A New Term Frequency with Gaussian Technique for Text Classification and Sentiment Analysis

Vuttichai Vichianchai & Sumonta Kasemvilas*

Hardware-Human Interface and Communications (H^2I-Comm) Laboratory,
Department of Computer Science, Faculty of Science, Khon Kaen University,
123 Mittraphap Road., Nai Mueang, Mueang Khon Kaen 40002, Thailand
*E-mail: sumkas@kku.ac.th

Abstract. This paper proposes a new term frequency with a Gaussian technique (TF-G) to classify the risk of suicide from Thai clinical notes and to perform sentiment analysis based on Thai customer reviews and English tweets of travelers that use US airline services. This research compared TF-G with term weighting techniques based on Thai text classification methods from previous researches, including the bag-of-words (BoW), term frequency (TF), term frequency-inverse document frequency (TF-IDF), and term frequency-inverse corpus document frequency (TF-ICF) techniques. Suicide risk classification and sentiment analysis were performed with the decision tree (DT), naïve Bayes (NB), support vector machine (SVM), random forest (RF), and multilayer perceptron (MLP) techniques. The experimental results showed that TF-G is appropriate for feature extraction to classify the risk of suicide and to analyze the sentiments of customer reviews and tweets of travelers. The TF-G technique was more accurate than BoW, TF, TF-IDF and TF-ICF for term weighting in Thai suicide risk classification, for term weighting in sentiment analysis of Thai customer reviews for Burger King, Pizza Hut, and Sizzler restaurants, and for the sentiment analysis of English tweets of travelers using US airline services.

Keywords: *customer reviews; clinical note; machine learning; natural language processing; suicide risk classification; tweets of travelers.*

1 Introduction

Term weighting refers to the process of indexing text in a document to determine the weight of each word in the document. The weight of a word in a document is vital for improving the data retrieval efficiency [1]. Term weighting is a commonly accepted technique in text classification for determining the importance of words in a document or query, because each word is not equally significant. This approach is important for data retrieval systems and demonstrates great potential for improving the data extraction performance of data retrieval systems [2]. Term weighting is a well-known step in creating feature vectors to enhance the accuracy of text classification [3]. Thai text classification in previous research is limited. Previous studies used traditional

Received February 19th, 2021, 1st Revision July 4th, 2021, 2nd Revision July 17th, 2021, Accepted for publication July 21ST, 2021.
Copyright © 2021 Published by IRCS-ITB, ISSN: 2337-5787, DOI: 10.5614/itbj.ict.res.appl.2021.15.2.4

term weighting and machine learning (ML) techniques, but their accuracy was not high. In addition, these studies utilized datasets from the Internet that did not explicitly identify classes of text [4-7].

Increasing the accuracy of text classification is a challenging issue for researchers. The accuracy of text classification and sentiment analysis depends on the selected term weighting technique, dataset, and ML technique [8-9]. An effective technique must be able to determine the appropriate weights for the words in a text. A good dataset must be reliable and have a clearly identifiable group of texts. The applied ML techniques should be highly accurate and reliable for text classification and sentiment analysis.

The aim of the present research was to introduce a new term frequency with a Gaussian technique (TF-G) and to compare its text classification accuracy with the accuracy of bag-of-words (BoW) [10], term frequency (TF), term frequency-inverse document frequency (TF-IDF) [2], and term frequency-inverse corpus document frequency (TF-ICF) [11]. The ML techniques compared in this study were decision tree (DT) [12], naïve Bayes (NB) [13], support vector machine (SVM) [14], random forest (RF) [15], and multilayer perceptron (MLP) [16-17] techniques to enhance the accuracy of text classification. The datasets in this study were derived from clinical notes in Thai from the outpatient department (OPD) of Khon Kaen Rajanagarindra Psychiatric Hospital, from Thai text of customer reviews for Burger King, Pizza Hut, and Sizzler restaurants [18], and from English text in the tweets of travelers that use US airline services [19].

2 Theoretical Background and Related Works

This section describes the theories and techniques applied for text classification, including Thai word segmentation, term weighting techniques, and feature extraction, as well as a literature review of works related to Thai and English text classification and sentiment analysis.

2.1 Thai Word Segmentation

The Thai language structure is different from that of the English language. The Thai language does not use spaces to separate words and full stops are not used to indicate the end of a sentence. Thus, word segmentation is an important step in the classification of Thai text.

2.1.1 Longest Matching

The longest matching method was used in the majority of early studies of Thai word segmentation. This approach involves scanning a sentence from left to right for the longest match with a dictionary entry at each point. If the algorithm cannot discover the rest of the words in a phrase using the selected match, the program will backtrack to locate the next-longest match and to continue finding more matches. Because of its greedy nature, this algorithm will fail to obtain the correct segmentation in many cases. For example, the text 'ไปหามเหสี' ('go see the queen') is segmented as 'ไป/หาม/เห/สี' ('go/carry/deviate/color') with the longest matching technique; this is incorrect. The correct result is 'ไป/หา/มเหสี' ('go/see/queen'), which the algorithm cannot obtain. The longest matching technique is approximately 90% accurate for Thai text [20].

2.1.2 Maximum Matching

To overcome the limitations of the longest matching algorithm, the maximum matching algorithm was developed. This algorithm generates all feasible segmentations for a sentence before selecting the sentence with the fewest words, which can be quickly achieved using dynamic programming. This technique outperforms the longest matching approach because it can obtain real maximum matches rather than guessing using local greedy heuristics. For example, in the text 'ไปหามเหสี' ('go see the queen'), when using the maximum matching technique, the word segmentations are (1) 'ไป/หาม/เห/สี' ('go/carry/deviate/color') and (2) 'ไป/หา/มเหสี' ('go/see/queen'). This method selects the segmented text result with the smallest number of words, which is (2). In the case that the results of all segment types have the same number of words, the result of the longest matching technique is selected. The maximum matching technique is 99% accurate for Thai text [21].

2.2 Term Weighting

The term weighting technique refers to a well-known preprocessing step in text classification that assigns appropriate weights to each term in documents with a feature vector structure to enhance the accuracy of text classification. This section describes various term weighting techniques, including the bag-of-words (BoW), term frequency (TF), term frequency-inverse document frequency (TF-IDF), and term frequency-inverse corpus frequency (TF-ICF) techniques for text classification.

2.2.1 Bag of Words (BoW)

BoW is a technique for extracting features from text [10]. This technique is simple to understand and implement. For example, for the text 'I am a mental patient, but I never commit suicide', the total number of words is 10; the frequency of 'I' is 2; the frequency of 'am' is 1; the frequency of 'a' is 1; the frequency of 'mental' is 1; the frequency of 'patient' is 1; the frequency of 'but' is 1; the frequency of 'never' is 1, the frequency of 'commit' is 1 and the frequency of 'suicide' is 1.

2.2.2 Term Frequency-Inverse Document Frequency (TF-IDF)

The TF, tf(t,d), is the number of words that appear in a document divided by the total number of words in a document and can be calculated with Eq. (1). The IDF, idf(t,D), is a calculation of the importance of a term across a collection of documents and can be calculated with Eq. (2). TF-IDF, tfidf(t,d,D), is a technique for assessing the importance of individual terms in a document by calculating the weight of each term. This approach is commonly applied in information retrieval and text mining and is based on Eq. (3) [2].

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{1}$$

$$\mathrm{idf}(t,D) = \log(\frac{N}{|\{d \in D, t \in d\}|}) \tag{2}$$

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) * \mathrm{idf}(t,d) \tag{3}$$

where $f_{t,d}$ is the frequency of term (t) in document (d), N is the total number of documents in the corpus $N = |D|$, and $\mathrm{tf}(t,d) \neq 0$. If the term is not in the corpus, a division-by-zero will occur. It is therefore common to adjust the denominator to $1 + |\{d \in D, t \in d\}|$.

2.2.3 Term Frequency-Inverse Corpus Frequency (TF-ICF)

The TF-ICF, tficf(t,C), is the inverse of the frequency of documents in each group of terms being considered and the total number of documents and can be calculated with Eq. (4) [11].

$$\mathrm{tficf}(t,C) = \log\left(f_{t,d} + 1\right) * \log(\frac{N_c}{df_{t,c}}) \tag{4}$$

where $f_{t,d}$ is the frequency of term (t) in document (d), N_c is the total number of documents in each class $c = |C|$, and $df_{t,c}$ is the number of documents in which term (t) appears in class (c).

2.3 Feature Extraction

Feature extraction has become an important part of natural language processing tasks [22]. Feature extraction extracts attributes from data as numbers or determines the weights of terms so that they can be inserted into a model (as a feature vector). The numerical value or weight of each term is obtained from term weighting techniques.

2.4 Text Classification and Sentiment Analysis Works

Some related works involving Thai and English text classification and sentiment analysis that use term weighting are shown in Table 1. Table 1 shows related work on Thai text classification and sentiment analysis. For example, Ref. [7] considered major depressive disorder (MDD) risk classification in 2019. This work applied the BoW technique for term weighting and ML techniques, including SVM and NB techniques, to classify the risk of MDD based on Thai texts on Facebook, with a total of 1,500 posts. The SVM technique was the most accurate, with an F-measure of 94%.

Table 1 Related works on Thai and English text classification (sorted from lowest to highest accuracy based on the F-measure).

Reference / Year	Technique		Dataset	Number of texts	F-measure, by technique
	Term weighting	Machine learning			
[4]/2018	TF-IDF applied to PoS	DT, NB and SVM	Customer reviews from three beauty websites (Thai)	2,770	70.90%, TF-IDF with bigram words
[5]/2014	BoW, TF and TF-IDF	SVM, NB, DT and *k*-NN	Social network posts (Thai)	1,800	77.86%, BoW with SVM
[18]/2017	PoS	DNN	Customer reviews from three restaurants (Thai)	644	87.34%, PoS with DNN
[6]/2010	TF-IDF and SVD	DT, SVM, NB and *k*-NN	Internet (Thai)	200	90.00%, TF-IDF and SVD with NB
[7]/2019	BoW	SVM and NB	Facebook posts (Thai)	1,500	94.00%, BoW with SVM
[23]/2015	PoS	NB, RF, and SVM	Product reviews collected from Amazon.com (English)	2,000,000	94% with SVM
[24]/2014	VADER, Hu-Liu04, LIWC, GI, ANEW, SWN, SCN, and WSD	NB,ME, and SVM	Tweets, movie reviews, product reviews, opinion news articles (English)	4,200, 10,605, 3,708, and 5,190	96% with VADER and SVM

Note: Valence Aware Dictionary for sEntiment Reasoning (VADER), Linguistic Inquiry Word Count (LIWC), General Inquirer (GI), Affective Norms for English Words (ANEW), SentiWordNet (SWN), SenticNet (SCN), Word-Sense Disambiguation (WSD), and Maximum Entropy (ME) [24].

In this research, we used the following techniques: Thai word segmentation with the maximum matching technique and term weighting with BoW, TF, TF-IDF, and TF-ICF. The latter methods were compared with TF-G. DT, NB, SVM, RF, and MLP models are techniques for Thai and English text classification and sentiment analysis.

3 Methodology

This section describes the datasets and conceptual framework.

3.1 Datasets

Three different datasets were utilized in this research.

(1) The first dataset contained 2,987 data samples, which were collected from clinical notes in Thai on mental illness history from the OPD of Khon Kaen Rajanagarindra Psychiatric Hospital between January 2016 and January 2019. The data were obtained from 1,552 clinical records in which nurses and psychiatrists indicated suicide risk and 1,435 clinical records in which nurses and psychiatrists indicated no suicide risk for 238 patients. The average length of the clinical notes was 105 words; the shortest note contained 2 words; the longest note contained 204 words.

(2) The second dataset contained data that were downloaded from Thai customer reviews of Burger King, Pizza Hut, and Sizzler restaurants, with a total of 644 posts, including 361 positive posts and 283 negative posts. The average length for the corpus was 230 words; the shortest post contained 2 words; the longest post contained 458 words [18].

(3) The third dataset contained data that were downloaded from the tweets of users of US airline services in February 2015 who expressed their feelings on Twitter (license: CC BY-NC-SA 4.0), with a total of 14,640 tweets, including 3,090 neutral tweets, 2,363 positive tweets and 9,187 negative tweets. The average length for the corpus was 21 words; the shortest tweet contained 2 words; the longest tweet contained 40 words [19].

3.2 Conceptual Framework

The conceptual framework (Figure 1) includes data preparation and cleaning, text preprocessing, feature extraction, text classification (suicide risk and sentiment), and a comparison of techniques used for text classification from clinical notes, customer reviews, and tweet sentiments of travelers using US airline services.

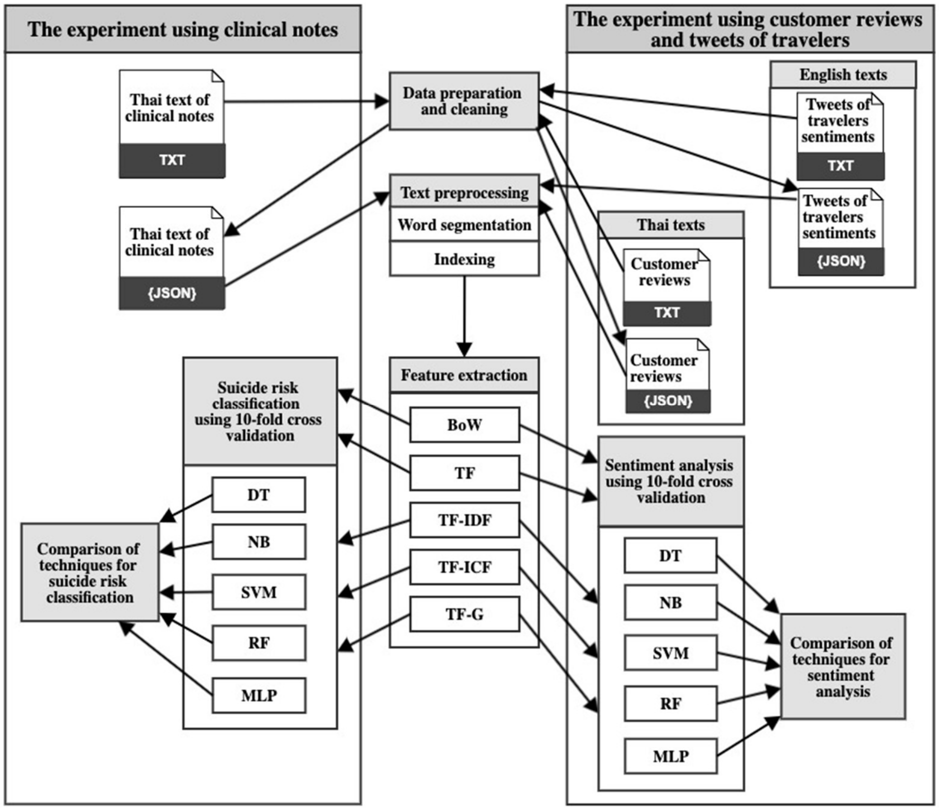
The experiment using clinical notes
Thai text of clinical notes
TXT
Thai text of clinical notes
{JSON}
Data preparation and cleaning
Text preprocessing
Word segmentation
Indexing
Feature extraction
BoW
TF
TF-IDF
TF-ICF
TF-G
Suicide risk classification using 10-fold cross validation
DT
NB
SVM
RF
MLP
Comparison of techniques for suicide risk classification
The experiment using customer reviews and tweets of travelers
English texts
Tweets of travelers sentiments
TXT
Tweets of travelers sentiments
{JSON}
Thai texts
Customer reviews
TXT
Customer reviews
{JSON}
Sentiment analysis using 10-fold cross validation
DT
NB
SVM
RF
MLP
Comparison of techniques for sentiment analysis

Figure 1 Conceptual framework of this research.

3.2.1 Data Preparation and Cleaning

This step was conducted to handle text in datasets before text preprocessing. The step can be divided into two substeps: data preparation and cleaning for the Thai language and data preparation and cleaning for the English language. The proposed method consists of the following processes:

3.2.1.1 Data Preparation and Cleaning in the Thai Language

Data preparation and cleaning for the Thai language involved the following steps. (1) Typos and misspelled words were manually corrected. For example, the misspelled words ‘หงิดหงุด’ and ‘โมห’ are meaningless. The correct words are ‘หงุดหงิด’, which means ‘irritable’, and ‘โมโห’, which means ‘angry’. (2) Abbreviations and medical vocabularies were changed to Thai text in clinical notes. For example, the text ‘admit’ was changed to ‘ให้พักรักษาในโรงพยาบาล’;

the text ‘MT’ was changed to ‘จิตบำบัด’ (‘MT’ stands for ‘mental treatment’); and the text ‘SE’ was changed to ‘ผลข้างเคียง’ (‘SE’ stands for ‘side effect’). (3) Special characters and numbers, such as ‘+’, ‘-’, ‘#’, and ‘@’, were removed from the datasets.

3.2.1.2 Data Preparation and Cleaning in the English Language

Data preparation and cleaning for the English language involved the following steps. (1) Typographical errors were manually corrected. For example, the misspelled words ‘wlere’ and ‘dslike’ were changed to ‘where’ and ‘dislike’. (2) Groups of characters that create emoji symbols, such as ‘úà’, ‘úîÔ∏è’, ‘úÖ’, and ‘≠êÔ∏è’, were removed. (3) Special characters and numbers, such as ‘+’, ‘-’, ‘#’, and ‘@’, were removed from the datasets.

3.2.2 Text Preprocessing

Text preprocessing consists of word segmentation and the identification of stopping words and stemming words.

3.2.2.1 Word Segmentation

This process is divided into two types of word segmentation because this study used Thai and English datasets. (1) The clinical notes and customer reviews are Thai language datasets, so word segmentation using the maximum matching technique was performed. (2) The travelers’ tweets about US airline services was an English language dataset, so word segmentation considered spaces in text.

3.2.2.2 Indexing

Indexing involves establishing a collection of unique words in a document system. This step creates index words by selecting the unique words from every text that has undergone text preprocessing.

3.2.3 TF-G Technique

The TF-G, tfg(t,G), is based on the idea that the distribution of words in each document category is related to the document category. Therefore, we propose a new technique for term weighting by applying Gaussian functions to calculate the distribution of words in each category, as shown in Eq. (5) and Eq. (6).

$$G = \frac{n_c+1}{\sqrt{\pi 2\left(\sigma_{i,c}+1\right)^2}} e^{-\left[\frac{\left(w_{i,j}-\mu_{i,c}\right)^2}{2\left(\sigma_{i,c}+1\right)^2}\right]} \tag{5}$$

$$\mathrm{tfg}(t, G) = \log\left(f_{t,d} + 1\right) * G \tag{6}$$

where G is the distribution value of term i in document j of category c; n_c is the number of documents in category c; $w_{i,j}$ is the frequency of term i in document j; $\mu_{i,c}$ is the mean of all terms i that are within a document in category c; $\sigma_{i,c}$ is the variance in term i within a document in category c; c is the category of a document; and $f_{t,d}$ is the frequency of term (t) in document (d).

For example, we consider a set of seven documents. The category is identified from the documents (where 1 is positive and 0 is negative), as shown in Table 2.

Table 2 Document sample.

Document No.	Texts	Label
1	‘อยากตาย เบื่อ ’(‘am bored and want to die’)	1
2	‘เบื่อไม่อยากกินยา ’(‘am bored and do not want to take medicine’)	1
3	‘ไม่อยากตาย เบื่อ ’(‘do not want to sleep and am bored’)	0
4	‘ไม่นอน เบื่อยา ’(‘do not want to sleep and am bored with medication’)	0
5	‘เบื่อ หงุดหงิด ’(‘am bored, frustrated’)	1
6	‘หงุดหงิด+อยากตาย+ไม่อยากนอน ’(‘am frustrated +want to die +do not want to sleep’)	1
7	‘นอนกินยา ’(‘take medicine while sleeping ’)	0

The documents in Table 2 are processed via the data preparation, data cleaning, and text preprocessing steps. The results are shown in Table 3.

Table 3 Results of text preprocessing.

Document No.	Texts	Label
1	อยาก/ตาย/เบื่อ	1
2	เบื่อ/ไม่/อยาก/กิน/ยา	1
3	ไม่/อยาก/ตาย/เบื่อ	0
4	ไม่/นอน/เบื่อ/ยา	0
5	เบื่อ/หงุดหงิด	1
6	หงุดหงิด/อยาก/ตาย/ไม่/อยาก/นอน	1
7	นอน/กิน/ยา	0

The next step is to create index terms from the results in Table 3. The index terms are ‘อยาก’, ‘ตาย’, ‘เบื่อ’, ‘ไม่’, ‘กิน’, ‘ยา’, ‘นอน’, and ‘หงุดหงิด’. The weights of the index terms are shown in Table 4. These weights are calculated with the BoW technique.

Table 4 Weights of the index terms obtained from the BoW technique.

Document No.	$w_{i,j}$ (Weights of index terms obtained using the BoW technique)								Label
	อยาก	ตาย	เบื่อ	ไม่	กิน	ยา	นอน	หงุดหงิด	
1	1	1	1	0	0	0	0	0	1
2	1	0	1	1	1	1	0	0	1
3	1	1	0	1	0	0	0	0	0
4	0	0	1	1	0	1	1	0	0
5	0	0	1	0	0	0	0	1	1
6	2	1	0	1	0	0	1	1	1
7	0	0	0	0	1	1	1	0	0

From Table 4, the next step is to calculate the mean and variance of index terms in the documents in each category. The results are shown in Table 5.

Table 5 Mean and variance of index terms in the documents in each category.

Mean/Variance	Index terms								Label
	อยาก	ตาย	เบื่อ	ไม่	กิน	ยา	นอน	หงุดหงิด	
$\mu_{i,P}$	1.00	0.50	0.75	0.50	0.25	0.25	0.25	0.50	1
$\mu_{i,N}$	0.33	0.33	0.33	0.67	0.33	0.67	0.67	0.00	0
$\sigma_{i,P}$	0.58	0.48	0.52	0.48	0.47	0.52	0.52	0.25	1
$\sigma_{i,N}$	1.07	0.51	0.51	0.51	0.19	0.38	0.51	0.58	0

The final step is to calculate the distribution value of index terms in the documents in each category. These values can be calculated by Eq. (5) and Eq. (6) with the values in Tables 4 and 5, respectively. The results are shown in Table 6.

Table 6 Distribution values of the index terms in the documents in each category.

Document No.	The weights of index terms obtained using the TF-G technique								Label
	อยาก	ตาย	เบื่อ	ปฏิเสธ	กิน	ยา	นอน	หงุดหงิด	
1	0.38	0.38	0.39	0.00	0.00	0.00	0.00	0.00	1
2	0.38	0.00	0.39	0.38	0.36	0.35	0.00	0.00	1
3	0.22	0.29	0.00	0.31	0.00	0.00	0.00	0.00	0
4	0.00	0.00	0.29	0.31	0.00	0.34	0.31	0.00	0
5	0.00	0.00	0.39	0.00	0.00	0.00	0.00	0.44	1
6	0.49	0.38	0.00	0.38	0.00	0.00	0.35	0.44	1
7	0.00	0.00	0.00	0.00	0.34	0.34	0.31	0.00	0

3.2.4 Suicide Risk Classification and Sentiment Analysis

This procedure uses a feature vector, and each weighted technique classified suicide risk from clinical notes, customer reviews for sentiment analysis, and traveler tweets regarding US airline services for sentiment analysis. Suicide risk and sentiment analysis was done with the DT, NB, SVM, RF, and MLP techniques.

These experiments were conducted in Python (version 2.7.16). The model evaluation method for all datasets was 10-fold cross-validation. The advantage of 10-fold cross-validation is that all feature vectors are used for both training and testing, and each feature vector is utilized for validation only once.

3.2.5 Comparison of Techniques for Thai Text Classification, Thai Text Sentiment Analysis, and English Text Sentiment Analysis

The comparison steps were: (1) finding the most accurate technique for suicide risk classification from Thai clinical notes, (2) finding the most accurate technique for sentiment analysis from Thai customer reviews, and (3) finding the most accurate technique for sentiment analysis from the tweets of travelers using US airline services.

4 Results

This section presents the experimental results of the risk of suicide classification from the clinical notes dataset (Table 7), sentiment analysis from the customer reviews dataset (Table 8), and sentiment analysis from the tweet dataset for travelers using US airline services (Table 9).

In Table 7, the TF-G technique was the most accurate technique in Thai suicide risk classification, with higher F-measure scores than other term weighting techniques based on 10-fold cross-validation. The TF-G technique with the RF technique was the most accurate technique in suicide risk classification, with precision, recall, and F-measure scores of 97.83%, 96.88%, and 97.51%, respectively. The results of the experiment for text sentiment analysis are shown in Table 8. The TF-G technique was the most accurate technique in Thai sentiment analysis, with higher F-measure scores than the other term weighting techniques based on 10-fold cross-validation. The TF-G technique with the RF technique was the most accurate in suicide risk classification, with precision, recall, and F-measure scores of 92.5%, 92.18%, and 92.21%, respectively.

Table 7 Results of suicide risk classification from clinical notes.

Machine learning	Term weighting	Precision	Recall	F-measure
DT	BoW	78.06	76.14	75.30
	TF	77.81	75.74	74.82
	TF-IDF	77.81	75.74	74.82
	TF-ICF	91.82	90.03	89.86
	TF-G	96.45	96.39	**96.23**
NB	BoW	75.78	75.25	74.98
	TF	71.95	70.56	70.31
	TF-IDF	72.30	71.01	70.78
	TF-ICF	66.79	66.51	66.19
	TF-G	76.77	75.89	**76.34**
SVM	BoW	77.70	76.78	76.26
	TF	77.53	76.61	76.10
	TF-IDF	75.70	74.95	74.46
	TF-ICF	72.91	72.59	72.35
	TF-G	78.22	77.58	**77.18**
RF	BoW	78.69	77.49	76.84
	TF	78.56	77.67	77.14
	TF-IDF	78.56	77.67	77.14
	TF-ICF	95.75	95.76	95.68
	TF-G	97.83	96.88	**97.51**
MLP	BoW	71.07	69.65	70.33
	TF	72.15	72.74	72.41
	TF-IDF	70.60	71.78	71.14
	TF-ICF	70.26	71.40	72.68
	TF-G	72.24	73.02	**73.21**

As the experimental results show in Tables 7 and 8, the TF-G technique for Thai text classification and Thai sentiment analysis was more accurate than the results of the BoW, TF, TF-IDF, and TF-ICF techniques with the DT, NB, SVM, RF, and MLP techniques.

In Table 9, the TF-G technique was the most accurate in English sentiment analysis, with higher F-measure scores than other term weighting techniques. The TF-G technique with the SVM technique was the most accurate technique in the sentiment analysis of travelers using US airline services, with precision, recall, and F-measure scores of 99.84%, 99.73%, and 99.78%, respectively, based on 10-fold cross-validation.

Based on the experimental results shown in Tables 7, 8, and 9, the results of the TF-G technique for text classification and sentiment analysis were more accurate than the results of the BoW, TF, TF-IDF, and TF-ICF techniques with the DT, NB, SVM, RF, and MLP techniques for Thai and English datasets.

Table 8 Results of the sentiment analysis from customer reviews.

Machine learning	Term weighting	Precision	Recall	F-measure
DT	BoW	79.20	77.82	77.84
	TF	76.69	76.31	76.15
	TF-IDF	76.69	76.31	76.15
	TF-ICF	87.53	86.34	85.25
	TF-G	90.90	91.22	**90.60**
NB	BoW	87.77	87.01	87.07
	TF	85.03	73.25	72.65
	TF-IDF	85.80	77.07	77.09
	TF-ICF	74.39	73.89	73.94
	TF-G	88.28	88.92	**88.80**
SVM	BoW	86.89	86.68	86.43
	TF	88.72	87.63	87.82
	TF-IDF	86.56	85.58	85.76
	TF-ICF	78.93	78.62	78.37
	TF-G	89.49	88.87	**88.86**
RF	BoW	85.99	78.76	78.90
	TF	86.08	78.82	79.21
	TF-IDF	86.08	78.82	79.21
	TF-ICF	85.41	74.86	74.26
	TF-G	92.50	92.18	**92.21**
MLP	BoW	79.76	79.73	89.60
	TF	93.64	81.17	86.64
	TF-IDF	92.96	74.43	82.37

	TF-ICF	83.01	83.28	82.83
	TF-G	90.23	90.31	**90.11**

Table 9 Results of the sentiment analysis from the tweets of travelers using the US airline services sentiment dataset [19].

Machine learning	Term weighting	Precision	Recall	F-measure
DT	BoW	72.38	64.06	64.01
	TF	72.02	60.09	61.34
	TF-IDF	73.15	61.43	63.62
	TF-ICF	86.14	69.21	70.03
	TF-G	80.99	81.75	**79.93**
NB	BoW	78.53	71.74	72.91
	TF	82.14	53.30	65.16
	TF-IDF	84.75	63.1	69.67
	TF-ICF	84.15	80.62	81.87
	TF-G	81.79	82.60	**82.13**
SVM	BoW	83.89	81.91	82.68
	TF	81.66	77.62	78.81
	TF-IDF	81.41	78.56	79.24
	TF-ICF	92.27	89.15	90.34
	TF-G	99.84	99.73	**99.78**
RF	BoW	68.32	63.69	65.93
	TF	62.40	65.32	63.85
	TF-IDF	64.45	65.89	65.20
	TF-ICF	87.82	90.32	89.12
	TF-G	88.35	93.12	**90.67**
MLP	BoW	68.41	67.34	67.84
	TF	70.88	69.26	70.02
	TF-IDF	67.53	68.44	67.96
	TF-ICF	77.94	73.38	75.5
	TF-G	81.58	75.79	**79.08**

5 Conclusion

In this paper, a new TF-G technique was proposed for creating feature vectors with well-known and effective ML techniques to enhance the accuracy of classifying text. Cases involving classification from Thai clinical notes, a sentiment analysis of Thai customer reviews [18], and a sentiment analysis of English tweets from travelers using US airline services [19] were investigated.

The datasets had different characteristics: the Thai clinical notes dataset was composed of patient data from the OPD of Khon Kaen Rajanagarindra Psychiatric Hospital; the Thai customer reviews conveyed the sentiments of customers after using the services at Burger King, Pizza Hut, and Sizzler restaurants; and the English tweets conveyed the sentiments of travelers after using US airline services. Although these datasets contained some typographical errors, they were reliable; notably, the clinical notes were patient health records, and the risk of suicide was diagnosed during each visit by nurses and psychiatrists. Additionally, customer reviews were clearly identified with positive and negative sentiments, and the tweets of traveler sentiments are clearly identified with neutral, positive, and negative connotations.

The proposed technique is a term weighting technique for creating feature vectors. This research compares the BoW, TF, TF-IDF, and TF-ICF techniques with TF-G in the extraction of features from Thai text and English text. Suicide risk classification and sentiment analysis were performed with the DT, NB, SVM, RF, and MLP techniques. The experimental results showed that TF-G with the RF technique was the most effective for suicide risk classification. The TF-G with the RF technique was the most accurate for positive and negative sentiment analysis. The TF-G with the SVM technique was the most accurate for neutral, positive, and negative sentiment analysis.

The TF-G method performed significantly better than the traditional term weighting techniques for all three datasets. Therefore, the TF-G technique is appropriate for Thai text classification, Thai text sentiment analysis, and English text sentiment analysis.

A limitation of this research was that we were not able to find a dataset from a previous research to compare with the results of our model to evaluate efficiency. Therefore, this research used clinical notes collected from an OPD, customer reviews downloaded from the Internet, and tweets from travelers downloaded from the Internet to conduct the experiments. In the future, we will use TF-G techniques to implement a new module to classify the risk of suicide from clinical notes. This module could be integrated with a hospital system to assist psychiatrists and nurses in their diagnoses.

Acknowledgment

We thank the director and staff of Khon Kaen Rajanagarindra Psychiatric Hospital for their help in collecting the data and for their valuable technical support in this research.

References

[1] Salton, G. & Buckley, C., *Term-weighting Approaches in Automatic Text Retrieval,* Inf., Process., Manage., **24**(4), pp. 513-523, 1988.

[2] Salton, G. & McGill, M., *Introduction to Modern Information Retrieval*, McGraw-Hill Book Company, New York, NY, 1983.

[3] Alsmadi, I. & Hoon, G.K., *Term Weighting Scheme for Short-text Classification: Twitter Corpuses*, Neural Computing and Applications, **31**(8), pp. 3819-3831, 2019.

[4] Inrak, P. & Sinthupinyo, S., *Applying Latent Semantic Analysis to Classify Emotions in Thai Text,* in 2010 2nd International Conference on Computer Engineering and Technology*,* IEEE, **6**, pp. V6-450, 2010.

[5] Chirawichitchai, N., *Emotion Classification of Thai Text Based Using Term Weighting and Machine Learning Techniques,* 2014 11th International Joint Conference on Computer Science and Software Engineering JCSSE, IEEE, pp. 91-96, May. 2014.

[6] Charoensuk, J. & Sornil, O., *A Hierarchical Emotion Classification Technique for Thai Reviews*, Journal of ICT Research and Applications, **12**(3), pp. 280-296, 2018.

[7] Hemtanon, S. & Kittiphattanabawon, N., *An Automatic Screening for Major Depressive Disorder from Social Media in Thailand*, 2019 10th National and International Research Conference and Presentation, **1**(10), pp. 103-113, 2019.

[8] Mazyad, A., Teytaud, F. & Fonlupt, C., *A Comparative Study on Term Weighting Schemes for Text Classification*, International Workshop on Machine Learning, Optimization, and Big Data, Springer, Cham., pp. 100-108, September. 2017.

[9] Mazyad, A., Teytaud, F. & Fonlupt, C., *Generating Term Weighting Schemes through Genetic Programming*, International Conference on Machine Learning, Optimization, and Data Science, Springer, Cham., pp. 92-103, September. 2018.

[10] McTear, M.F., Callejas, Z. & Griol, D., *The Conversational Interface*, Cham: Springer, **6**(94), pp. 102, 2016.

[11] Reed, J.W., Jiao, Y., Potok, T.E., Klump, B.A., Elmore, M.T. & Hurson, A.R., *TF-ICF: A New Term Weighting Scheme for Clustering Dynamic Data Streams*, 2006 5th International Conference on Machine Learning and Applications ICMLA'06, IEEE, pp. 258-263, 2006.

[12] Han, J., Kamber, M. & Pei, J., *Mining Frequent Patterns, Associations, and Correlations.,* Data Mining: Concepts and Techniques, pp. 227-283, 2006.

[13] Chen, J., Huang, H., Tian, S. & Qu, Y., *Feature Selection for Text Classification with Naïve Bayes*, Expert Systems with Applications, **36**(3), pp. 5432-5435, 2009.

[14] Suykens, J.A. & Vandewalle, J., *Least Squares Support Vector Machine Classifiers*, Neural Processing Letters, **9**(3), pp. 293-300, 1999.

[15] Liaw, A. & Wiener, M., *Classification and Regression by Randomforest*, R news, **2**(3), pp. 18-22, 2002.

[16] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. & Gao, J., *Deep Learning-based Text Classification: A Comprehensive Review*, arXiv preprint arXiv:2004.03705, 2020.

[17] Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N. & Kingsbury, B., *Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups*, IEEE Signal Processing Magazine **29**(6), pp. 82-97, 2012.

[18] JagerV3, sentiment_analysis_thai/corpus, https://github.com/JagerV3/sentiment_analysis_thai, 2017. (22 June 2021)

[19] Twitter US Airline Sentiment, *Analyze How Travelers in February 2015 Expressed Their Feelings on Twitter*, https://www.kaggle.com/crowdflower/twitter-airline sentiment?select=Tweets.csv, 2018. (23 June 2021)

[20] Poovorawan, Y. & Imarom, V., *Dictionary-based Thai Syllable Segmentation*, 9th Electrical Engineering Conference, 1986.

[21] Sornlertlamvanich, V., *Word Segmentation for Thai in Machine Translation System*, Machine Translation, NECTEC, pp. 556-561, 1993.

[22] Sammons, M., Christodoulopoulos, C., Kordjamshidi, P., Khashabi, D., Srikumar, V. & Roth, D., *Edison: Feature Extraction for NLP, Simplified*, Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 4085-4092, 2016.

[23] Fang, X. & Zhan, J., *Sentiment Analysis Using Product Review Data*, Journal of Big Data, **2**(1), pp. 1-14, 2015.

[24] Hutto, C. & Gilbert, E., *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of social media Text*, 2014 International AAAI Conference on Web and Social Media, **8**(1), May. 2014.