on social media analytics to extract useful models or knowledge from the data. One of the most interesting topics in social media analytics is text classification on social media data. However, since social media data has a diverse and complex data structure, text analysis and classification are considered a challenging issue that requires a specific technique to implement. The objective of this review paper is to collect and review research related to the automatic classification of Thai text on social media by presenting and explaining the process of text classification on various issues. These include data collection and data sources, amount of data and data preparation for research, feature extraction methods, text classification automated modeling methods, efficacy evaluation and measurement methods, the results of text classification, and summary of the overall trend of research on the topic.

KEYWORDS

Lexicon-Based Approaches, Machine Learning, Social Media Text, Thai Text Classification

INTRODUCTION

Social media networks, the main means of communication between people across the world, can meet the needs of its users in many ways (Wichitboonyarak, 2011; Al-Ibrahim & Alzamil 2019; Saggu & Sinha 2020; Verma et al., 2021). For example, people can express opinions, exchange information, or communicate via text, audio, and video (Chandran & Madhu 2021; Tanantong et al., 2021). In addition, users can access social networks like Facebook, Twitter, and Instagram through tablets, smartphones and notebooks (“Social Networks: An Introduction,” 2009). People can use social media to share their feelings or opinions through posts, comments, and short messages (Aggarwal & Zhai, 2012; Champihom, 2018). According to DataReportal (2020), 75% of the Thai population average three hours of social media use per day per person. Thailand has the eighth largest number of Facebook users in the world, with approximately 41 million users (Positioning, 2020a). On Twitter, Thai users ranked 15th in the world, with approximately 5.7 million accounts (Positioning, 2020b). Popular social media platforms in Thailand include Facebook, Instagram, Pantip, YouTube, and expanding news sites. The number of users and amount of information on social networks are increasing at a

rapid pace (Ikonomakis et al., n.d.; Tampakas, 2005). Social media analytics can identify patterns or extract useful knowledge from the significant amount of data on these platforms (AR_Group, 2020)

Text classification is an important and widely used social media analytics technique to understand and automatically classify data. This method flags messages according to the relevance of content. Groups or categories of messages are predefined. Text classification techniques can be applied to messages on social media platforms and divided into a text classification by topic group and sentiment analysis group (Thongied & Netisopakul, 2017; MonkeyLearn, 2020b). Classification by topic divides text according to topics defined by the researcher, such as the classification of type of news (Viriyavisuthisakul et al., 2015; Jotikabukkana et al., 2016). This, in turn, can categorize the news into types like economics, entertainment, foreign news, etc. Type of research can categorize messages into legal topics like defamation and nondefamation (Rao & Spasojevic, 2016; Arreerard & Senivongse, 2018; Yuenyong et al., 2018). Research that classifies messages into types of goods and services can identify groups who will or will not buy or use a product (Chumwatana, 2015; Thetmueang & Chirawichitchai, 2017; Apichai et al., 2018). The sentiment analysis classification focuses on the feelings or intention of the author. This can be divided into two types. The first classification targets the author's feelings toward a product, service, person, place, or event. These feelings may be positive feelings, negative feelings, or neutral sentiments (Pugsee & Niyomvanich, 2015; Songram, 2016; Kuhmanee et al., 2017; Kunpattanasopon, 2018; Tanantong et al., 2020). The second classification targets the author's intended emotion. Examples include happiness, sadness, loneliness, shock, and fear (Sarakit et al., 2015; Hemtanon & Kittiphattanabawon, 2019; Panawas, 2019).

Text classification on social media can be categorized by topic, feeling, or author intention. After defining the purpose of the message type, the prepared text is used to develop a model for automatic text classification. This can be divided into two main development approaches. First, a text classification model is built using a lexical-based approach. The development of the model classifies messages based on rules created by the researcher. For example, the searching and extracting of keywords in text can be categorized into messages based on mood or sentiment. The message will receive a negative classification if it contains the words "bad" or "nasty." It will be classified as positive if the message contains the words "delicious" or "good." The lexical classification method is used on research that categorizes text by topic, emotion, or feeling (Pugsee & Niyomvanich, 2015; Wansopha et al., 2015; Songram, 2016; Tapang, 2016; Masdisornchote, 2016). The machine learning (ML) approach uses model development from the sample dataset where the answer is assigned. The prepared data is then extracted and features can be used by various algorithms in the development of automated models, including: naïve baye (NB); decision tree (DT); nearest neighbor (k-NN); logistic regression analysis (LR); support vector machine (SVM); and deep learning (DL). Research studies have developed a text classification model based on ML methods using NB, DT, k-NN, LR, and SVM (Sanguansat, 2016; Sodanil, 2016; Songpan, 2017; Thetmueang & Chirawichitchai, 2017; Kuhmanee et al., 2017; Apichai et al., 2018; Arreerard & Senivongse, 2018; Thanasopon et al., 2019; Inplang & Thongkam, 2020) and research using DL (Rao & Spasojevic, 2016; Akkaradamrongrat et al., 2019).

This article presents a review of the literature and research in social media automatic message classification published between 2015 and 2020. Several methods were used to classify messages on social media through the research and development of models. The current study can be used as a guideline for future study and development.

Next, the article will discuss the collection, data generation, preparation, and extraction of key features. Then, the article will explore text classification techniques and efficacy evaluation before presenting an automated method for analyzing and classifying text on social media. Finally, the article will provide a summary of the literature review.

DATA COLLECTION AND PREPARATION

Text classification from social media is the process of searching and extracting knowledge from messages on social media platforms, including Facebook and Pantip. The extracted data is used for various reasons, including the classification or categorization of the following (Bhoyar, 2012; Phawattanakul, 2012; Pinmuang & Thongkam, 2017; Laowsungsuk et al., 2017):

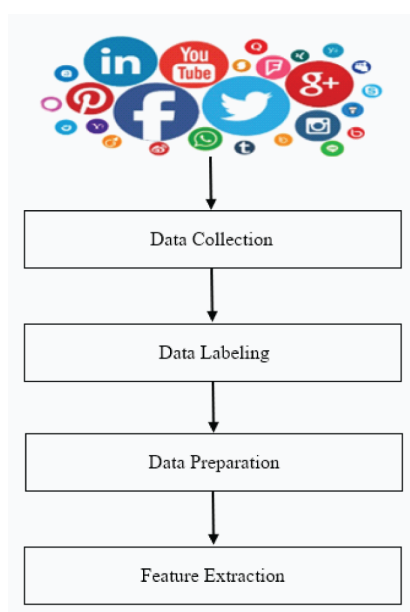
- General e-mail vs. junk or spam.
- Restaurant customer reviews by product and satisfaction (positive, neutral, or negative).
- Opinions on online payment services (PromPantip Pay) using Facebook data.

There are three forms of text classification (MonkeyLearn, 2020b):

1. **Topic Classification:** This is the identification and classification of messages into defined topics like education, technology, and daily life (Sornlertlamvanich et al., 2015).
2. **Sentiment Analysis and Classification:** This is the analysis and classification of a message via message content per the author's intent, whether positive or negative (Tanantong et al., 2020).
3. **Language Classification:** This is the classification of text according to the language in which the original text was written (Abramov & Mehler, 2011).

There are challenges in the analysis and classification of messages on social networks. For example, in Thai, sentence structure and grammar are inconsistent and complex (Boonkwan, 2016). Moreover, online communication is usually in a short form and, therefore, may not consider grammatical correctness (Lhasiw et al., 2021). Thus, Thai text should be prepared and optimized before entering the extraction process (Songmuang et al., 2021). As shown in Figure 1, the main steps to classification requires the following steps: (1) data collection; (2) data labeling; (3) data preparation; (4) feature extraction (Jotikabukkana et al., 2016).

Figure 1. Data collection and preparation



Data Collection

The method of collection and managing data is complex because information on social media is diverse in terms of data type and source (Senasang, 2020). The literature review of research related to the analysis and classification of Thai text on social media from 2015 to 2020 found that messages were collected and extracted from various sources (Senate, 2018), including social networking sites (Facebook, Twitter, and Instagram), video-sharing sites (YouTube and Twice), online boards and forums (Pantip and Pramool), and blogs (Wongnai and Agoda).

Amount of Data

Many amounts of data were used to research methods to classify Thai text on social media. These could be clustered according to the amount of available data.

1. Research employing data in experiments ranging from 100 to 999 messages was led by Chaoprasit and Lekcharoen (2017). Over 515 blog posts were used to classify text by topic. Songram (2016) classified feelings using data from 467 Facebook messages.
2. Research employing data in experiments ranging from 1,000 to 9,999 messages included:
 - a. Jenkarn and Ketcham (2020) used Facebook, YouTube, Twitter, and Instagram to classify 1,100 messages based on topic.
 - b. Viriyavisuthisakul et al. (2015) adopted data from 1,638 Pantip messages.
 - c. Akkaradamrongrat et al. (2019) classified messages based on emotion through data from 2,928 Facebook messages.
 - d. Sarakit et al. (2015) applied data from 5,848 YouTube posts.
 - e. Pugsee and Niyomvanich (2015) used data from 7,220 blog posts.
 - f. In addition, Thai text classification studies applied the thousand-digit data included Chaoprasit and Lekcharoen (2017), Apichai et al. (2018), Arreerard and Senivongse (2018).
3. Research employing data in experiments ranging from 10,000 to 99,999 messages included:
 - a. Kuhamanee et al. (2017) used data from 10,000 Twitter messages.
 - b. Sanguansat (2016) applied data from Pantip 55,539 messages.
 - c. Sornlertlamvanich et al. (2015) classified feelings from 17,069 Twitter messages.
 - d. Additional research studies, included Jotikabukkana et al. (2016), Phan and Tay (2017), and Khamphromma et al. (2019), employed the ten-thousand-digit data.
4. Research employing data in experiments ranging from 100,000-999,999 messages on social media included:
 - a. Trakultaweekoon and Klaithin (2016) used data from 148,227 Twitter and Pantip messages.
 - b. Eamwiwat et al. (2019) used data from 323,196 Twitter and Facebook messages.
 - c. Deerojanadej et al. (2016), Kongyoung et al. (2019), Kunpattanasopon (2018), and Rao and Spasojevic (2016) used the hundred-thousand-digit data.

Regarding the research on Thai text classification on social media, over 60% of the research used data in experiments that ranged from 1,000 to 9,999 messages.

Methods of Data Collection

Many methods are used to collect Thai text from social media. Data collection by programming can be divided into two methods. The first works in conjunction with the service provider's website application programming interface (API), acting as a middleman between users and the API creator (Rodphothong et al., 2018). This method can only be used on social media platforms that have developed API to connect and collect data. Examples include Twitter, Facebook, and YouTube (Sornlertlamvanich et al., 2015; Sarakit et al., 2015; Deerojanadej et al., 2016; Jotikabukkana et al., 2016; Klaithin & Haruechaiyasak, 2016; Songram et al., 2016; Katchapakirin et al., 2018; Sirihattasak, 2019).

The second programming method is manual programming. One example, Python, retrieves information about the stock of Siam Commercial Bank (SCB), Krung Thai Bank (KTB), and Kasikorn Bank (KBANK) to access product reviews (Thetmueang & Chirawichitchai, 2017; Sangsavate et al., 2019). Data collection via software packages allows users to collect information from social media. For example, WebHarvy can retrieve game comments and RapidMiner can retrieve Twitter news messages (Jotikabukkana et al., 2016; Inplang & Thongkam, 2020).

Data Labeling

Data labeling is required to create a model for automatically classifying texts into topics emotions, or feelings (Services, 2021). Defining the answer is an important step to set the message answer.

Determining Data Labeling

Decisions must be made by experts in each field. Tanantong et al. (2020) used 4,608 responses to identify positive, negative, and neutral messages. Tapang (2016) divided their 4,027 responses into three groups: (1) ironic messages; (2) unironic messages; and (3) inconclusive messages. Five experts determined the answer by awarding one point per message for irony, zero points for unironic, and one-half point to messages that were inconclusive. Other studies would determine statement outcomes through expert decisions (Apichai et al., 2018; Arreerard & Senivongse, 2018; Charoensuk & Sornil, 2018).

Answers Based on Keywords or Special Symbols

Keywords or special symbols can be used to determine messages, as evidenced in Jotikabukkana et al. (2016), Kunpattanasopon (2018), Songram et al. (2016), Vateekul and Koomsubha (2016), Sirihattasak (2019), and Jitrlada and Chingchai (2019). When compared to the determination of the expert, this method is more convenient when processing large amounts of data. This was beneficial when Kongyoung et al. (2019) processed 132,938 messages. The method designated keywords like “jaizaa.com,” “news.jarm.com,” and “liekr.com” as “clickbait.” Other keywords, including “Morning News,” “TV3,” “Thai PBS,” and “Voice TV,” were listed as “non-clickbait.” Vateekul and Koomsubha (2016) used 22,000 messages to categorize messages by considering emoticons within the text. If the emoticon was found to express a positive feeling, it was deemed a positive message. On the contrary, if the emoticon expressed a negative feeling, it was a negative statement.

Data Preparation

Data preparation refers to the process that changes raw data into an appropriate format before analysis. Data preparation can differ based on the type of data and purpose of analysis (TechTarget, 2021). For the messages from social media (e.g., Facebook, Twitter, and Pantip), the data preparation process was used to access suitable text for analysis and processing (Phawattanakul, 2012).

Word Segmentation

Word segmentation is performed by separating the meaning of each word, which is the basic process for text analysis divided into subwords (Runnaphongsa & Urathamkun, 2006). As shown in Figure 2, these words are then used to extract characteristics. There are three techniques for word segmentation:

1. **Rule-Based Approach:** This technique follows rules established with Thai grammar principles. It considers consonants, vowels, orthography, and new paragraphs (Mahatthanachai, 2017). The technique saves time and processing space. A lexicon or dictionary are not required. However, it is less accurate and cannot handle words beyond the rules.
2. **Dictionary-Based Approach:** This technique compares the words stored in the dictionary to the document. This type of truncation will have differing results. It will depend on the term stored

in the referenced dictionary. This technique is accurate because more words can be added to the dictionary. However, it requires more space to store the words.

3. **Corpus-Based Approach:** This technique applies statistical principles and probabilities to compare terms with a repository of solutions (Haruechaiyasak et al., 2008). The surrounding context should be considered to increase the efficiency of cutting, for example “sit | expose | wind” and “do | eye | round | wink” (Plookpedia, 2019). However, it may take more time and space to process. The accuracy depends on the number, quality, and variety of text in the solution corpus.

Other techniques were applied along with these three to develop a tool (see Table 1).

Table 1. Tools for word segmentation

Tools	Methods	Details
Lexto	Dictionary-based Approach	The Lexto cutting technique (Sangkheetrakarn, n.d.) selects the longest word from a dictionary. This method is processed quickly. Lexto was developed by Thailand’s National Electronics and Computer Technology Center (Nectec). Research includes Desai and Mehta (2016), Sirihattasak (2019), and Thanasopon et al. (2019).
LextoPlus	Dictionary-based Approach	The LextoPlus program was developed from Lexto to support normalized words in social media (Center, 2016). For example, “very goodddddd” converts to “very good” (and can be added to the dictionary). Research includes Kongyoung et al. (2019) and Saensuk et al. (2019).
Kucut	Corpus-based Approach	Kucut is a Python-language library for cutting Thai text via corpus and machine learning. Research includes Charoensuk and Sornil (2018), Sanguansat (2016), Vateekul and Koomsubha (2016), and Viriyavisuthisakul et al. (2015).
PYThaiNLP	Dictionary-based Approach	PYThaiNLP is a Python library package for text manipulation and analysis. There are various functionalities like word break, Thai word cut, spell check, and typo correction. Research includes Sangsavate et al. (2019).
Deepcut	Corpus-based Approach	Deepcut is a Python package library for word wrapping (“word segmentation by deep learning [AI]”). It is based on surrounding words or word context. It uses machine learning by Nectec and S-Sens. Research includes Yuenyong et al. (2018).

Figure 2. Word segmentation and elimination of stopwords in Charoensuk and Sornil (2018)

Original messages (in Thai): “ไม่เห็นได้ผลเลยใช้มาจะ 6 เดือนแล้ว”
Word segmentation results (in Thai): ไม่ เห็น ได้ ผล เลย ใช้ มา จะ 6 เดือน แล้ว
Original messages: “I don’t see any results at all. I have used for _6_ bottles already”
Word segmentation results: don’t see any results at all used have for _ 6 _ bottles already

Stopword Removal

Stopwords are words found frequently in each document. They are insignificant to the classification of messages (Choonui, 2012). The elimination of a stopword is the removal of a stopword from the text (see Figure 2). It is an important step to reduce word size and size of the generation attribute. This will result in effective next steps. Stopwords can be divided into the following types:

1. A special symbol type of stopword (special character) refers to marked characters like “+,” “ü,” “-,” “\$,” and “%.” It also includes characters that do not exist on the keyboard, such as “α,” “β,” and “π.” Research on special symbols includes Tapang (2016), Masdisornchote (2016), Charoensuk and Sornil (2018), and Sirihattasak (2019).
2. A numeric stopword refers to symbols that represent real numbers like “12,” “53,” and “VIII.” Research includes Songram et al. (2016), Vateekul and Koomsubha (2016), and Chumwatana and Wongkolkitsilp (2019).
3. A prepositional stopword is the word used before nouns, pronouns, verbs, or adverbs. It tells the status or expresses the relationship between words, pronouns, or sentences like “with,” “in,” “of,” “with,” and “by.” Examples of research examining prepositional stops include Songram et al. (2016) and Thetmueang and Chirawichitchai (2017).
4. A pronoun stopword refers to a word used to refer to a noun, such as “she,” “he,” and “it” (Songram et al., 2016; Thetmueang & Chirawichitchai, 2017).
5. A modifying stopword is a word used to modify a noun, pronoun, verb, or adverb. This word provides clarity and detail, such as “huge,” “white,” “hot,” and “cool.” Research includes Pinmuang and Thongkam (2017) and Thetmueang and Chirawichitchai (2017).

There are more stopwords that eliminate URLs attached to messages (Desai & Mehta, 2016; Jitrlada & Chingchai, 2019). Vateekul and Koomsubh (2016) specified a single character in a message as a stopword. Desai and Mehta (2016) determined the words beginning with the # hashtags as stopwords.

Feature Extraction

Computers cannot directly analyze and process natural language text or documents; therefore, it is necessary to convert text or documents into a feature format. In doing so, a mathematical model or ML model can be developed to analyze and process text in an automated way. This process, called feature extraction, will extract key features that represent or help classify text or documents. Such features are available in a variety of formats like word, phrase, term frequency weighting (TF), and term frequency-inverse document frequency (TF-IDF). Each feature is suitable for different processing (Songram, 2016; Songram et al., 2016).

Word

The word attribute feature uses words from the text that have been cut through the process. Word feature extraction is performed in a simple, less time-consuming process. It can be applied for work on text analysis and classification, such as social media opinion classification, by extracting word-of-text attributes (Tanantong et al., 2020). It is then compared with a positive word dictionary and negative word dictionary. The statement is considered positive if the text matches more words in the positive dictionary than the negative dictionary. However, the use of word attribute does not consider word placement, word efficiency, or quality in constructing a discriminant model. The verbal feature has also been used in research samples (Chumwatana, 2015; Masdisornchote, 2016; Laosungsk et al., 2017; Saensuk et al., 2019).

Boolean Weighting

Boolean weighting values represent extracted words in each passage with the truth value relative to the extracted words in the entire passage. Each word in the text is represented by 1 or 0 (the word appears or does not appear in the text, respectively). Boolean weighting is an easily extractable feature. However, this feature does not consider word placement. Panawas (2019) used Boolean weighting to classify social media messages based on the mood of the message.

Term Frequency Weighting (TF)

This weighted value accounts for the frequency of words in the text. This feature is like Boolean weighting; however, in Boolean weighting, word frequency is used instead of word weight. TF does not consider word placement or word performance in constructing text classification. Examples include Viriyavisuthisakul et al. (2015), Desai and Mehta (2016), Pinmuang and Thongkam (2017) and Hemtanon and Kittiphattanabawon (2019). The method for calculating TF is as follows:

$$tf(word, text) = \frac{f(word, text)}{\sum f(word, text)}$$

$word$ = the word

$text$ = the message on social media

$f(word, text)$ = the frequency of interested words in the text

$\sum f(word, text)$ = the total number of words in the text

Term Frequency–Inverse Document Frequency (TF-IDF)

TF-IDF is a statistical value used to assess the importance of words in a document or text. The importance of a word is directly proportional to the number of times the word appears in the document or text. It is reduced in importance by the frequency of that word in the entire document or text (Thanasopon et al., 2019). TF is used along with IDF. TF-IDF is used in Sornlertlamvanich et al. (2015), Tapang (2016), Jotikabukkana et al. (2016), Savigny and Purwarianti (2017), Kunpattanasopon (2018), and Thanasopon et al. (2019). The importance of the words to the document or text is evaluated to reduce the problem of using words that have no influence on the text classification model. However, the position of words is not considered by the IDF formula:

$$idf(word, all_{text}) = \log \frac{N}{df(word)}$$

N = the total number of messages

$df(word)$ = the number of messages that contain words of interest

Word Embedding

Word embedding is a feature for processing natural language. Proposed by Mikolov et al. (2013), this is a technique to turn words into vectors using ML algorithms. The word embedding feature considers the meaning and location of a word in calculating the attribute value. Word embedding can be implemented in many ways, such as Word2Vec, Global Vectors (GloVe), and FastText. Word2Vec can create word embedding via Skip Gram and Common Bag of Words (CBOW).

GloVe is a way to convert words into vectors by calculating the similarity of the meaning of the two words. This algorithm finds words that are related to each other, putting them in a group of words that are similar in relation to Word2Vec. GloVe also looks at the relationship of the words used in conversational contexts. When used in different contexts, it can give different meanings.

FastText is an evolution of Skip Gram and Word2Vec. FastText was released in 2016 by Facebook. It is effective; however, it uses more memory and time to analyze and classify the text as compared with other features. Examples include Claypo and Jaiyen (2015), Akkaradamrongrat et al. (2019), Eamwivat et al. (2019), Piyaphakdeesakun (2019), and Sirihattasak (2019).

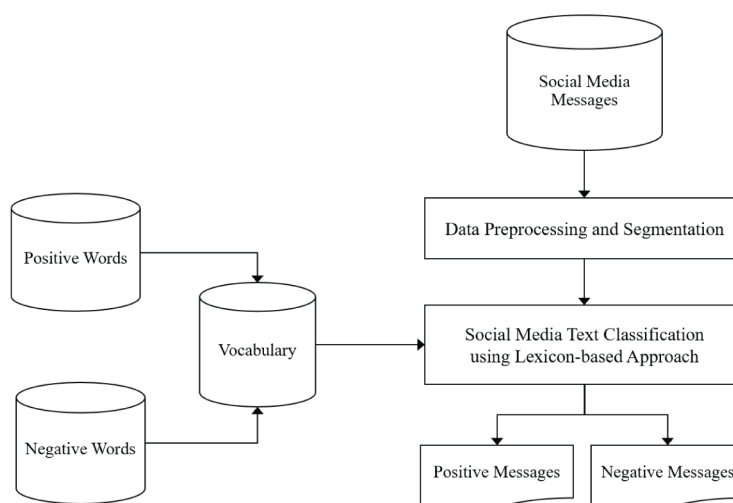
TECHNIQUES FOR CLASSIFYING TEXT AND EVALUATING PERFORMANCE

Text classification technique is a procedure used to indicate the group that the message belongs. For example, messages are classified as the expression of opinions to the Thai government, whether positive or negative. Text classification methods are divided into two methods (Suwanpipob, 2019), lexicon-based approach and machine learning approach.

Lexicon-Based Approach (LB Approach)

The LB classification method establishes rules or requirements for classifying text. The algorithm searches for polarity words like “bad,” “nasty,” “tasty,” or “good.” Expressions in the text are calculated by algorithm and scored from -1 to +1. After combining the scores of all the words, it uses the mathematical and statistical methods to determine the final scores to identify the sentiment of the text (Taboada et al., 2011), as shown in Figure 3. The scores and classifications of texts depend on the vocabulary in the dictionary for the analysis and classification of English sentiments (Suwanpipob, 2019).

Figure 3. Text classification framework by a LB approach



Jaihuek and Mungsing (2018) used an automatic scoring program to create a dictionary for the keywords of answers. The individual received points if the answer that matched the keyword was found

(see Table 2). Saensuk et al. (2019) developed a method for characterization and classifying opinions on smartphone features into positive and negative statements. Different types of words (“pretty” or “bad”) can be positive and negative, respectively. Nouns (N) or noun phrases (NP) (“camera,” “battery,” and “color”) are represented by various numbers (see Table 3). For example, machine was represented by 13, camera was represented by 1, and screen was represented by 14. Then, the words in the comment text were compared with the comments to create rules for classifying the message based on the comment terminal. If the comment terminals in a message were positive (+) or negative (-), they were all evaluated as positive or negative, respectively. However, if a message contained both polarized words, both positive (+) and negative (-) comments were evaluated as negative.

Table 2. Automated score assessment for subjective tests

Question	Keywords	Answer	Scores
What is the center processing unit?	Computer brain and CPU	CPU and taking good cares	66
Provide an example of four importing devices.	Mouse, keyboard, scanner, and barcode reader	Mouse, keyboard, and microphone	60

Table 3. Identification of opinions using a rule-based system

Word Segmentation	Rule-based System	Opinion
machine degrade fast	13 - +	13 -
machine good good	13 + +	13 +
camera bad bad	1 - -	1 -
camera not nice	1 ! +	1 -
screen resolution little	14 + -	14 -
machine camera good	13 1 +	13 + 1 +

Machine Learning Approach (ML Approach)

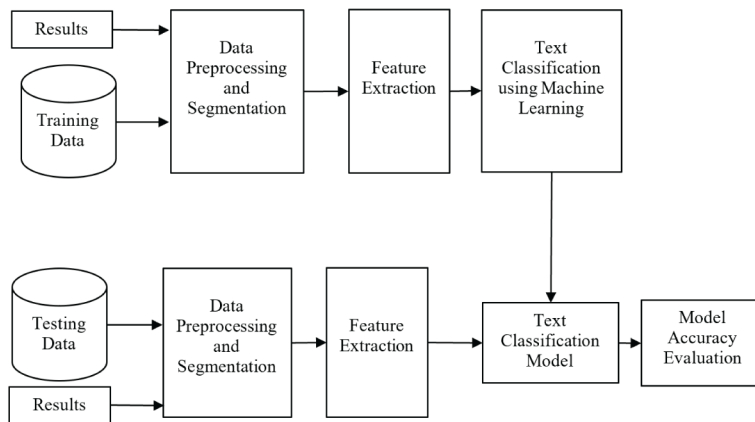
Data analysis and classification uses a method of learning and extracting knowledge from data sets to create automated models. ML is divided into supervised learning, unsupervised learning, and reinforcement learning (Shalev-Shwartz & Ben-David, 2009). The supervised learning method has been widely applied to develop models for automatically classifying messages on social media. First, it collects and generates data types of solutions. Second, it prepares and divides the data into training and test datasets. Third, it extracts data features. Fourth, it creates a model for classifying text. Finally, fifth, it evaluates the performance of the model by using the test data. Details are shown in Figure 4 (Vateekul & Koomsubha, 2016). There are techniques that are commonly used by data classification methods for the ML approach.

Naïve Bayes (NB)

The NB algorithm is a learning method to create classifiers based on probability theory and on Bayes’ theory (Thangaraj & Sivakami, 2018). It aims to find the probability of the emergence of an event by guessing from events that already took place. NB, a ML algorithm, can be processed quickly. NB uses mathematical principles to classify the data; therefore, the data to be analyzed and classified

must be in numerical form or vector of numbers. Research includes Klaithin and Haruechaiyasak (2016), Chaoprasit and Lekcharoen (2017), Arreerard and Senivongse (2018), and Chumwatana and Wongkolkitsilp (2019). Studies on the classification of message sentiments includes Kuhamanee et al. (2017), Hemtanon and Kittiphattanabawon (2019), Panawas (2019), Piyaphakdeesakun et al. (2019), Sangsavate et al. (2019), and Thanasopon et al. (2019).

Figure 4. ML approach



Decision Tree (DT)

DT is an algorithm that results in a tree-like graph structure. The node holds various properties that define and decide the direction of the data (Jittapu, 2007). The highest node is the root node. The branch is the value of the attributes in the extracted node. The number of branches is equal to the number of possible values in each node. The leaf is a group of results in data classification. DT is an algorithm that can be easily understood and interpreted by humans (Pinmuang & Thongkam, 2017). The DT graph reveals which properties of the data determine the classification, as well as the value of the individual properties. It is useful because it can analyze data with more accurate classification of information. Research on DT for data classification includes text classification research by Sodanil (2016), Chaoprasit and Lekcharoen (2017), and Songpan (2017). Research on sentiment classification includes Sodanil (2016), Kuhmanee et al. (2017), Kunpattanasopon (2018) and Panawas (2019).

K-Nearest Neighbors (k-NN)

k-NN is a distance-based algorithm used to classify data. This algorithm considers the distance between the characteristics of the prepared dataset and the new dataset (Yuenyong et al., 2018). It also predicts the solution of the new dataset by class closest to the sample data (Apichai et al., 2018). The solution is selected based on the data type with the largest number of members out of k members, where k is the number of the data. For example, 1-NN, 3-NN, and 5-NN are considered for the 1, 3, and 5 members of the sample data set, respectively. Research on text classification according to a given topic includes Chaoprasit and Lekcharoen (2017). Research on text sentiment classification includes Inplang and Thongkam (2020).

Logistic Regression (LR)

LR analysis is based on the concept of regression analysis, which is the study of the relationship of variables to estimate the value of the dependent variable by considering the value and the relationship of the independent variable. Subsequently, the regression analysis technique was developed to be LR, which is a qualitative statistical technique that uses statistical analysis (Trakultaweekoon & Klaithin, 2016). The probability of an event occurrence is forecasted from a set of independent variables.

The two types of LR, binary LR and multinomial LR (Sucheewa, 1996), differ in terms of dependent variables. LR applied to the dependent variables are divided into two subgroups (0 and 1). The multigroup LR applied to the dependent variables are more than two subgroups. The research using LR to classify messages includes research by Sanguansat (2016) and Thanasopon et al. (2019). These are methods presented for dividing opinions of messages (positive, negative, neutral) as the opinions on banking and tourism businesses in Thailand, respectively.

Support Vector Machine (SVM)

SVM is a classification method presented by Cortes and Vapnik (1995). The standard process starts from analyzing data and creating a straight line to divide the data into groups as specified. Then, the straight line is considered the optimal line for segmenting data (Sanguansat, 2016). The SVM algorithm has the advantage of being able to adjust to a range of parameters, resulting in a highly efficient model compared to other algorithms. However, the model is both difficult to understand and to explain the answers or results (Thangaraj & Sivakami, 2018). Examples of work that applies SVM include Sodanil (2016), Katchapakirin et al. (2018) and Kongyoung et al. (2019). Research on the classification of message sentiments includes Thetmueang and Chirawichitchai (2017), Charoensuk and Sornil (2018), and Kunpattanasopon (2018).

Deep Learning (DL)

DL gives computers the ability to compute and process data that mimics the work of the human brain. The neural network (NN) algorithm consists of three layers (Han & Zhao, 2016). The input layer receives data (the source variable). The hidden layer, in the middle, analyzes and processes data based on mathematical principles. The output layer makes decisions and predicts the outcome of the data (dependent variable). DL was developed from the NN algorithm by connecting several hidden layers of the NN (Pattansarn & Pattansarn, 2020). DL is derived from using more than two layers of NN in modeling. Therefore, it can be concluded that the more hidden layers are connected, the deeper the learning structure is created.

For tasks in automatic text classification, the most popular algorithm is the long-short-term memory (LSTM) algorithm. When new events go into memory, the brain will choose whether to receive it due to its importance. When the brain chooses to accept new events to the memory system, it is necessary to forget some events in the past. The basic structure of LSTM is a forget gate to simulate a “forget” event and a memory gate to simulate a “remember” event. LSTM research focuses on categorizing messages based on emotion from messages on social media like sentiment analysis of products and services and sentiment-based text analysis (Rao & Spasojevic, 2016; Vateekul & Koomsubha, 2016; Akkaradamrongrat et al., 2019).

Performance Evaluation of Text Classification Model

Evaluation of the model’s effectiveness in classifying text on social media can be determined and calculated from the values in the fusion matrix table (Kowsari et al., 2019), as shown in Table 4. The values in the table can calculate accuracy, precision, recall, and F-Measure.

Table 4. Performance evaluation of text classification model

Confusion Matrix		Result	
		Class A text	Class B text
Predicted result	Class A text	True Positive	False Positive
	Class B text	False Negative	True Negative

Accuracy (ACC)

The ACC value indicates the ability to classify how likely a new text is predicted to be correct (Thetmueang & Chirawichitchai, 2017). Accuracy can be calculated from the following equation:

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

when:

- True Positive (TP) is the number of texts correctly predicted in Class A.
- True Negative (TN) is the number of texts correctly predicted in Class B.
- False Positive (FP) is the number of texts incorrectly predicted in Class B.
- False Negative (FN) is the number of texts incorrectly predicted in Class B.

Precision (PRE)

PRE is ability of the text classification model that indicates the correct prediction accuracy of the model. High-precision models return the same prediction results in different messages with similar forms and features. On the contrary low-precision model will give different prediction results although the text has similar forms or characteristics (Varasri et al., 2014). The precision can be calculated from the following:

$$Precision = \frac{TP}{TP + FP}$$

Recall (REC)

REC is a measure of the effectiveness of the prediction (Inplang & Thongkam, 2020). Recall values are calculated using the ratio of the correct predicted outcomes (TP) to the total number of predictions (TP + FN). The recall value can be obtained from the following equation:

$$Recall = \frac{TP}{TP + FN}$$

F-Measure

F-Measure calculates the ratio between precision and recall. The F-Measure value indicates the predictive performance and efficiency of the model (Lisirikul, 2018; Suwanpipob, 2019). The F-Measure value can be obtained from the following equation:

$$F - Measure = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right)$$

RESEARCH PRESENTING AN AUTOMATED METHOD FOR ANALYZING AND CLASSIFYING THAI TEXT ON SOCIAL MEDIA

The analysis and classification of Thai text on social media can be divided into two categories according to research objectives. The first, topic classification focuses on the research for developing methods to identify and classify messages into different topics. For example, messages can be divided into reliable and unreliable (Songram et al., 2016), vulgar and nonvulgar (Chaoprasit & Lekcharoen, 2017), defamatory and nondefamatory of the government (Arreerard & Senivongse, 2018), and cyber-bullying and non-cyber-bullying messages (Jenkarn & Ketcham, 2020). The second, sentiment analysis and classification, presents a method for analyzing and classifying messages by considering the content of the message, as well as how the author feels or intends to communicate. Examples include classifying a message as positive, negative, or neutral (Pugsee & Niyomvanich, 2015; Kuhamanee et al., 2017; Inplang & Thongkam, 2020). The emotional classification of the text can be divided into angry, confused, fearful, happy, sad, and surprised (Sarakit et al., 2015).

Topic Classification

Topic classification is the categorization of texts according to the topic specified by the researcher (MonkeyLearn, 2020b). This research can be divided into two groups according to the method to classify the text, namely LB and ML. Topic classification using LB has been used in many research studies. Wansopha et al. (2015) designed and developed a system to analyze suicide risk via Facebook texts. The texts were compared to words in the vocab storage, classifying whether it is a risky vs. risk-free text. Then, it measured user satisfaction with the developed system. The results of the assessment of satisfaction and overall efficiency in using the system was 3.56 points (score range 0-5). Tapang (2016) collected a total of 4,027 Twitter texts related to Internet services. The texts were divided into 108 malicious texts and 3,913 normal texts. However, there were six texts that the researcher could not categorize and, therefore, did not process. The LexTo tool was used for the word wrap. The study developed methods for distinguishing sneaky texts from regular messages by calculating text probabilities based on the N-gram model.

From the review of research published between 2015 and 2020, it is found that the number of Thai texts on social media classified using LB is still smaller than ML methods. The data reported in MonkeyLearn (2020a) indicated that ML classification often produces higher efficiency results than LB. Therefore, ML classification tends to be more popular. Viriyavisuthisakul (2015) developed a model for classifying news articles on social media. The study collected 1,638 news articles from social media via TF and TF-IDF features. k-NN was used to model and compare the performance of 10 similarity computational methods. The results showed that k-NN text classification using Bray Curtis, cosine, and correlation similarity computation showed correlation results. Accuracy was highest at 56.47%, 54.66%, and 53.23%, respectively.

Klaithin and Haruechaiyasak (2016) collected 24,779 opinion texts about traffic from Twitter. The texts were prepared and improved by removing duplicate tweets and deleting URLs contained in

texts. After the initialization process, 6,131 texts remained. Then, the texts were categorized and made expert solutions. The texts were classified into texts on road accidents, texts that break traffic news, texts for help, and texts that relay feelings about traffic. Then, the LexTo tool was used to wrap the words. TF-IDF attributes were extracted via NB for the classification of the investigator selection text. It was found that the data classification model had an accuracy of 76.4%. Rao and Spasojevic (2016) presented a model for text classification. It compared the authenticity of the model in 31 languages, collected 336,000 Twitter texts, and divided the statements into those that caused companies to sue the writer and those that did not cause companies to sue the writer. It used word embedding as a feature and LSTM to classify the texts. The study found that the model had an average accuracy of 87%. Songram (2016) presented the development of an application for detecting Thai texts that lead to deception on social media Facebook. The researchers collected 2,378 texts by programming in PHP language with the Facebook graph API tool. Then, the data were divided into 1,189 unreliable texts and 1,189 reliable texts. The texts were extracted for Boolean weighting, TF, and TF-IDF attributes. SVM and k-NN were used as part of the model development. It was found that experiments via the vector algorithm combined with the Boolean weighting feature yielded the highest accuracy (98.82%).

Chaoprasit and Lekcharoen (2017) developed a model to detect Thai profanity on social media by collecting text from blog websites. The researchers randomly selected 515 texts from a total of 1,214, categorized as 199 vulgar texts and 316 nonvulgar text. The researchers used TF-IDF feature to develop models, and then used NB, DT, and k-NN to classify messages. The accuracy is 96%, 96%, and 95%, respectively. Thetmueang and Chirawichitchai (2017) analyzed 2,890 texts about online product reviews (1,573 texts from Agoda blogs and 1,317 texts from Twitter). The researcher experimented with RapidMiner for modeling. The researchers extracted TF-IDF attributes and compared the model's performance using NB, DT, SVM, and k-NN in text classification. The accuracy was 83.31%, 79.92%, 83.38%, and 71.26%, respectively.

Yuenyong (2018) collected 96,737 texts using the Twitter API. The texts were divided into six groups: (1) obscene texts (1,360 texts); (2) texts referring to sex toys (915 texts); (3) texts related to prostitution (2,261 texts); (4) texts related to drugs (546 texts); (5) texts referring to gambling (1,655 texts); and (6) common texts (90,000 texts). Deepcut was used for data preparation and text extraction. TF-IDF was used for feature extraction. The model was created using the LSVM algorithm for classifying the text. The average accuracy was 97.23%. Apichai et al. (2018) collected texts about the incident of the Tham Luang Khun Nam Nang Non cave accident. Text was pulled from 2,100 Twitter texts (written in Thai and Japanese). The texts were clustered into 158 remedial texts, 560 news texts, 1,078 commentary texts, and 304 general texts. The researchers then selected word embedding features and compared the efficiency of NB and SVM in text classification. The overall efficiency (F-Measure) was 64%. Arreerard and Senivongse (2018) used 1,034 Facebook-related comments about Thai government. Then, a lawyer was asked to determine the answer. The texts were divided into 446 defamatory statements and 588 nondefamatory statements. The researcher chose the word feature to compare the efficiency between NB and the vector contributor. The accuracy was 68% and 74%, respectively.

Kongyoung et al. (2019) developed a system to detect social media ads by collecting data from 132,938 Twitter texts. Propaganda included 60,393 texts with the words "jaizaa.com," "bkkchanel.info," and "liekr.com." There were 72,545 texts with the words "ThaiPBS," "Voice_TV," and "Thairath_News," which were considered nonpropaganda. Then, word embedding was done to create classifying models from NB, SVM, and CNN algorithms. The researchers reported the F-Measure results of the algorithms as 88.94%, 91.45%, and 95.25%, respectively. Chumwatana (2019) collected messages from 1,000 social media texts related to real estate business. The texts were divided into two groups. The first included texts that show customers' intention to buy products. The second included texts that show when customers have no intention to buy the product. By extracting TF-IDF attributes and using NB and SVM to create the text discriminant model, it was found that the accuracy of the SVM was 78.1%. NB was 63.4%.

Jenkarn and Ketcham (2020) collected 1,100 comments that expressed opinions to people like celebrities, athletes, or famous criminals. The texts were divided into 538 cyberbullying texts and 562 noncyberbullying texts. The researchers then extracted TF-IDF attributes and classified the texts using NB, k-NN, and SVM to develop cyberbullying prevention on social media. The experimental results found that SVM provided an 83.91% accuracy, which is higher than NB and k-NN rates of 74.64% and 72.82%, respectively. In addition, research related to the classification of social media messages by topic using ML techniques included Sornlertlamvanich et al. (2015), Jotikabukkana et al. (2016) and Piyaphakdeesakun et al. (2019).

The researchers studied data from Thai text from social media sources like Twitter, Facebook, Pantip and YouTube. Most of the data used for processing is in the thousand digits (1,000 to 9,999 messages). After collecting the texts, the topic classification varies depending on research objectives like classification of legal related texts, inappropriate texts detection, and texts related to purchasing decisions. The report on accuracy is between 53.23% and 98.82%, as detailed in Table 5.

Table 5. Research methods for classifying social media texts by topics

Researcher(s)	Data Source	Number of Data (Texts)	Classification Topic	Feature	Algorithm	Measurement (Highest Results: %)
Viriyavisuthisakul et al. (2015)	Pantip	1,638	News	TF, TF-IDF	KNN	ACC = 56.47
Wansopha et al. (2015)	FB	NA	Risk of Suicide	Word	LB	NA
Rao and Spasojevic (2016)	TW, FB	336,000	Defamation	Word Embedding	LSVM	ACC = 87.00
Songram et al. (2016)	FB	2,378	Scams	Boolean,TF, TF-IDF	SVM, KNN, NB	ACC = 98.82 (KNN)
Tapang (2016)	TW	4,027	Ironic Texts	TF-IDF, Word	LB	NA
Klaithin and Haruechaiyasak (2016)	TW	6,131	Opinions on Traffic	TF-IDF	NB	ACC = 76.40
Chaoprasit and Lekcharoen (2017)	Blog	515	Rude Text	TF-IDF	NB, DT, KNN	ACC = 96 (NB, DT)
Thetmueang and Chirawichitchai (2017)	TW, Blog	2,890	Opinions on Online Product Reviews	TF-IDF	SVM, DT, KNN, NB	ACC = 83.38 (SVM)
Yuenyong et al. (2018)	TW	96,737	Illegal Text	TF-IDF	LSVM	ACC = 97.23
Apichai et al. (2018)	TW	2,100	Texts about “Tham Luang” Cave Incident	Word Embedding	NB, SVM	F1 = 64.00 (SVM)
Arreerard and Senivongse (2018)	FB	1,034	Defamation	Word, TF	NB, SVM	PRE = 72.00 (SVM), ACC = 74.00 (SVM), F1 = 64.00 (SVM)
Chumwatana and Wongkolkitsilp (2019)	TW, FB,Pantip	1,000	Possibility to Buy Products	TF-IDF	NB, SVM	ACC = 78.10 (SVM)

Table 5 continued on next page

Table 5 continued

Researcher(s)	Data Source	Number of Data (Texts)	Classification Topic	Feature	Algorithm	Measurement (Highest Results: %)
Kongyoung et al. (2019)	TW	132,938	Clickbait	Word Embedding	CNN, NB, SVM	PRE = 95.15, (CNN) REC = 95.34, (CNN) F1 = 95.245 (CNN)
Jenkin and Ketcham (2020)	YT, TW, FB, IG	1,100	Cyberbullying Messages	TF-IDF	SVM, KNN, NB	PRE = 84.23, (SVM) REC = 83.80, (SVM) ACC = 83.91, (SVM) F1 = 84.01 (SVM)

Sentiment Analysis and Classification

Sentiment classification research is the classification of messages according to the feelings or intentions of the writer. Its purpose is to explore and develop businesses and services. The research can be divided into two categories. First, the text classification uses feelings toward a product, service, person, place, or event (positive, negative, and neutral). The second is the classification of messages based on the emotion or feeling that the writers want to communicate and express (happy, sad, lonely, shocked, and fear).

Messages can be classified based one's feelings about a product, service, person, place, or event. Pugsee and Niyomvanich (2015) developed a system to determine the choice of recipes using comments. A total of 7,222 texts from blogging sites about food were analyzed and categorized based on users' feelings about a recipe. It consisted of 6,620 positive texts, 54 negative texts, and 548 neutral texts. The researchers developed an automatic text classification based on the SentiWordNet dictionary comparison, with an accuracy of 93.08%.

Masdisornchote (2016) collected 1,090 texts expressing opinions about mobile phones via the Siamphone Website. The texts were segmented by the LexTo word tokenization and compared with the sentiment dictionary to classify into positive and negative statements. An overall efficiency (F-Measure) was 86.1%. Songram (2016) used a Graph API tool to collect 467 texts from Facebook during the Coup d'état in Thailand. Three experts classified the texts, which can be divided into 259 positive texts and 208 negative texts. A classification based association (CBA) was used to create rules to classify the texts. The result showed an accuracy of 77.75%. Vateekul and Koomsubha (2016) collected 3,813,173 texts from Twitter. They brought them into the word segmentation process via Kucut to consider the emoticons. A positive emoticon was identified as a positive emotion and feeling. A negative emoticon indicated a negative emotion and feeling. A random sample of 22,000 texts were used to create a model for classification by extracting the embedded feature using the word2vec. This study used CNN, NB, and SVM. The accuracy was 75.35%, 74.05%, and 74.71%, respectively. Sanguansa (2016) used 55,539 texts to express customer opinions in the field of business (positive, negative, and neutral texts). The researchers extracted TF, TF-IDF, and word-embedding attributes. They used NB, LR, and SVM to create a text classification model. It was found that the embedding attribute classification using the LR model yielded the highest accuracy at 85.12%.

Kuhamanee et al. (2017) collected 10,000 opinion texts about traveling in Thailand. Twitter data was analyzed to formulate guidelines for promoting the tourism industry of Bangkok. The researchers used the obtained data to divide the text into positive and negative feelings, as well as a sense of neutrality. TF-IDF attributes were then extracted. The performance of the text classification models by NB, DT, SVM, and artificial NN was compared. It found that the accuracy of each was 55.06%, 79.83%, 80.11%, and 80.33%, respectively.

Inplang and Thongkam (2020) studied the comments on games. PlayerUnknown's Battlegrounds (PUBG) on Google Play used WebHarvy to retrieve 3,798 texts. The texts were prepared by deleting non-Thai text, cutting words, and removing stop words. There were 3,123 texts left after going through the text preparation process (1,875 positive texts, 299 negative texts, and 949 neutral texts). The number of texts per type was balanced with a new synthetic method, the oversampling technique called SMOTE. This method added a small group of classes to the same level as the other. To model the text classification in this research, RapidMiner was used with various algorithms, including DT, NB, k-NN, and SVM. It was found that the k-NN algorithm had the highest accuracy at 99.87%. Tanantong et al. (2020) collected 4,608 texts from social media platforms (Facebook, Instagram, and Pantip) to classify opinions on the mobile network operator. The texts were divided into negative and positive. The research team used LB approaches to create a dictionary of terms on topics related to mobile network service. They calculated text scores, finding that the accuracy was the highest at 87.7%.

The research for categorizing text by sentiment varies in terms of data and issues. The purpose of text analysis often identifies and classifies opinions on social media (positive or negative) toward a product, service, person, place, or event. Another interesting way to analyze and categorize text is a classification based on the direct emotions or feelings of the writers. Sarakit et al. (2015) classified 5,848 comments on YouTube videos by categorizing according to the mood of the messages (anger, confusion, fear, happiness, sadness, and surprise). LexTo was used as word segmentation. TF and TF-IDF were used to extract attributes. NB, DT, and SVM were used to create a model for text classification. It was found that the TF-based NB algorithm had the highest accuracy (84.48%).

Panawas (2019) collected 1,800 texts to present an analysis of the feelings of Thai people on social media. The texts were divided according to emotion (anger, fear, love, sadness, and shock). The performance of text classification was compared according to attributes like TF, TF-IDF, and Boolean weighting. The classification efficiency was evaluated according to NB, k-NN, and Ensemble. It found that the Ensemble algorithm using Boolean weighting as a feature achieved the highest accuracy (76.04%). Hemtanon and Kittiphattanabawon (2019) used text recognition techniques to screen patients with depression from 1,500 Facebook posts. The posts were divided into those from people at risk of depression (negative messages) and those from people who are not at risk (positive messages). It provided a ratio of 50% for positive and negative posts. The texts were wrapped with LexTo and extracted with TF attributes. NB and SVM were used to create a text recognition model. F-Measure values were 93% and 94%, respectively. In addition, research on sentiment analysis and classification was performed by Chumwatana (2015), Claypo and Jaiyen (2015), Deerosejanadej et al. (2016), and Haruechaiyasak et al. (2018). Trakultaweekoon and Klaithin (2016) classified texts by sentiment toward a product, service, person, place, or event. Charoensuk and Sornil (2018) classified texts based on the emotion or feeling that the writers wanted to communicate.

From the literature review on sentiment analysis and classification, the researcher found that the study collected Thai text from social media platforms like Twitter, Facebook, Pantip, and YouTube. Text can be classified as 1. First, the classification was based on how someone feels about a product, service, or event (positive, negative, and neutral). Second, the classification is based on the emotion or feeling the writers want to communicate or express (i.e., happiness, sadness, loneliness, shock, and fear). The researchers collected data in the areas of medical, business, and services to explore customers' opinions about a product, service, or event. The researcher aimed to develop and improve the product, service, or event. The results show accuracy rates between 77.75% and 99.87%, as detailed in Table 6.

Table 6. Research methods for classifying social media messages by emotion and feeling

Researcher(s)	Data Source	Number of Data (Texts)	Classification Topic	Feature	Algorithm	Measurement (Highest Results: %)
Sarakit et al. (2015)	You Tube	5,848	Opinions on YouTube Videos	TF-IDF	NB, DT, SVM	ACC = 84.48 (NB)
Pugsee and Niyomvanich (2015)	TW	7,222	Opinions on Foods	Word	LB	ACC = 93.08
Masdisornchote (2016)	TW, FB, Pantip	1,090	Opinions on Mobile Phones	Word	LB	PRE = 86.11, REC = 86.23, F1 = 86.10,
Vateekul and Koomsubha (2016)	TW	22,000	Opinions on Twitter	Word Embedding	LSVM, NB, SVM, DCNN	ACC = 75.35 (DCNN)
Songram (2016)	FB	467	Opinions on Politics	Word	LB	ACC = 77.75
Sanguansat (2016)	Pantip	55,539	Opinions on Retail Sales, Banks, and Communication Systems	TF, TF-IDF, Word Embedding	LR, NB, SVM	ACC = 85.12 (LR)
Kuhamanee et al. (2017)	TW	10,000	Opinions on Tourist Attractions in Thailand	TF-IDF	NB, DT, SVM, ANN	ACC = 80.33 (ANN)
Panawas (2019)	TW, FB, Pantip	1,800	Opinions on Online Social Feeds	TFIDF, TF, Boolean	NB, KNN, DT, Ensemble	ACC = 76.04 (Ensemble)
Hemtanon and Kittiphattanabawon (2019)	FB	1,500	Opinions on Depression	TF	NB, SVM	PRE = 96 (SVM), REC = 93 (NB), F1 = 94 (SVM)
Inplang and Thongkam (2020)	Blog	3,123	Opinions on Games, PlayerUnknown's Battlegrounds (PUBG)	TF	SVM, NB, DT, KNN	PRE = 99.75 (KNN), REC = 100 (KNN), ACC = 99.87 (KNN)
Tanantong et al. (2020)	TW, FB, Pantip	4,608	Opinions on Social Media	Word	GLC, SLC	ACC = 87.7 (SLC)

CONCLUSION

This research presents a literature review on the classification of Thai text on social media. It serves as a research and application guideline for those interested in researching and developing methods for automatic message classification on social media. The classification of Thai text from social media can be divided into sentiment analysis and classification and topic classification. There are several steps to develop an automatic text classification method. In the first step, text is collected and prepared (cutting and eliminating stopwords). The researcher collects messages from various social networking services like Twitter, Facebook, Instagram, Pantip, and YouTube. Data can be collected

manually by computer using instant program. The research found that most of the researchers use text in the thousand-digit data.

The process to prepare text occurs after the data collection. It wraps the text using a specific technique or tool to break the text into words. Then, the words are used to eliminate the word stops (words that are unnecessary for creating a text classification model). Feature extraction is a continuation of the data preparation following the collection process in which key features are extracted from the text in preparation for modeling. Examples of features used in text classification research include word, Boolean weighting, term frequency weighting, TF-IDF, and word embedding. Next, the researcher applies the features that have been selected and prepared to be applied in model development. It can be divided into the LB approach and modeling with ML approach. The final step for conducting research on Thai text classification on social media is the evaluation of the model's performance with the measurement tools and datasets provided for testing. Many studies reported performance results with various measuring tools, with accuracy ranging between 53.23% and 98.82%, precision between 72.00% and 99.75%, recall between 86.23% and 100.00%, and F-measure between 64.00% and 95.24%.

The literature reviews found the use of diverse methods on the development of Thai text classification models on social media. These can be applied as a guideline for the analysis and classification of text, articles, or other forms of Thai documents (both online and offline). However, due to the Thai language's complex structure and various interpretations, social media networks have become a main source of communication. The emergence of slang terms and new words conveys meaning and understanding within a specific group. The development of a Thai text classification model faces challenges both in terms of model performance and the implementation of developed models in real-world environments. For example, a significant amount of labeled social media data is required for training and evaluating the classification model. However, the annotation of social media data is time-consuming and expensive. Therefore, further studies should examine the use of semi-supervised learning and transfer learning for improving the model performance of text classification on social media.

ACKNOWLEDGMENT

This work was supported by Thammasat Research Unit in Data Innovation and Artificial Intelligence.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

REFERENCES

- Abramov, O., & Mehler, A. (2011). Automatic language classification by means of syntactic dependency networks. *Journal of Quantitative Linguistics*, 18(4), 291–336. doi:10.1080/09296174.2011.608602
- Aggarwal, C., & Zhai, C. (2012). *Mining text data*. Publisher.
- Akkaradamrongrat, S., Kachamas, P., & Sinthupinyo, S. (2019). Text generation for imbalanced text classification. In *Proceedings of the International Joint Conference on Computer Science and Software Engineering (IJCSS)* (pp. 181-186). Academic Press.
- Al-Ibrahim, F., & Alzamil, Z. A. (2019). Big data contextual analytics study on Arabic tweets summarization. *International Journal of Knowledge and Systems Science*, 10(4), 18–34. doi:10.4018/IJKSS.2019100102
- Apichai, C., Chan, K., & Suzuki, Y. (2018). Classification of Thai tweets: Mining treasures from Tweet heap. In *Proceedings of the International Conference on Systems and Informatics (ICSAI)* (pp. 311-315). doi:10.1109/ICSAI.2018.8599398
- AR Group. (2020). *What is social media analytics and why is it important?* <https://www.ar.co.th/kp/th/575>
- Arreerard, R., & Senivongse, T. (2018). Thai defamatory text classification on social media. In *Proceedings of the International Conference on Big Data, Cloud Computing, Data Science & Engineering* (pp. 73-78). Academic Press.
- Bhojar, K. (2012). Email mining: A review. *IJCSI International Journal of Computer Science Issues*, 9(1).
- Boonkwan, P. (2016). *Text mining and Thai NLP*. National Electronics and Computer Technology Center. <https://www.slideshare.net/datascienceth/text-mining-and-thai-nlp>
- Champhom, S. (2018). *Basic knowledge about social media*. <https://sites.google.com/site/suwinaycapihxm/bi-ngan-srup-khwam-ru-beuxng-tn-keiyw-kab-sangkkm-xxnlin>
- Chandran, J., & Madhu, V. V. (2021). A novel triangle count-based influence maximization method on social networks. *International Journal of Knowledge and Systems Science*, 12(4), 92–108.
- Chaoprasit, N., & Lekcharoen, S. (2017). Development of Thai language profanity investigation model for online media using data mining technique. In *Proceedings of the Rangsit Graduate Research Conference (RGRC)* (pp. 1432-1441). Academic Press.
- Charoensuk, J., & Sornil, O. (2018). A hierarchical emotion classification technique for Thai reviews. *Journal of Information and Communication Technology ITB*, 12, 280–296.
- Choonui, W. (2012). *Feature reduction using FCA for Web page classification* [Master's thesis]. Prince of Songkla University, Thailand.
- Chumwatana, T. (2015). Using sentiment analysis technique for analyzing thai customer satisfaction from social media. In *Proceedings of the International Conference on Computing and Informatics* (pp. 659-664). Academic Press.
- Chumwatana, T., & Wongkolkitsilp, K. (2019). Using classification technique for customer relationship management based on Thai social media data. In *Proceedings of the International Conference on Computer and Automation Engineering* (pp. 7-11). Academic Press.
- Claypo, N., & Jaiyen, S. (2015). Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection. In *Proceedings of International Computer Science and Engineering Conference (ICSEC)* (pp. 394-397). Academic Press.
- Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. Kluwer Academic Publishers.
- Deerosejanadej, C., Promon, S., & Achalakul, T. (2016). A framework for mining Thai public opinions. In *Big data* (pp. 355-339). Academic Press.
- Desai, M., & Mehta, M. (2016). Techniques for sentiment analysis of Twitter data: A comprehensive survey. In *Proceedings of International Conference on Computing, Communication and Automation (ICCCA2016)* (pp. 149-154). Academic Press.

Eamwivat, C., Thanasutives, P., Saetia, C., & Chalothorn, T. (2019). Using label noise filtering and ensemble method for sentiment analysis on Thai social data. In *Proceedings of the International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). Academic Press.

Han, Y., & Zhao, X.-M. (2016). A hybrid sequential feature selection approach for the diagnosis of Alzheimer's Disease. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (pp. 1216-1220). Academic Press.

Haruechaiyasak, C., Kongthon, A., Palingoon, P., & Trakultaweekoon, K. (2018). S-Sense: A sentiment analysis framework for social media monitoring applications. *Information Technology Journal*, 14(1), 11–22.

Haruechaiyasak, C., Kongyoung, S., & Damrongrat, C. (2008). LearnLexTo: A machine-learning based word segmentation for indexing Thai texts. In *Proceedings of the 2nd ACM workshop on Improving Non-English Web Searching* (pp. 85-84). Academic Press.

Hemtanon, S., & Kittiphattanabawon, N. (2019). An automatic screening for major depressive disorder from social media in Thailand. *Proceedings of the International Academic Conference of Suan Sunandha Rajabhat University*.

Ikonomakis, E., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. *WSEAS Transactions on Computers*, 4(8), 966–974.

Inplang, W., & Thongkam, J. (2020). Opinion analysis on PlayerUnknown's Battlegrounds (PUBG) mobile games using text mining. *Journal of Science & Technology MSU*, 39(5), 523–531.

Jaihuek, S., & Mungsing, S. (2018). Development of automatic subjective score test with word matching. *Kasalongkham Research Journal Chiangrai Rajabhat University*, 12(2), 81–93.

Jenkarn, N., & Ketcham, M. (2020). Thai-textual cyberbullying detection using support vector machines. *Science Technology and Innovation*, 1, 26–38.

Jitrlada, R., & Chingchai, E. (2019). Analysing Thai social media content to improve customer satisfaction. In *Proceedings of the International Conference on Electricity Distribution* (pp. 1-5). Academic Press.

Jotikabukkana, P., Sornlertlamvanich, V., & Manabu, O. (2016). Social media text classification by enhancing well-formed text trained model. *Institut Teknologi Bandung Journal*, 10, 177–196.

Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P., & Kaewpitakkun, Y. (2018). Facebook social media for depression detection in the Thai community. In *Proceedings of the International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1-6). Academic Press.

Kempi, S. (2020, February 18). *Digital 2020: Thailand*. <https://datareportal.com/reports/digital-2020-thailand>

Khamphomma, W., Soin, H., & Imthongkham, P. (2019). A hybrid sentiment analysis model for Thailand hotel review using K-means and K-NN. In *Proceedings of the National Conference on Informatics, NCIs, Walailak University* (pp. 1-11). Academic Press.

Klaithin, S., & Haruechaiyasak, C. (2016). Traffic information extraction and classification from Thai Twitter. In *Proceedings of International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1-6). Academic Press.

Kongyoung, S., Rugchatjaroen, A., & Kaothanthong, N. (2019). Automatic feature extraction and classification model for detecting Thai clickbait headlines sing convolutional neural network. *Information Modelling and Knowledge Bases*, 312, 184–194.

Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Multidisciplinary Digital Publishing Institute*, 10, 1–68.

Kuhamanee, T., Talmongkol, N., & Chaisuriyakul, K. (2017). Sentiment analysis of foreign tourists to Bangkok using data mining through online social network. In *Proceedings of the International Conference on Industrial Informatics (INDIN)* (pp. 1068-1073). Academic Press.

Kunpattanasopon, N. (2018). *Thai news article sentiment classification based on user comments on online social media* [Master's thesis]. Prince of Songkla University, Thailand.

- Laowsungsuk, P., Jinda, A., & Sitthisarn, S. (2017). Sentiment analysis of restaurant reviews on review Web sites. *Thaksin University Journal*, 20(1).
- LexToPlus. (2016, Oct 19). <https://www.nectec.or.th/innovation/innovation-software/lextoplus.html>
- Lhasiw, N., Sanglerdsinlapachai, N., & Tanantong, T. (2021). A bidirectional LSTM model for classifying Chatbot messages. In *Proceedings of the 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)* (pp. 1-6). Academic Press.
- Lisirikul, C. (2018). *Opinion analysis system to business by text mining on Twitter (Special Problem)*. King Mongkut's University of Technology North Bangkok.
- Mahatthanachai, C. (2017). *PTTSF word parsing techniques* [Master's thesis]. Mahidol University, Thailand.
- Masdisornchote, M. (2016). A sentiment analysis framework in implicit opinions for Thai language. In *Proceedings of Annual Conference of the IEEE Industrial Electronics Society* (pp. 357-361). Academic Press.
- MonkeyLearn. (2020a). *Text classification with machine learning & NLP*. <https://monkeylearn.com/text-classification/>
- MonkeyLearn. (2020b). *What is text classification?* <https://monkeylearn.com/what-is-text-classification/>
- Panawas, N. (2019). Thai sentiment analysis on social media using majority voting-based ensemble method. *Sripatum Chonburi Journal*, 15(1), 51–67.
- Pattansarn, N., & Pattansarn, N. (2020). Image processing for classifying the quality of the Chok-Anan Mango by simulating the human vision using deep learning. *Journal of Information Science and Technology*, 10, 24–29.
- Phan, M. C., & Tay, Y. (2017). Lazada product title quality challenge: A bag of features for short text classification. *Proceedings of the The Conference on Information and Knowledge Management (CIKM)*.
- Phawattanakul, K. (2012). *Suggestion mining from reviewers' reviews of television programs* [Master's thesis]. National Institute of Development Administration.
- Pinmuang, N., & Thongkam, J. (2017). *Classifying Thai opinions on online media using text mining* [Master's thesis]. Faculty of Informatics, Mahasarakham University, Thailand.
- Piyaphakdeesakun, C., Facundes, N., & Polvichai, J. (2019). Thai comments sentiment analysis on social networks with deep learning approach. In *Proceedings of the International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)* (pp. 1-4). Academic Press.
- Plookpedia. (2019). *Thai word segmentation*. <https://www.trueplookpanya.com/blog/content/61491>
- Positioning. (2020a). *Thai people are addicted to "social media." Facebook totals over 41 million, ranked 8th in the world*. <https://positioningmag.com/1092090>
- Positioning. (2020b). *Thai people use "Twitter" more than "Facebook"*. <https://positioningmag.com/1159478>
- Pugsee, P., & Niyomvanich, M. (2015). Sentiment analysis of food recipe comments. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*, 9, 182-193.
- PyThaiNLP User Guide. (n.d.). <https://pythainlp.readthedocs.io/en/latest/pythainlp-1-4-thai/#pythainlp-14>
- Rao, A., & Spasojevic, N. (2016). *Actionable and political text classification using word embeddings and LSTM*. arXiv 2016, arXiv:1607.02501.
- Rodphothong, W., Manerat, P., Peechapol, C., & Srimuang, C. (2018). The development of the job appointment management system. *Proceedings of the SPUC National and International Conference*.
- Runnaphongsa, K., & Urathamkun, P. (2006). *Improved rule-based and new dictionary for Thai word segmentation* [Master's thesis]. Khon Kaen University, Thailand.
- Saensuk, M., Songram, P., & Chompoowiset, P. (2019). Feature-based polarity specification of SmartPhone on Facebook. *Journal of Technology Management Rajabhat Maha Sarakham University*, 6(2).
- Saggu, A. K., & Sinha, A. (2020). Social influence analysis for information diffusion in complex commercial network. *International Journal of Knowledge and Systems Science*, 11(1), 22–59.

Sangkheettrakarn, C. (n.d.). *Thai word segmentation tool*. shorturl.at/uyJX4

Sangsavate, S., Tanthanongsakkun, S., & Sinthupinyo, S. (2019). Stock market sentiment classification from FinTech News. In *Proceedings of the International Conference on ICT and Knowledge Engineering (ICT&KE)* (pp. 1-4). Academic Press.

Sanguansat, P. (2016). Paragraph2Vec-based sentiment analysis on social media for business in Thailand. In *Proceedings of the International Conference on Knowledge and Smart Technology (KST)* (pp. 175-178). Academic Press.

Sarakit, P., Theeramunkong, T., Haruechaiyasak, C., & Okumura, M. (2015). Classifying emotion in Thai YouTube comments. In *Proceedings of International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)* (pp. 1-5). Academic Press.

Savigny, J., & Purwarianti, A. (2017). Emotion classification on Youtube comments using word embedding. In *Proceedings of the International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA)* (pp. 1-5). Academic Press.

Senasang, T. (2020). Opinions of adolescents on the use of social media: A case study of faculty of social sciences students. *Kasetsart University Political Science Review Journal*, 7(1), 19–33.

Senate, T. S. o. T. (2018). *Strategies for promoting through social media using Facebook*. The Secretariat of the Senate of Thailand.

Services, A. W. (2021). *What is data labeling for machine learning?* <https://aws.amazon.com/th/sagemaker/groundtruth/what-is-data-labeling/>

Sirihattasak, S. (2019). *Annotation and classification of toxicity for Thai Twitter* [Master's thesis]. Tokyo Metropolitan University.

Social Networks: An Introduction. (2009). Development and learning in organizations. *An International Journal*, 23(4).

Sodanil, M. (2016). Multi-language sentiment analysis for hotel reviews. In *Proceedings of the International Conference on Mechanics and Industrial Engineering* (pp. 1-4). Academic Press.

Songmuang, P., Sirisup, C., & Suebsriwichai, A. (2021). Missing link prediction using non-overlapped features and multiple sources of social networks. *Information*, 12(5), 214.

Songpan, W. (2017). The analysis and prediction of customer review rating using opinion mining. In *Proceedings of the International Conference on Software Engineering Research, Management and Applications* (pp. 71-77). Academic Press.

Songram, P. (2016). Opinion classification of politics on social network using associative classification. *Science and Technology RMUTT Journal*, 6, 83–93.

Songram, P., Choopool, A., & Thipsanthia, P. (2016). Analysis of deception leads by text on social network (Research plan). Maha Sarakham University, Thailand.

Sornlertlamvanich, V., Pacharawongsakda, E., & Charoenporn, T. (2015). Understanding social movement by tracking the keyword in social media. *Proceedings of the Multiple Approaches Lexicon (MAPLEX)*.

Sucheewa, S. (1996). Logistic regression analysis: Concept, analysis and interpretation. *Journal of Research Methodology*, 8(1).

Suwanpipob, W., Arch-int, N., & Wattana, M. (2019). A sentiment classification from review corpus using linked open data and sentiment lexicon. In *Proceedings of International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 19-23). Academic Press.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307.

Tanantong, T., & Ramjan, S. (2021). An association rule mining approach to discover demand and supply patterns based on Thai social media data. *International Journal of Knowledge and Systems Science*, 12(2), 1–16.

- Tanantong, T., Sanglerdsinlapachai, N., & Donkhampai, U. (2020). Sentiment classification on Thai social media using a domain-specific trained lexicon. In *Proceedings of the International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 580-583). Academic Press.
- Tapang, K. (2016). *Sarcasm classification in Twitter using probability of tweets* [Master's thesis]. Chulalongkorn University, Thailand.
- TechTarget. (2021). *What is data preparation? An in-depth guide to data prep*. <https://www.techtarget.com/searchbusinessanalytics/definition/data-preparation>
- Thanasopon, B., Buranapanitkij, J., & Netisopakul, P. (2019). Mining social media crowd trends from Thai text posts and comments. *Malaysian Journal of Computer Science*, 96-108.
- Thangaraj, M., & Sivakami, M. (2018). Text classification techniques: A literature review. *Journal of Information, Knowledge, and Management*, 13, 117-135.
- Thetmueang, R., & Chirawichitchai, N. (2017). Thai sentiment analysis of product review online using support vector machine. *Engineering Journal of Siam University*, 18(1).
- Thongied, K., & Netisopakul, P. (2017). Concept of creation and utilization of the Sentiment Corpus. *KMITL Journal of Information Technology*, 5.
- Trakultaweekoon, K., & Klaithin, S. (2016). SenseTag A tagging tool for constructing Thai sentiment lexicon. In *Proceedings of the International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1-4). Academic Press.
- Vateekul, P., & Koomsubha, T. (2016). A study of sentiment analysis using deep learning techniques on Thai Twitter data. In *Proceedings of the International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp. 1-6). Academic Press.
- Verma, J. V., Tanwar, S., Garg, S., & Rathod, A. D. (2021). Crowdsourced social media reaction analysis for recommendation. *International Journal of Knowledge and Systems Science*, 12(1), 1-19.
- Viriyavisuthisakul, S., Sanguansat, P., & Charnkeitkong, P. (2015). A comparison of similarity measures for online social media Thai text classification. In *Proceedings of the International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 1-6). Academic Press.

Tanatorn Tanantong received a bachelor's degree and master's degree in Computer Engineering from Suranaree University of Technology, Thailand in 2005 and 2008, respectively. In 2010, he received a scholarship from Thailand Research Fund (TRF) under The Royal Golden Jubilee Ph.D. Program and got his PhD in Computer Science from Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand in 2015. He is currently an Assistant Professor at Department of Computer Science (CS), Faculty of Science and Technology, Thammasat University, Thailand. He is also the head of Thammasat Research Unit in Data Innovation and Artificial Intelligence. In 2016, he was a visiting professor at School of Computing, Informatics, and Decision Systems Engineering, Arizona State University in USA. In 2017, he was also a Post-Doctoral researcher at Japan Advanced Institute of Science and Technology in Japan. In recent years, he has received the Ph.D. Dissertation Award (Good Level) from The National Research Council of Thailand in 2019. His interesting research includes Artificial Intelligence, Data Mining, Medical & Health Informatics, and Hospital Information Systems. He can be contacted at email: tanatorn@sci.tu.ac.th.

Monchai Parnkow received the Bachelor's degree in Statistics from Faculty of Science and Technology, Thammasat university, Thailand in 2018. He is studying in a master's degree under the supervision of Assistant Professor Dr. Tanatorn Tanantong at Department of Computer Science, Faculty of Science and Technology, Thammasat university, Thailand. He is also a research assistant at Thammasat Research Unit of Data Innovation and Artificial Intelligence. His research interests is focused on the Social Media Mining. He can be contacted at email: monchai.parn@dome.tu.ac.th.