JIT Mahasarakham University, Thailand
http://jit.it.msu.ac.th

Journal of Applied Informatics and Technology, 6 (2), 158-178

Research Article

Thai Word Segmentation using a Replacing the English Alphabet Approach to Enhance Thai Text Sentiment Analysis

Vuttichai Vichianchai[1], Sumonta Kasemvilas[2,*]

[1] Intelligent Data Analytic Laboratory, Faculty of Informatics, Mahasarakham University, Mahasarakham 44150, Thailand

[2] Hardware-Human Interface and Communications (H^2I-Comm) Laboratory, College of Computing, Khon Kaen University, Khon Kaen 40002, Thailand

[*] Corresponding Author: Sumonta Kasemvilas, sumkas@kku.ac.th

Received:
12 October 2023
Revised:
27 December 2023
Accepted:
12 January 2024

Keywords:

Misspelled Words, Text Classification, New Maximum Matching, Deepcut, Thai Writing Structure

Abstract: Thai word segmentation is an important method used that is in several document analysis applications. Dictionary-based techniques are popular for Thai word segmentation because of their high accuracy. However, these techniques are prone to errors, especially when some words are not in the dictionary. A solution to this problem is to add more vocabulary to the dictionary. Moreover, traditional techniques cannot be applied to segment misspelled words. Therefore, this research proposes a new Thai word segmentation method that replaces Thai letters with English letters. Replacing the English alphabet (REA) is a novel approach for generating short English character sequences using various formats with the same Thai writing structures. This approach improves the accuracy of Thai word segmentation, thus increasing the accuracy of Thai text classification and sentiment analysis. An evaluation is performed using Thai social media messages and Thai post comments on Pantip. These datasets are labeled by their sentiments (positive, neutral, or negative). The performance of the REA approach with the TF-G and RF techniques is better than that of the other methods, and the experimental results may be acceptable upon comparison with those of earlier well-known studies.

1. Introduction

In English, each character is written on the same level, and each word is separated by a space. In Thai, however, word boundaries are not indicated. Thai words are processed correctly according to linguistic principles. According to related research, Thai word cutting methods has been developed over a long period of time. The related approaches can be divided as follows: rule-based techniques (Charnyapornpong, 1983), dictionary-based techniques (Poovorawan & Imarom, 1986; Sornlertlamvanich, 1993; Urathammakul & Runapongsa, 2006), statistical-based techniques (Kawtrakul, Thumkanon, & Seriburi, 1995; Meknavin, Charoenpornsawat, & Kijsirikul, 1997; Kooptiwoot, 1999; Mahatthanachai *et al*., 2015), and machine learning-based techniques (Chaonithi & Prom-on, 2016; Kittinaradorn *et al*., 2019; Kongyoung, Rugchatjaroen, & Kosawat, 2018; Sunkpho & Hofmann, 2020; Tapsai *et al*., 2021; Soisoonthorn, Unger, & Maliyaem, 2023). These 4 types of methods provide relatively high word separation efficiency.

Dictionary-based techniques for Thai word segmentation (Poovorawan & Imarom, 1986; Sornlertlamvanich, 1993) are the most popular segmentation techniques because of their high accuracy (90-99%); however, they have several major limitations. For instance, the longest matching technique is prone to errors when words are not in the dictionary. A solution to this problem is to add words to the dictionary (Urathammakul & Runapongsa, 2006), but this approach expands the dictionary. In addition, the available techniques presented in several studies are complex and require large corpora for learning (Kawtrakul, Thumkanon, & Seriburi, 1995; Meknavin, Charoenpornsawat, & Kijsirikul, 1997; Kooptiwoot, 1999; Mahatthanachai *et al*., 2015; Chaonithi & Prom-on, 2016; Kittinaradorn *et al*., 2019; Kongyoung, Rugchatjaroen, & Kosawat, 2018; Sunkpho & Hofmann, 2020; Tapsai *et al*., 2021; Soisoonthorn, Unger, & Maliyaem, 2023).

The state-of-the-art word segmentation methods employ deep learning, which can learn from both data and mistakes. As an approach that uses convolutional neural networks (CNNs), Deepcut (Kittinaradorn *et al*., 2019) is a prominent Thai word segmentation method that has demonstrates an experimental accuracy of 96.57% and an F-score of 96.34%. Attacut (Chormai, Prasertsom, & Rutherford, 2019) is an improved version of Deepcut that combines character and syllable embeddings as characteristics. To achieve increased overall accuracy, the authors of (Boonkwan & Supnithi, 2018) combined models for both tasks to perform word segmentation and POS tagging. First, word boundaries are generated, and they subsequently serve as inputs for tagging. The next layer uses bidirectional recurrent neural networks (RNNs) to integrate the character-level n-gram characteristics with their surrounding contexts. In terms of the F-score metric, the accuracy of their approach exceeded 90%. In addition, the most recent state-of-the-art methods use sparse distributed representations (SDRs) (Soisoonthorn,

Unger, & Maliyaem, 2023). Two techniques are employed. 1) THDICTSDR improves the dictionary-based approach by utilizing SDRs to learn the surrounding context and combining this step with n-grams to select the correct word; and 2) THSDR uses SDRs instead of a dictionary. The results of an experiment revealed that this technique attained better performance than that of other approaches, with an F-score of 96.78% for learning all words. It could also achieve an F-score of 99.48%, which is higher than that produced of Deepcut when all sentences are learned.

However, traditional techniques cannot be employed to solve misspelled word problems. For example, a Thai social media message, "ต้องไปหาซื้ออะเธอว" (Wisesight Sentiment Corpus, 2019), can be correctly translated to "You have to go buy it". The misspelled words are "อะ" and "เธอว", which have no meaning. The correct words are "นะ", which is "a particle used at the end of a clause to show that it is a command or an entreaty (imperative mood)", and "เธอ", which means "you". Another example is "ไปๆๆๆ เก็วตัวแปป" (Wisesight Sentiment Corpus, 2019), which can be correctly translated to "Let us go. Save money for a while". The misspelled words are "เก็ว", "ตัง", and "แปป", which are meaningless, and the correct words are "เก็บ", "ตังค์", and "แป๊บ", which mean "save", "money", and "for a while", respectively. Such misspelled words cannot be segmented using traditional techniques. This problem results in decreases in the classification and sentiment analysis accuracies attained for Thai text. This is because word segmentation is an important preparation step for classification and text sentiment analysis tasks.

Therefore, to address this problem, this study proposes replacing the English alphabet (REA) for Thai word segmentation by replacing the Thai alphabet with the English alphabet through several term weighting techniques. The remainder of this work is structured as follows. Section 2 includes a review of the literature, Section 3 describes the datasets used in this study, and Section 4 describes the research methods. Section 5 contains the experimental results and a discussion. Finally, in Section 6, the conclusions are presented.

2. Literature Review

2.1 Thai Word Segmentation

Several approaches have been proposed for Thai word segmentation. Currently, the most popular techniques used in word processing cases for classification and sentiment analysis are techniques that solve the problem concerning words that do not appear in dictionaries and deep learning techniques (Esichaikul & Phumdontree, 2018; Eamwiwat *et al*., 2019; Pitichotchokphokhin *et al*., 2020; Thong-iad & Netisopakul, 2020). Both types of techniques use the segmentation method of the PyThaiNLP module with the newmm engine and the Deepcut engine. The new maximum matching technique (newmm) identifies a segmented result that contains all possible words and chooses the segmented text with the least number of words; this

technique was developed further in research conducted by Virach Sornlertlamvanich (Sornlertlamvanich, 1993) by adding words that did not appear in the dictionary. For example, in the text "นายกลับบ้าน" (meaning "Boss goes home"), when using the newmm technique, the word segmentation outcomes are (1) "นายก/ลับ/บ้าน" ("Prime minister/secret/home"), (2) "นาย/กลับ/บ้าน" ("Boss/go back/home"), and (3) "นาย/กลับบ้าน" ("Boss/go home"). Thus, the word segmentation result for this text is "นาย/กลับบ้าน". In the case in which all segmentation results have the same number of words, the algorithm chooses the word segmentation result with the longest matching method. The Deepcut technique involves the use of a Thai word segmentation library constructed via the deep neural network (DNN) technique (Kittinaradorn *et al*., 2019) (trained on the BEST corpus containing articles, novels, news, and encyclopedias (Kosawat *et al*., 2018)). This technique assigns tags to characters: ("ก, ข, ฃ, ค, ฅ, ง, จ, ช, ซ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, ฝ, พ, ภ, ม, ย, ญ, ร, ล, ว, ศ, ษ, ส, ห, and อ" with "c"), ("ค, ฉ, ผ, ฟ, ฌ, ห, and ฮ" with "n"), (" ะ, ิ, ี, ึ, ื, ุ, ู, ั, า, and ำ" with "v"), ("เ, แ, โ, ไ, and ใ" with "w"), (" ่, ้, ๊, and ๋" with "t"), ("์, ๆ, and ฯ" with "s"), ("0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ๐, ๑, ๒, ๓, ๔, ๕, ๖, ๗, ๘, and ๙" with "d"), (" " and ' ' with "q"), (space with "p"), ("a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, and z" with "s_e"), ("A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, and Z" with "b_e"), and (other with "o") to conduct learning in the text corpus (n-gram characters with their tags). For example, in the text "นายกลับบ้าน" (which means "Boss goes home"), when using the Deepcut technique, the word segmentation result is "นาย/กลับบ้าน" ("Boss/go home").

2.2 Thai Text Classification and Sentiment Analysis Techniques

In past studies, TW and machine learning (ML) techniques were employed in Thai text classification research. The BoW and TF-IDF techniques have been popularly utilized to classify Thai text from various datasets. For example, in (Inrak & Sinthupinyo, 2010), the authors combined singular value decomposition (SVD) with the BoW and TF-IDF techniques for FE and employed the decision tree (DT), naïve Bayes (NB), support vector machine (SVM), and k-nearest neighbor (k-NN) techniques for emotion classification using 200 sentences acquired from the internet; an accuracy of 0.90 was attained. In (Chirawichitchai, 2014), the authors combined the BoW, IF-IDF, and TF techniques for FE with the SVM, NB, DT and k-NN techniques to classify Thai text derived from 1,800 Thai web board posts, and they found that the BoW technique with an SVM was the most accurate approach, with an F-score of 0.779. In (Charoensuk & Sornil, 2018), the authors applied TF-IDF to the part-of-speech tagging (PoS) technique and combined it with the DT, SVM, and multinomial NB (MNB) techniques for classifying the sentiments of Thai text extracted from 2,770 Thai customer reviews, and they found that the SVM technique had

the highest F-score of 0.946. In (Hemtanon & Kittiphattanabawon, 2019), the authors employed the BoW technique for FE with the NB and SVM techniques to classify major depressive disorder (MDD) from 1,500 Thai Facebook posts and found that the SVM technique was the most accurate method, with an F-score of 0.94. In (Vichianchai & Kasemvilas, 2021), the authors combined the BoW, TF, and IF-IDF techniques; the TF-ICF and TF-G techniques for FE; and the DT, NB, SVM, random forest (RF) and multilayer perceptron (MLP) methods to classify Thai text from 2,987 clinical records and 644 Thai customer reviews and found that the RF technique was the most accurate method for classifying suicide risk, with an F-score of 0.975.

3. Datasets

In this study, we use two datasets to evaluate the performance of our proposed technique. 1) The first dataset is a Thai social media dataset containing 25,769 messages, including dictionary words, misspelled words, translations, and nouns. We remove 575 question sentiment messages and 393 numeric messages, emoji symbols and English character messages. This dataset is in the public domain under the Creative Commons Zero v1.0 universal license (Wisesight Sentiment Corpus, 2019). 2) The other dataset concerns Thai post comments on Pantip and includes 6,306 posts; Pantip is a popular Thai web board. This dataset includes dictionary words, misspelled words, translations, and nouns. We remove 151 numeric messages, emoji symbols and English character messages (Text Classification Corpus, 2021). Finally, these datasets are summarized and shown in Table 1.

4. Methodology

The proposed research methodology consists of four main processing steps. These methods include replacing the English alphabet for Thai word segmentation, term weighting, a conceptual framework, and determining the experimental settings. Each step is presented in more detail as follows.

Table 1. Class summary of the datasets

Class	Number of messages	
	Thai social media messages	Thai post comments from Pantip
Positive	6,715	385
Neutral	14,371	5,521
Negative	4,683	400

Figure 1. Levels of Thai word writing for "การเรียนรู้เป็นสิ่งสำคัญ"

4.1 Replacing the English Alphabet for Thai Word Segmentation

In this section, we explain the algorithm for replacing the English alphabet (REA) to perform Thai word segmentation. There are four levels of Thai writing and syllable structures (see Figure 1): level 1 consists of "์", "่", "้", "๊" and "๋"; level 2 consists of "ิ", "ี", "ึ", "ื", "ั", "็" and "ํ"; level 3 consists of "ก-ฮ", "ฤ-ฦ", "ไ", "ใ", "เ", "แ", "ะ", "ำ", "โ", "า", and "ๅ"; and level 4 consists of "ุ" and "ู". Thai words must be written in a specific order, starting at level 3 and followed by levels 4, 2, and 1 (Vichianchai, 2014). An example is, "การเรียนรู้เป็นสิ่งสำคัญ" (meaning "Learning is important").

The concepts of the proposed Thai word segmentation method are derived from the order of the writing levels of Thai words and their syllable structures. Our new method, REA, can generate short strings of English letters from a variety of formats that have the same Thai writing and syllable structures.

REA is a method that replaces 70 Thai letters (i.e., 46 consonants, 17 vowels, 4 tones, and 3 others) with 17 English letters, as shown in Table 2. REA letters can be combined to represent Thai words and reduce the use of written Thai words; for example, (1) "cc" can be used as a substitution for the following words: "พบ" (meaning "found") and "ลบ" (meaning "delete"). (2) "ccc" can be used as a substitution for the following words: "ขนม" (meaning "dessert") and "กลม" (meaning "round"). (3) "cAc" can be used as a substitution for the following words: "การ" (meaning "things or things to do"), "งาน" (meaning "work") and "มาก" (meaning "a lot"). In addition, the REA patterns are derived from Thai words and transliterated words that include "์", namely, (1) "VcccR" can be used as a substitute for the following words: "เกณฑ์" (meaning "criterion") and "เบอร์" (meaning "number"); (2) "cuccR" can be used as a substitute for the following words: "พันธ์" (meaning "breed") and "ฉันท์" (meaning "stanza"); and (3) "VcAcR" can be used as a substitute for the following words: "เสาร์" (meaning "Saturday") and "เซาว์" (meaning "South").

Table 2. English letters that replace Thai letters in REA

English alphabet using REA	Thai alphabet
c	ก, ข, ฃ, ค, ฅ, ฆ, ง, จ, ฉ, ช, ซ, ฌ, ฎ, ฏ, ฐ, ฑ, ฒ, ณ, ด, ต, ถ, ท, ธ, น, บ, ป, ผ, ฝ, พ, ฟ, ภ, ม, ย, ญ, ร, ล, ว, ศ, ษ, ส, ห, ฬ, อ, ฮ, ฤ, and ฦ
a	ะ
A	า and ๅ
i	ิ and ี
e	ึ
E	ื
v	ุ and ู
V	เ and แ
o	โ
I	ไ and ใ
u	ั
U	ำ
R	์
T	็
t	่, ้, ๊, and ๋
Y	ๆ
L	ฯ

The method proposed in this research utilizing REA for Thai word segmentation is based on Figure 2.

Figure 2 shows the process of segmenting Thai words with REA, which is described as follows:

1. First, REA letters are loaded into an REA array.

2. The REA format is used for comparing segmented Thai words and consists of 1314 patterns, such as cAc (i.e., “นาย”, “งาน”, and “จาน”, ccuc (i.e., “กลับ”, “หลับ”, “ผลัก”, and “หนัง”, and ctAc (i.e., “บ้าน”, “ป้าย”, and “ย้าย”).

3. The Thai text that needs to be segmented and the Thai letters replaced with English letters are read according to Table 2.

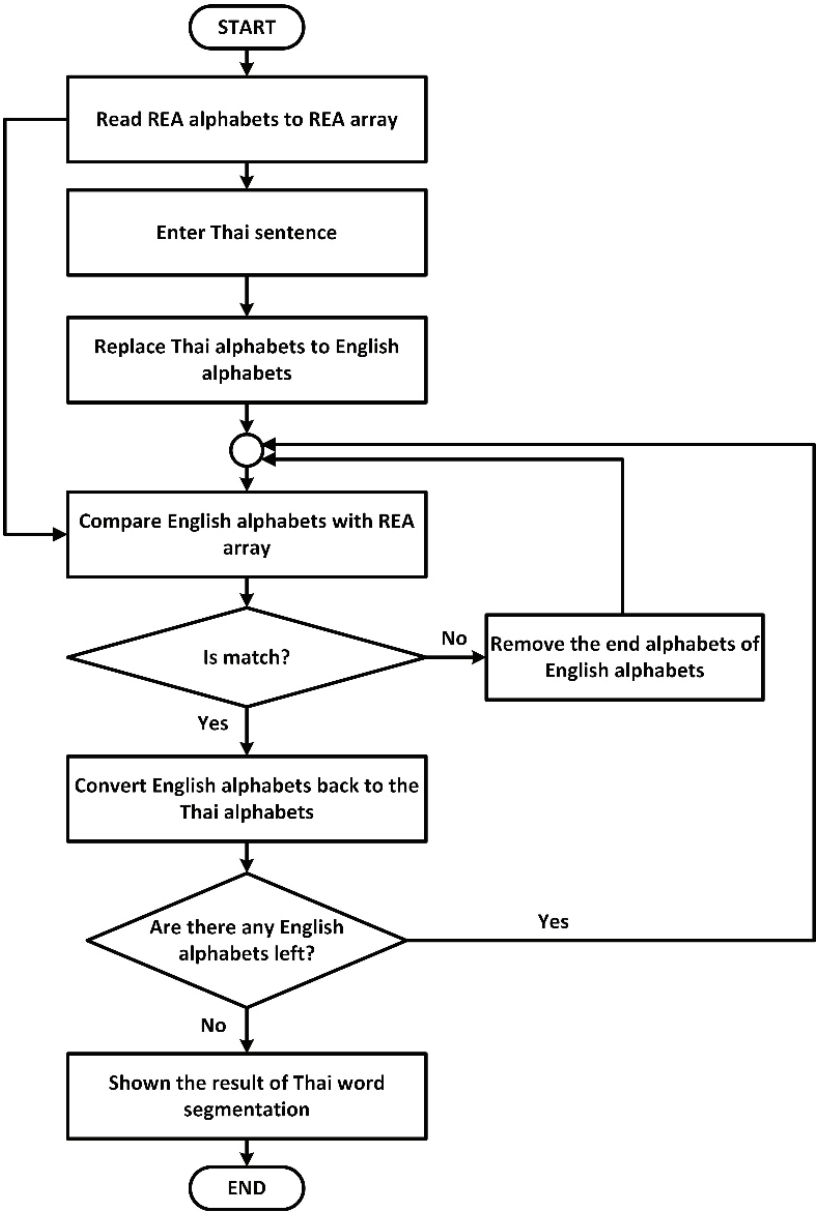

Figure 2. Process of Thai word segmentation with REA

4. The letters replaced in step 3 with the REA letters are matched. If they match, those letters are converted back to the Thai letters in the previous text. Otherwise, the end letters of the English letters are removed and compared until the text can be compared. This process is repeated to compare the remaining English letters against the REA letters until the entire word is segmented.

The following examples involve the word segments in the sentence “นายกลับบ้าน” (meaning “Boss goes home”), as shown in Table 3, which yield “นาย/กลับ/บ้าน” as the word segmentation result.

Table 3. Example of Thai word segmentation with "นายกลับบ้าน"

English letters must be replaced by word segments	Results comparison with REA	Segmented words compared to REA words	Thai letters cut off from the end of the sentence	English letters cut off at the ends of sentences
cAcccucctAc	No		น	c
cAcccucctA	No		บ้าน	ctAc
cAcccuc	No		บบ้าน	cctAc
cAcc	No		ลับบ้าน	cucctAc
cAc	No		กลับบ้าน	ccucctAc
cAc	Yes	cAc ("นาย")		ccucctAc (Will be compared to segment words)
ccucctAc	No		น	c
ccucctA	No		บ้าน	ctAc
ccuc	Yes	ccuc ("กลับ")		ctAc (Will be compared to segment words)
ctAc	Yes	ctAc ("บ้าน")		There is no text left to be segmented. The word segmentation results are shown

4.2 Term Weighting

The term weighting (TW) technique is a well-known preprocessing step for text classification in which appropriate weights are assigned to each term in documents with a feature vector structure to achieve improved text classification accuracy. The TW technique is employed in research to classify text as follows: 1) bag-of-words (BoW) is a technique for extracting features from text by weighting an index term found in a sentence as 1 and an index term not found in a sentence as 0; 2) the term frequency-inverse document frequency (TF-IDF) is the number of words that appear in a document divided by the total number of words in the document; 3) the term frequency-inverse corpus frequency (TF-ICF) is the inverse of the document frequency in each group of terms being considered and the total number of documents; 4) the term frequency and inverse gravity moment (TF-IGM) modifies and improves the TF-IDF; and 5) the term frequency with Gaussian (TF-G) techniques finds the weights of terms from the log values of the reviewed term frequencies multiplied by the distribution value of the

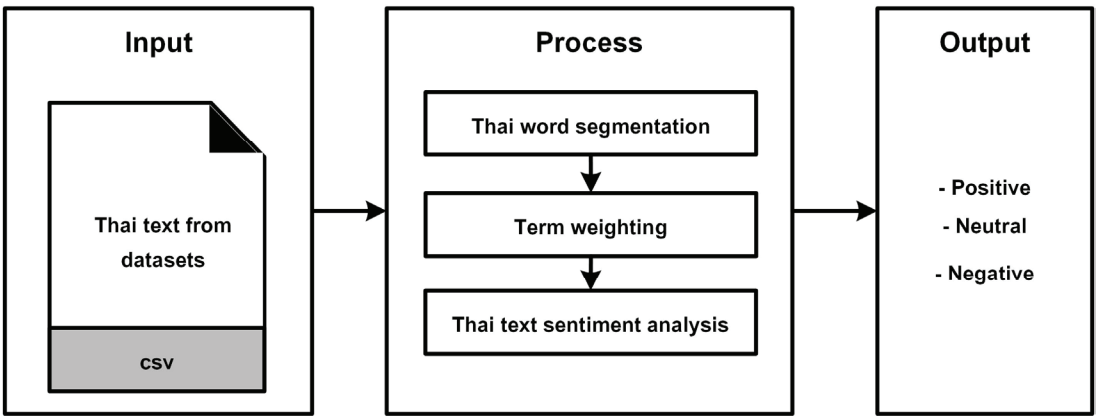

Figure 3. Conceptual framework of this research

terms in the reviews of each class (Chen *et al*., 2016; Dogan & Uysal, 2019; Vichianchai & Kasemvilas, 2021).

4.3 Conceptual Framework

The conceptual framework (see Figure 3.) includes an input, a process, and an output. In this research, we use the datasets from Section 3 as the input data for the processing steps, which include Thai word segmentation, term weighting, and sentiment analysis via machine learning to predict three classes of outcomes (positive, neutral, and negative). Each step is presented in more detail as follows.

In this research, we compare the proposed Thai word segmentation technique (REA) with other popular techniques (newmm and Deepcut) in text sentiment analysis cases because the latter techniques are the most popular word processing methods for Thai text classification and sentiment analysis (Esichaikul & Phumdontree, 2018; Eamwiwat *et al*., 2019; Pitichotchokphokhin *et al*., 2020; Thong-iad & Netisopakul, 2020). In addition, newmm is a state-of-the-art Thai word segmentation method based on dictionaries, and Deepcut is a state-of-the-art Thai word segmentation method based on machine learning. This step imports datasets for word segmentation by processing each technique and then uses the results produced by each word segmentation technique to find weights in the term weighting step.

The term weighting step takes the text obtained from the Thai word segmentation task by deriving the term weights from the following techniques.

1) BoW: Determining the weights of terms contained in the text yields a weight of 1 for words that are found in the text and a weight of 0 for words that are not found in the text.

2) TF-IDF: A phrase is identified by multiplying its frequency in a message by its log value over a group of messages in the dataset.

3) TF-ICF: The weights of terms are calculated based on the log values of the term frequencies in a message, where each class of terms is considered, and the total number of messages.

4) TF-IGM: The term weights are determined by multiplying the inverse gravitational moment of the interclass message frequency counts by class-specific counts.

5) TF-G: The weights of terms are calculated by multiplying the log value of a term's frequency in a message by the term's distribution value in each class's messages.

The final step is to import the term weights and sentiment classes (positive, neutral, and negative) into each machine learning technique. We select popular machine learning techniques from Section 2.2, including MNB, an SVM, an RF, and an MLP.

To summarize, in this step, we design two main experiments: 1) comparing the performance of the newmm, Deepcut, and REA techniques for Thai word segmentation; and 2) comparing the performance of Thai word segmentation techniques (newmm, Deepcut, and REA) and TW techniques (BoW, TF-IDF, TF-ICF, TF-IGM, and TF-G) with that of ML techniques (MNB, SVM, RF, and MLP) for Thai text sentiment analysis. Here, the precision, recall, F-score, and processing time metrics are used to evaluate the Thai word segmentation performance achieved by each method. The precision, recall, F-score, and root mean square error (RMSE) metrics are used to evaluate the Thai text classification performance of each approach (Forman, 2003).

5. Results and Discussion

5.1 Results

5.1.1 Thai Word Segmentation Results

To evaluate the efficiency of the tested Thai word segmentation approaches, we employed a word counting method. A segmentation example involving the text, "ผมขับรถยนตไปทำงาน", which can be correctly written as "ผมขับรถยนต์ไปทำงาน" (meaning "I drive a car to work"), is provided as follows. (1) The Thai word segmentation results obtained using the newmm technique were "ผม/ขับรถ/ยนต/ไป/ทำงาน", giving true-positive, false-positive, and false-negative values of 4 ("ผม","ขับรถ","ไป", and "ทำงาน"), 0, and 1 ("ยนต"), respectively. The precision, recall, and F-score were 1.00, 0.80, and 0.8889, respectively. (2) The Thai word segmentation results obtained using the Deepcut technique were "ผม/ขับ/รถ/ยนต/ไป/ ทำ/งาน", yielding true-positive, false-positive, and false-negative values of 6 ("ผม", "ขับ", "รถ", "ไป", "ทำ", and "งาน"), 0, and 1 ("ยนต"), respectively. The precision, recall, and F-score were 1.00, 0.8571, and 0.9231, respectively. (3) The Thai word segmentation results obtained using the REA technique were "ผม/ขับ/ รถยนต/ไป/ทำงาน", yielding true-positive, false-positive, and false-negative values of 4 ("ผม", "ขับ",

“ไป”, and “ทำงาน”), 1 (“รถยนต”), and 0, respectively. The precision, recall, and F-score were 0.80, 1.00, and 0.8889, respectively.

The correct word segmentation result for this text is “ผม/ขับ/รถยนต/ไป/ทำงาน”, (“I/drive/car/go/work”). The Thai word segmentation results obtained with the REA technique were more accurate in terms of recall than those of the newmm and Deepcut techniques.

Table 4 shows that the REA technique for Thai word segmentation was more accurate than the newmm and Deepcut techniques for Thai social media messages (negative, neutral, and positive). REA obtained recall and F-score values of 0.9553 and 0.9367, respectively. Additionally, the proposed Thai word segmentation method was faster than the newmm and Deepcut techniques were, with an average time requirement of 0.35 seconds per message. On the other hand, the newmm technique had the highest precision—0.9357—for word segmentation. The proposed method (REA) has higher accuracy than the newmm and Deepcut techniques because the 1,314 patterns covered a wide range of written Thai words.

Table 4. Thai word segmentation results

Thai word segmentation techniques	Precision	Recall	F-score	Average time processing (second)
Newmm	0.9357	0.8979	0.9112	0.41
Deepcut	0.9177	0.8652	0.8932	21.58
REA	0.9197	0.9553	0.9367	0.35

Table 5. Thai text sentiment analysis Results

Techniques	Thai social media messages				Thai post comments on Pantip			
	Precision	Recall	F-score	RMSE	Precision	Recall	F-score	RMSE
newmm+BoW+MNB	0.510	0.519	0.512	0.276	0.560	0.569	0.562	0.277
newmm +TF-IDF+MNB	0.539	0.537	0.537	0.272	0.589	0.587	0.587	0.273
newmm +TF-ICF+MNB	0.319	0.323	0.319	0.323	0.369	0.373	0.369	0.324
newmm +TF-IGM+MNB	0.507	0.516	0.510	0.276	0.557	0.566	0.560	0.277
newmm +TF-G+MNB	0.455	0.463	0.459	0.298	0.505	0.513	0.509	0.299
newmm +BoW+SVM	0.510	0.495	0.495	0.287	0.560	0.545	0.545	0.288
newmm +TF-IDF+SVM	0.527	0.505	0.507	0.276	0.577	0.555	0.557	0.277

Table 5. Thai text sentiment analysis Results (cont.)

Techniques	Thai social media messages				Thai post comments on Pantip			
	Precision	Recall	F-score	RMSE	Precision	Recall	F-score	RMSE
newmm +TF-ICF+SVM	0.818	0.793	0.797	0.178	0.868	0.843	0.847	0.179
newmm +TF-IGM+SVM	0.493	0.477	0.477	0.287	0.543	0.527	0.527	0.288
newmm +TF-G+SVM	0.870	0.846	0.848	0.172	0.920	0.896	0.898	0.173
newmm +BoW+RF	0.548	0.474	0.469	0.286	0.598	0.524	0.519	0.287
newmm +TF-IDF+RF	0.491	0.418	0.387	0.296	0.541	0.468	0.437	0.297
newmm +TF-ICF+RF	0.903	0.870	0.874	0.165	0.953	0.920	0.924	0.166
newmm +TF-IGM+RF	0.573	0.512	0.516	0.286	0.623	0.562	0.566	0.287
newmm +TF-G+RF	0.923	0.898	0.901	0.163	0.973	0.948	0.951	0.164
newmm +BoW+MLP	0.479	0.475	0.475	0.294	0.529	0.525	0.525	0.295
newmm +TF-IDF+MLP	0.445	0.449	0.424	0.296	0.495	0.499	0.474	0.297
newmm +TF-ICF+MLP	0.447	0.456	0.448	0.295	0.497	0.506	0.498	0.296
newmm +TF-IGM+MLP	0.480	0.477	0.476	0.297	0.530	0.527	0.526	0.298
newmm +TF-G+MLP	0.482	0.478	0.477	0.297	0.532	0.528	0.527	0.298
Deepcut+BoW+MNB	0.460	0.474	0.458	0.297	0.510	0.524	0.508	0.298
Deepcut+TF-IDF+MNB	0.491	0.491	0.491	0.297	0.541	0.541	0.541	0.298
Deepcut+TF-ICF+MNB	0.332	0.337	0.334	0.332	0.382	0.387	0.384	0.333
Deepcut+TF-IGM+MNB	0.457	0.474	0.459	0.296	0.507	0.524	0.509	0.297
Deepcut+TF-G+MNB	0.430	0.442	0.430	0.296	0.480	0.492	0.480	0.297
Deepcut+BoW+SVM	0.462	0.449	0.450	0.296	0.512	0.499	0.500	0.297
Deepcut+TF-IDF+SVM	0.524	0.516	0.516	0.276	0.574	0.566	0.566	0.277
Deepcut+TF-ICF+SVM	0.913	0.905	0.906	0.163	0.963	0.955	0.956	0.164
Deepcut+TF-IGM+SVM	0.469	0.456	0.456	0.286	0.519	0.506	0.506	0.287
Deepcut+TF-G+SVM	0.921	0.912	0.913	0.163	0.971	0.962	0.963	0.164
Deepcut+BoW+RF	0.503	0.474	0.467	0.286	0.553	0.524	0.517	0.287
Deepcut+TF-IDF+RF	0.520	0.439	0.417	0.286	0.570	0.489	0.467	0.287
Deepcut+TF-ICF+RF	0.931	0.912	0.914	0.161	0.981	0.962	0.964	0.162
Deepcut+TF-IGM+RF	0.505	0.474	0.472	0.286	0.555	0.524	0.522	0.287

Table 5. Thai text sentiment analysis Results. (cont.)

Techniques	Thai social media messages				Thai post comments on Pantip			
	Precision	Recall	F-score	RMSE	Precision	Recall	F-score	RMSE
Deepcut+TF-G+RF	0.940	0.926	0.927	0.159	0.990	0.976	0.977	0.160
Deepcut+BoW+MLP	0.470	0.467	0.467	0.286	0.520	0.517	0.517	0.287
Deepcut+TF-IDF+MLP	0.410	0.414	0.392	0.299	0.460	0.464	0.442	0.300
Deepcut+TF-ICF+MLP	0.397	0.400	0.397	0.296	0.447	0.450	0.447	0.297
Deepcut+TF-IGM+MLP	0.469	0.463	0.461	0.294	0.519	0.513	0.511	0.295
Deepcut+TF-G+MLP	0.477	0.470	0.470	0.295	0.527	0.520	0.520	0.296
REA+BoW+MNB	0.514	0.522	0.516	0.275	0.564	0.572	0.566	0.276
REA+TF-IDF+MNB	0.542	0.539	0.540	0.276	0.592	0.589	0.590	0.277
REA+TF-ICF+MNB	0.343	0.343	0.343	0.299	0.393	0.393	0.393	0.300
REA+TF-IGM+MNB	0.516	0.517	0.517	0.276	0.566	0.567	0.567	0.277
REA+TF-G+MNB	0.505	0.509	0.498	0.276	0.555	0.559	0.548	0.277
REA+BoW+SVM	0.518	0.499	0.499	0.277	0.568	0.549	0.549	0.278
REA+TF-IDF+SVM	0.611	0.593	0.595	0.278	0.661	0.643	0.645	0.279
REA+TF-ICF+SVM	0.917	0.915	0.916	0.161	0.967	0.965	0.966	0.162
REA+TF-IGM+SVM	0.497	0.481	0.481	0.279	0.547	0.531	0.531	0.280
REA+TF-G+SVM	0.927	0.917	0.918	0.157	0.977	0.967	0.968	0.158
REA+BoW+RF	0.561	0.498	0.499	0.279	0.611	0.548	0.549	0.280
REA+TF-IDF+RF	0.584	0.507	0.531	0.278	0.634	0.557	0.581	0.279
REA+TF-ICF+RF	0.935	0.920	0.919	0.157	0.985	0.970	0.969	0.158
REA+TF-IGM+RF	0.575	0.516	0.520	0.276	0.625	0.566	0.570	0.277
REA+TF-G+RF	0.945	0.930	0.929	0.153	0.995	0.983	0.979	0.154
REA+BoW+MLP	0.508	0.505	0.505	0.277	0.558	0.555	0.555	0.278
REA+TF-IDF+MLP	0.528	0.495	0.475	0.276	0.578	0.545	0.525	0.277
REA+TF-ICF+MLP	0.449	0.457	0.457	0.279	0.499	0.507	0.507	0.280
REA+TF-IGM+MLP	0.470	0.467	0.466	0.281	0.520	0.517	0.516	0.282
REA+TF-G+MLP	0.535	0.510	0.510	0.278	0.585	0.560	0.560	0.279

5.1.2 Thai Text Sentiment Analysis Results

The results of the sentiment analysis experiment are shown in Table 5. The bold numbers indicate that the precision, recall, and F-score values of the other techniques using REA were greater than those of the other techniques using newmm and Deepcut for Thai text classification. The REA and TF-G algorithms with the RF technique were the most accurate techniques for Thai text classification, with higher precision, recall, F-score, and accuracy values than those of the other techniques based on 10-fold cross-validation. The REA and TF-G algorithms with the RF technique were the most accurate Thai text sentiment analysis methods, with precision, recall, F-score, and RMSE values of 0.945, 0.93, 0.929, and 0.153, respectively, for Thai social media messages. The REA and TF-G algorithms with the RF technique were the most accurate Thai text sentiment analysis methods, with precision, recall, F-score, and RMSE values of 0.995, 0.983, 0.979, and 0.154, respectively, for the Thai post comments on Pantip.

5.2 Discussion Concerning Thai Word Segmentation and Thai Text Classification

Three main types of approaches have been employed for Thai word segmentation, including rule-based methods (Thairatananond, 1981; Charnyapornpong, 1983), dictionary-based methods (Poovorawan & Imarom, 1986; Sornlertlamvanich, 1993; Urathammakul & Runapongsa, 2006), and methods that learn from large Thai text corpora (Kawtrakul, Thumkanon, & Seriburi, 1995; Meknavin, Charoenpornsawat, & Kijsirikul, 1997; Kooptiwoot, 1999; Mahatthanachai *et al*., 2015; Chaonithi & Prom-on, 2016; Kittinaradorn *et al*., 2019; Kongyoung, Rugchatjaroen, & Kosawat, 2018; Sunkpho & Hofmann, 2020). These techniques have been proven to provide high word segmentation accuracy; for example, rule-based techniques are 85-96% accurate (Thairatananond, 1981; Charnyapornpong, 1983), dictionary-based techniques are 90-99% accurate (Poovorawan & Imarom, 1986; Sornlertlamvanich, 1993; Urathammakul & Runapongsa, 2006), and learning techniques are 79-97% accurate. Therefore, we chose the newmm and Deepcut techniques (Sornlertlamvanich, 1993; Kittinaradorn *et al*., 2019), which have high Thai word segmentation accuracy, for comparison with the REA technique.

In our study, we employed a Thai social media message dataset that was suitable for experimentation because it contains text with many misspelled words. An example message from this collection is “ไปๆๆๆ เก็วตังแปป” (Wisesight Sentiment Corpus, 2019), which means “Let us go. Save money for a while”. The misspelled words are “เก็ว”, “ตัง”, and “แปป”, which have no meaning, whereas “เก็บ”, “ตังค์”, and “แป๊บ” mean “save”, “money”, and “for a while”, respectively. The results of the experiments showed that the REA technique was more accurate than the MM and Deepcut techniques due to the following limitations.

The limitations of the MM technique are that the program can only segment words contained in dictionaries. Misspelled words, transliteration words and proper names cannot be segmented by employing the MM technique (Sornlertlamvanich, 1993). The Deepcut technique is applicable for segmenting Thai words because the program learns from a large text corpus (Kosawat *et al*., 2018; Kittinaradorn *et al*., 2019). Thus, ambiguous words, transliterated words, and proper names can be segmented if the utilized text corpus contains those particular words. However, misspelled words remain indecipherable for this method because misspelled words are not contained in the learned text corpus.

REA is capable of segmenting misspelled words because the English alphabet set is comprehensive and has a similar writing-level syllable structure to that of the Thai word set. For example, “ไปๆๆๆ เก็วตังแปป” (Wisesight Sentiment Corpus, 2019) is a message with misspelled words. The Thai word segmentation result obtained using the newmm technique was “ไป/ๆๆๆ/ /เก็ว/ตัง/ แป/ป”, producing a true-positive value of 0.67. The Thai word segmentation result obtained using the Deepcut technique was “ไป/ๆ/ๆ/ ๆ/ /เก็ว/ตังแปป”, yielding a true-positive value of 0.82. The Thai word segmentation result obtained using the REA technique was “ไป/ ๆ/ๆ/ๆ/เก็ว/ตัง/แปป”, providing a true-positive value of 1.

Additionally, the proposed method for Thai word segmentation was faster than the newmm and Deepcut techniques for text containing approximately 935 characters, and in the experimental results, when the number of characters exceeded 1,000, the segmentation speed of the proposed method was slower than that of the MM technique because the proposed technique needed time to convert Thai characters to English characters (see Table 2).

REA can increase the accuracy of Thai text sentiment analysis because accurate word segmentation, which is an important preparation step, solves the problem of misspelled words, resulting in misspelled words not being segmented into other words and possibly having incorrect meanings. We observed that the true-positive rates produced by the newmm and Deepcut techniques decreased, which resulted in inaccurate word weights in the vectors and reduced text classification accuracy.

The REA and TF-G methods with the RF technique were the most accurate method for analyzing the sentiments of Thai social media messages and Thai post comments on Pantip because TF-G is a technique that calculates word weights from the distribution of words in each category, whether they are misspelled or correct (Vichianchai & Kasemvilas, 2021). The RF technique adds multiple trees and divides the features of each tree according to its specific features to increase the diversity and independence of each tree. In addition, an RF is a technique that reduces the overfitting caused by the use of many features

in DTs and achieves improved accuracy; this technique is flexible when applied to classification and regression problems, and it can work well with categorical values (Horsuwan *et al*., 2019; Wongpatikaseree *et al*., 2021).

6. Conclusion

This research aimed to develop a new Thai word segmentation method. REA is a novel approach for generating short English character strings in various formats so that they have the same Thai writing and syllable structures. The proposed method is better at segmenting Thai words, especially misspelled words, than competing approaches.

Therefore, the proposed technique is more appropriate for performing Thai word segmentation on documents that contain misspelled words. Presently, we often find articles with misspelled words, such as social media messages, news content, and web-based messages. The proposed Thai word segmentation method was faster than the newmm technique for addressing text containing approximately 935 characters. Additionally, the REA technique can enhance the accuracy of Thai text sentiment analysis, especially for Thai texts with many misspelled words.

Although REA can represent Thai words for word segmentation comparisons, it may not be possible to segment future Thai words using the current REA construction rules.

Due to the limitations of our framework, the diagnosis process focused only on misspelled words. In a future study, the proposed framework can be extended to word processing and correcting misspelled words to improve text before it is sent, enabling the TW and ML techniques to perform Thai text classification and sentiment analysis tasks.

References

Boonkwan, P. & Supnithi, T. (2018). Bidirectional deep learning of context representation for joint word segmentation and POS tagging. *In: Le, NT., van Do, T., Nguyen, N., Thi, H. (eds) Advanced Computational Methods for Knowledge Engineering. ICCSAMA 2017. Advances in Intelligent Systems and Computing, vol 629*. Springer, Cham. https://doi.org/10.1007/978-3-319-61911-8_17

Chaonithi, K. & Prom-on, S. (2016). A hybrid approach for Thai word segmentation with crowdsourcing feedback system. *Proceeding of the 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2016),* Chiang Mai, Thailand, 28 June 2016, 1-6. https://doi.org/10.1109/ECTICon.2016.7561298

Charnyapornpong, S. (1983). *A Thai syllable separation algorithm* [Master's Thesis, Asian Institute of Technology].

Charoensuk, J. & Sornil, O. (2018). A hierarchical emotion classification technique for Thai reviews. *Journal of ICT Research and Applications*, *12*(3), 280-296. https://doi.org/10.5614/itbj.ict.res.appl.2018.12.3.6

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications*, 66, 245-260. https://doi.org/10.1016/j.eswa.2016.09.009

Chirawichitchai, N. (2014). Emotion classification of Thai text based using term weighting and machine learning techniques. *Proceeding of the 11th International Joint Conference on Computer Science and Software Engineering (JCSSE 2014)*, 91-96. https://doi.org/10.1109/JCSSE.2014.6841848

Chormai, P., Prasertsom, P., & Rutherford, A. (2019). *Attacut: A fast and accurate neural thai word segmenter*. *arXiv*, preprint arXiv:1911.07056. https://doi.org/10.48550/arXiv.1911.07056

Dogan, T. & Uysal, A. K. (2019). Improved inverse gravity moment term weighting for text classification. *Expert Systems with Applications*, 130, 45-59. https://doi.org/10.1016/j.eswa.2019.04.015

Eamwiwat, C., Thanasutives, P., Saetia, C., & Chalothorn, T. (2019). Using label noise filtering and ensemble method for sentiment analysis on Thai social data. *Proceeding of the 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP 2019),* Chiang Mai, Thailand, 30 October 2019, 1-6. https://doi.org/10.1109/iSAI-NLP48611.2019.9045419.

Esichaikul, V., & Phumdontree, C. (2018). Sentiment analysis of Thai financial news. *Proceedings of the 2nd International Conference on Software and e-Business (ICSEB 2018)*, Zhuhai, China, 18 December 2018, 39-43. https://doi.org/10.1145/3301761.3301773

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3, 1289-1305. https://dl.acm.org/doi/10.5555/944919.944974

Hemtanon, S. & Kittiphattanabawon, N. (2019). An automatic screening for major depressive disorder from social media in Thailand. *Proceeding of the 10th National & International Conference,* 103-113. http://journalgrad.ssru.ac.th/index.php/8thconference/article/view/1880

Horsuwan, T., Kanwatchara, K., Vateekul, P., & Kijsirikul, B. (2019). A comparative study of pretrained language models on Thai social text categorization. *In: Nguyen, N., Jearanaitanakij, K., Selamat, A., Trawiński, B., Chittayasothorn, S. (eds) Intelligent Information and Database Systems. ACIIDS 2020. Lecture Notes in Computer Science(), vol 12033*. Springer, Cham. https://doi.org/10.1007/978-3-030-41964-6_6

Inrak, P. & Sinthupinyo, S. (2010). Applying latent semantic analysis to classify emotions in Thai text. *Proceeding of the 2nd International Conference on Computer Engineering and Technology (ICCET 2010),* Chengdu, China, 16 April 2010, 450-454. https://doi.org/10.1109/ICCET.2010.5486137

Kawtrakul, A., Thumkanon, C., & Seriburi, S. (1995). A statistical approach to Thai word filtering. *Proceeding of the 2nd Symposium on Natural Language Processing (SNLP'95)*, 2-4 August 1995, 398-406.

Kittinaradorn, R. *et al*. (2019). Deepcut: A Thai word tokenization library using deep neural network. Retrieved 10 November 2023. Retrieved from https://github.com/rkcosmos/Deepcut.

Kongyoung, S., Rugchatjaroen, A., Kosawat, K. (2018). TLex+: A hybrid method using conditional random fields and dictionaries for Thai word segmentation. *In: Theeramunkong, T., Skulimowski, A., Yuizono, T., Kunifuji, S. (eds) Recent Advances and Future Prospects in Knowledge, Information and Creativity Support Systems. KICSS 2015. Advances in Intelligent Systems and Computing, vol 685*. Springer, Cham. https://doi.org/10.1007/978-3-319-70019-9_10

Kooptiwoot, C. (1999). Segmentation of ambiguous Thai words by inductive logic programming. Bangkok: Chulalongkorn University.

Kosawat, K. *et al*. (2009). BEST 2009: Thai word segmentation software contest. *Proceeding of the Eighth International Symposium on Natural Language Processing (SNLP 2009)*, Bangkok, Thailand, 20 October 2009, 83-88. https://doi.org/10.1109/SNLP.2009.5340941

Mahatthanachai, C., Malaivongs, K., Tantranont, N., & Boonchieng, E. (2015). Development of thai word segmentation technique for solving problems with unknown words. *Proceeding of the International Computer Science and Engineering Conference (ICSEC 2015), Chiang Mai, Thailand, 23 November 2015,* 1-6. https://doi.org/10.1109/ICSEC.2015.7401423

Meknavin, S., Charoenpornsawat, P., & Kijsirikul, B. (1997). Feature-based Thai word segmentation. *Proceedings of Natural Language Processing Pacific Rim Symposium (NLPRS'97), Phuket, Thailand, 2 December 1997,* 41-46.

Pitichotchokphokhin, P., Chuangkrud, P., Kalakan, K., Suntisrivaraporn, B., Leelanupab, T., & Kanungsukkasem, N. (2020). Discover underlying topics in Thai news articles: A comparative study of probabilistic and matrix factorization approaches. *Proceeding of the 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON 2020), Phuket, Thailand, 24 June 2020,* 759-762. https://doi.org/10.1109/ECTI-CON49241.2020.9158065

Poovorawan, Y., & Imarom, V. (1986). Thai syllable separater by dictionary. Proceedings of the 9th National Conference on Electrical Engineering, Khon Kaen, Thailand.

Soisoonthorn, T., Unger, H., & Maliyaem, M. (2023). Thai word segmentation with a brain-Inspired sparse distributed representations learning memory. *Computational Intelligence and Neuroscience*, 2023. https://doi.org/10.1155/2023/8592214

Sornlertlamvanich, V. (1993). Word segmentation for Thai in machine translation system. *Machine translation*. [In Thai].

Sunkpho, J., Hofmann, M. (2020). Thai words segmentation using an unsupervised learning technique. *In: Meesad, P., Sodsee, S. (eds) Recent Advances in Information and Communication Technology 2020. IC2IT 2020. Advances in Intelligent Systems and Computing, vol 1149.* Springer, Cham. https://doi.org/10.1007/978-3-030-44044-2_9

Tapsai, C. *et al*. (2021). TLS-ART-MC, A new algorithm for Thai word segmentation. *In: Thai Natural Language Processing. Studies in Computational Intelligence, vol 918*. Springer, Cham. https://doi.org/10.1007/978-3-030-56235-9_3

Text classification corpus. (2021). Pittawat2542/krathu-500. Retrieved 10 November 2023. Retrieved from https://github.com/Pittawat2542/krathu-500.

Thairatananond, Y. (1981). Towards the design of a Thai text syllable analyzer. [Master's Thesis, Asian Institute of Technology].

Thong-iad, K. & Netisopakul, P. (2020). Comparison of Thai sentence sentiment tagging methods using Thai sentiment resource *In: Boonyopakorn, P., Meesad, P., Sodsee, S., Unger, H. (eds) Recent Advances in Information and Communication Technology 2019. IC2IT 2019. Advances in Intelligent Systems and Computing, vol 936.* Springer, Cham. https://doi.org/10.1007/978-3-030-19861-9_9

Urathammakul, P. & Runapongsa, K. (2006). Improved rule-based and new dictionary for Thai Word segmentation. *Proceedings of the 3rd Joint Conference on Computer Science and Software Engineering*, Bangkok, Thailand, 34-40. [In Thai]

Vichianchai, V. (2014). The comparison of Thai word segmentation with Thai writing structures and syllable structures. *Journal of Science and Technology Mahasarakham University*, 33(5), 503-509. [In Thai]

Vichianchai, V. & Kasemvilas, S. (2021). A new term frequency with Gaussian technique for text classification and sentiment analysis. *Journal of ICT Research & Applications*, 15(2), 152-168. https://doi.org/10.5614/itbj.ict.res.appl.2021.15.2.4

Wisesight Sentiment Corpus. (2019). PyThaiNLP/wisesight-sentiment. Retrieved 18 September 2022. Retrieved from https://github.com/PyThaiNLP/wisesight-sentiment.

Wongpatikaseree, K., Kaewpitakkun, Y., Yuenyong, S., Matsuo, S., & Yomaboot, P. (2021). Emocnn: Encoding emotional expression from text to word vector and classifying emotions—A case study in thai social network conversation. *Engineering Journal*, 25(7), 73-82. https://doi.org/10.4186/ej.2021.25.7.73