Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 3, June 2026, pp. 827∼834
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i3.pp827-834

On exploring text mining approaches to sentiment analysis based on the combination of word-based and ontology-based approaches

Suthira Plansangket, Supaporn Kansomkeat, Supasit Kajkamhaeng
Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand

Article Info

Article history:

Received Jul 28, 2025
Revised Mar 21, 2026
Accepted May 26, 2026

Keywords:

CSDF
ontoCSDF
Sentiment analysis
Text mining
TF-IDF

ABSTRACT

Currently, sentiment analysis plays an important role in business. Entrepreneurs try to understand customer needs for products and services. If they know about the needs, they can create the marketing plans or strategy plans in their business that help improve products and services. Therefore, this study explores two novel approaches to improve the classification accuracy of sentiment analysis data using a combination of a word-based approach (TF-IDF or CSDF) and an ontology-based approach (ontoSen) to provide two new methods, called ontoTF-IDF and ontoCSDF. The experimental results show that CSDF method had the best classification accuracy among all the methods in this study: ontoCSDF did not improve further the classification accuracy of sentiment analysis data. Furthermore, ontoTFIDF method improved the classification by IBk algorithm significantly ($p<0.05$).

Corresponding Author:

Suthira Plansangket
Division of Computational Science, Faculty of Science, Prince of Songkla University
Songkhla, Thailand
Email: suthira.p@psu.ac.th

1. INTRODUCTION

Sentiment analysis plays an important role in business. To understand customer needs for products and services is concerned by the entrepreneurs. For example, how people think about a particular movie or a hotel, etc. If a business is aware of the needs, it can create marketing or strategy plans to improve products and services. In addition, if we can improve sentiment analysis for feedback of feelings by the users, then the products and services can be improved, giving a competitive advantage to the business.

This study aims to explore two novel methods to improve the classification accuracy of sentiment analysis data, using combinations of word-based and ontology-based approaches. Normally, word-based approaches use a bag-of-words representation with statistics. This ignores both order and meaning of the words. On the other hand, an ontology-based approach focuses on the order, semantics, and relationships of words. Therefore, this study aimed to explore the combination these two approaches for an improved sentiment analysis by text mining.

2. LITERATURE REVIEW

Affective computing is study for developing systems and devices that can recognize, process, interpret, and simulate human affects. It is an interdisciplinary field combining computer science, psychology, and

cognitive science [1]. One of the motivations for such research is the ability to give machines emotional intelligence, including simulated empathy. The machine should interpret the emotional states of humans and adapt its behaviour to them, giving appropriate responses to those emotions. To detect the emotions, we start from the sensors that collect data on the status or physical behaviour of humans, without interpretation. The data from the sensors are comparable to the data from human senses reflecting other human's emotions. For example, a camera can detect facial expressions, posture, and physical movements; and a microphone can detect speech. Furthermore, there are other sensors that can directly collect physiological data, such as body temperature, brain waves, and electrical conductivity of the skin. In addition, emotional recognition needs to extract meaningful patterns from the collected signals, which can be learned by various machine learning algorithms, such as those for speech recognition, face detection, and natural language processing (NLP). After that, these can be interpreted as emotions.

Sentiment analysis is a subset of affective computing. It is about NLP that learns from the sentences in documents, and then translates to express feelings, thinking, or attitudes. Sentiment analysis is separated into two major types: positive and negative [2]. Some research also considers a type called neutral [3]. Sentiment analysis consists of two major techniques. First, machine learning or data mining is used to find patterns and relationships inside a large dataset for sentiment analysis, by using mathematical, statistical, and pattern recognition approaches [3]-[5]. Second, the types of words in a document fall into two categories: positive words and negative words. After that, each word can be given a score and summed to a total score. If the positive words dominate over the negative words, then the sum is positive and we can conclude that this document is positive [6].

Natural language processing plays an important role in sentiment analysis. Benamara *et al.* [7] proposed using a combination of adjectives and adverbs in sentiment analysis. The results show that this combination helps improve the performance of sentiment analysis using pearson correlation. Furthermore, Kouloumpis *et al.* [5] studied the parts of speech; nouns, verbs, adjectives, and adverbs in sentiment analysis by data mining. The data representation in this research used a bag of words [8], which accounts for each word separately without considering their order. The method of data representation by term frequency–inverse document frequency (TF-IDF) [9] is well-know and a standard method. However, Plansangket and Gan [10] have proposed a novel data representation called the class specific document frequency (CSDF). It is a data representation method which focuses on the important words in text mining. These are the words that often appear in documents of the same class, but rarely appear in documents of a different class. The experimental results show that CSDF gives a better classification accuracy than the TF-IDF.

TF-IDF and CSDF are a bag of words model. Each word is independently accounted without reference to word order. However, the order of words is important in meaningful sentences. A different order of words gives a semantic difference. For example, " the rabbit ran faster than the turtle " and " the turtle ran faster than the rabbit " are totally different semantically, although made up of the same words. Therefore, some studies have used ontology in text mining for information retrieval [11] and sentiment analysis [12].

Ontology is a study of object characteristics, regarding what is this object, how many types of this object are there, and structure, properties, events, processes of this object, and relations between objects in the real world. In addition, ontology is the reliable reality in the real world [13]. Furthermore, ontology is the principal foundation of knowledge representation that collects all knowledge, semantics, and relationships from different sources [14]. Ontology separates things into different types; therefore, it relates to text mining and sentiment analysis.

Nowadays, extensive research focuses on integrating ontology with machine learning and deep learning techniques. For example, Sharma and Kumar [15] proposed a new hybrid semantic indexing approach for unstructured text documents was introduced by integrating machine learning with domain ontology. The method enhanced its capability to identify concepts that are semantically related to document content through the use of a machine learning–based skip-gram model. Experimental results demonstrated that the proposed method outperformed state-of-the-art techniques on the evaluated datasets, achieving an average accuracy improvement of 29%. Moreover, Jain *et al.* [16] presented a new ontology-based natural language processing techniques that combine feature extraction with deep learning–based classification. A comparative analysis was conducted between the proposed approach and existing techniques, with the proposed method achieving superior results.

Although the use of machine learning and deep learning can yield good results, their complexity and the significant amount of time required for learning led the researchers to choose a simpler and more convenient approach for learning. This study aims to explore two novel approaches to improve the classification accuracy of sentiment analysis data using combined word-based approach, namely TF-IDF or CSDF, and ontology-based approach, namely ontoSen. The two new methods called ontoTF-IDF and ontoCSDF may help to improve the performance of sentiment analysis.

3. METHOD

Methodologies in this research which aims to improve the performance of sentiment analysis were separated into three major approaches.

3.1. Data representation

Data representation using alternatively TF-IDF and CSDF method is used in both training and testing set. First, the TF-IDF scores [17], [18] are calculated as (1):

$$TFIDF_{ji} = \begin{cases} (1 + log_2 TF_{ji}) \times log_2 \frac{N}{DF_i}, & \text{if } TF_{ji} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $TFIDF_{ji}$ is a score for term i in document j, TF_{ji} is the frequency for term i in document j, N is the total documents in the training set, and DF_i is the number of documents that term i appears in, in the training set.

Second, CSDF score [10] of term i in class k is defined as (2):

$$CSDF_{ik} = \begin{cases} \frac{DF_{ik}}{N_k} / \frac{(DF_i - DF_{ik})}{(N - N_k) + 1}, & \text{if } TF_{ik} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where DF_{ik} is the count of term i appearing in the training set and in class k, DF_i is the total number of documents that term i appears in in the training set, N_k is the total number of documents in the training set and in class k, and N is the total number of documents in the training set. However, we do not know the values of DF_{ik} and N_k in documents of the test set. Therefore, CSDF values of term i in both training and testing documents are defined by the variance value of original CSDF for term i in class k, as shown in (3).

$$CSDF_i = var(CSDF_{ik}) \quad (3)$$

3.2. The ontology

The data representation using ontology in this research was applied as in the research of Wójcik and Tuchowski [19]. The ontology or knowledge base is the collection of all facts about the objects in the real world, including relationships between them. The structure of this collection is a hierarchical structure that is ranked one above the other by importance. For example, the root class is thing, with phone as its direct subclass. This class serves as the parent for categories that represent the main features of smartphones. Subsequent levels in the hierarchy describe less significant or more detailed phone characteristics. Each class is assigned a single level based on its position in the hierarchy. The phone class is designated as the first-level class. Its immediate subclasses form the second level, while their descendants correspond to the following hierarchical levels. Therefore, the total sentiment is defined as follows:

$$OntoSen_i = \frac{\sum_{j=1}^{n} \left[\frac{OntoSen_i(j)}{level_i(j)} \right]}{n} \quad (4)$$

where $OntoSen_i$ is the total ontology score of term *i* from each document *j*, *n* is the total number of documents, $OntoSen_i(j)$ is an ontology score of term *i* of document *j*, and $level_i(j)$ is the relationship level of term *i* of document *j*.

3.3. ontoTFIDF and ontoCSDF

ontoTFIDF and ontoCSDF are two novel methods proposed in this research. They are methods that combine word-based approach and ontology-based approach, aiming to calculate combinations of statistical as semantic scores for a word. *ontoTFIDF* and *ontoCSDF* are defined as follows:

$$ontoTFIDF_i = \alpha \times TFIDF_i + (1 - \alpha) \times OntoSen_i \tag{5}$$

$$ontoCSDF_i = \alpha \times CSDF_i + (1 - \alpha) \times OntoSen_i \tag{6}$$

where $ontoTFIDF_i$ and $ontoCSDF_i$ are the sums of $TFIDF_i$ or $CSDF_i$ scores and $ontoSen_i$ score, respectively, and α is a weight used to balance (tune) the scores of two methods from 10% to 90% to find the best combination.

4. EXPERIMENTAL DESIGN

The data set used in this research is sentiment analysis data on opinions regarding Amazon products, collected by Johns Hopkins University's Department of Computer Science [20] . There are four categories in this dataset; books, DVDs, electronics, and kitchenware. In the book category, there are 1,025 documents in training set and 80 documents in the testing set. For DVD categories, there are 1,017 documents in the training set and 106 documents in the testing set. For electronics, there are 4,721 documents in the training set and 1,182 documents in the testing set. Finally, in kitchenware category, there are 4,120 documents in the training set and 1,031 documents in the testing set. Therefore, the overall number of documents in the training set is 10,883 documents and 2,399 documents in the testing set. The split between training and testing sets 80% –20%. Each category in these data is separated into four classes: very satisfying (best), satisfaction (good), lower than expectation (bad), dissatisfaction (worst). There is no neutral class in this study.

The bag of words model is applied in this research. There are 351,672 words in the training set and 100,066 words in the testing set. Data cleansing or data preprocessing is an important process removing all the stop words and special characters, and retaining only necessary data for the analysis. The training set was subjected to preprocessing. Starting with removing all meaningless words, stop words, and special characters from 351,672 words the set came down to about 200,000 words. After that, an NLTK library in Python called nltkstopword was used to remove all stop words; bringing the word count to 45,613 words.

A challenge of text mining is to select the right features for classification [17], [21], [22]. Feature section decreases the size of the data, removes duplicate data, and removes all unrelated words. It helps to easily find the hidden patterns in data for classification. Overall there were 10,883 documents; and there were overall 45,613 words. The overall document count is four times less than the overall feature count, so that feature section [22] is very necessary in this case. Starting with choosing meaningful words both in English dictionary and in ontology. This experiment uses NLTK library in Python called PyEnchant to select words in the English dictionary from 45,613 words this selected 10,403 words. After that, the features are reduced by choosing part of speech. The study [7] found that sentiment analysis related to nouns, verbs, adjectives, and adverbs. This experiment uses a NLTK library in Python namely nltkpos–tag to choose nouns, verbs, adjectives, and adverbs. This step decreased the number of features from 10,403 words to 3,080 words.

Normally, there are two feature selection approaches; filter approach and wrapper approach. However, due to the limited time, this study chose the filter approach ignoring the wrapper approach that would need a lot of time to find the best features. Function CfsSubsetEval [23] in the Weka package [24] was applied to the filter approach in this study. It uses correlation-based feature subset selection [25] algorithm to select the features that relate and effectively might classify the data, and to remove duplicate data. This function returns a ranking of the features from the best features that are associated with the classes, in descending order. Therefore, from the 3,080 features, only the top 400 features were chosen. After that, there are some documents that are not related to these features. Therefore, some documents were removed from this experiment. To summarize, there 4,903 documents remaining from the 10,883 documents in the training set. These are separated into four classes; very satisfaction 1,142 documents, satisfaction 1,171 documents, lower than expected 1,257 documents, and dissatisfaction 1,331 documents. On the other hand, in test data from initial 2,399 documents 1,241 were retained, separated into four classes; very satisfaction 283 documents, satisfaction 315 documents, lower than expected 317 documents, and dissatisfaction 326 documents as shown in Table 1.

The selected 400 retained features were used in this experiment. The data in these files is the frequency data from the original dataset, both in training and testing sets. After that, data representation was applied by calculating the scores for TF-IDF, CSDF, ontoTFIDF, and ontoCSDF, following the equations in methodology section. The ontoTFIDF and ontoCSDF scores were calculated for alpha values from 0.1 to 0.9. The final step is classifying the data, and comparing the classification accuracies of five alternative classification methods; SMO or support vector machine (SVM), Naïve bayes, J48 or decision tree, IBk or k-nearest neighbors (kNN) algorithm, and logistic regression.

Table 1. Statistics of training and testing data

Sentiment	Best	Good	Worse	Bad	Total
Training set	1142	1171	1259	1331	4903
Testing set	326	283	317	315	1241
Total	1468	1454	1576	1646	6144

5. EXPERIMENTAL RESULTS

The experimental results are shown in Table 2, Table 3, and Figure 1. To ensure that our predictive models are correct, these were tested by the training data. The experimental results show that the TF-IDF gave poor classification accuracy already with training data of less than 60%. This shows that the TF-IDF data representation is not appropriate for creating a good predictive model, in this case study. On the other hand, the experimental results with CSDF gave classification accuracy of prediction in training data as 100% with IBk or kNN approaches, 95.38% with J48 or decision tree approach, and over 44% but less than 60% with Naïve bayes, logistic regression, and SMO approaches. Therefore, CSDF data representation may allow an appropriate prediction model by using IBk or J48 methods.

Table 2. Classification accuracies on using TF-IDF, CSDF, ontoTFIDF, and ontoCSDF data

Classification method	Document representation method	Classification accuracy: Prediction with training data	Classification accuracy: Prediction with testing data
Naïve Bayes	TFIDF	41.79	40.53
	CSDF	44.54	43.23
Logistic	TFIDF	50.68	41.5
	CSDF	58.31	51.68
SMO	TFIDF	48.91	41.02
	CSDF	59.93	55.7
IBk	TFIDF	59	37.39
	CSDF	100	96.47
J48	TFIDF	44.1	38.2
	CSDF	95.38	96.56

Table 3. Classification accuracies on using TF-IDF, CSDF, ontoTFIDF, and ontoCSDF data (Continued)

Classification method	Document representation method	Classification accuracy: α values of ontoTFIDF and ontoCSDF								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Naïve Bayes	TFIDF	40.85	40.94	40.94	41.34	40.94	38.76	37.79	35.38	36.1
	CSDF	34.04	39.9	41.43	41.43	40.12	40.77	40.77	41.26	41.26
Logistic	TFIDF	41.82	41.74	41.74	41.1	41.18	41.5	41.42	41.66	41.02
	CSDF	44.46	44.63	44.87	44.71	45.12	45.2	45.37	45.86	48.07
SMO	TFIDF	40.69	41.42	41.1	41.02	40.61	40.77	40.69	40.85	40.94
	CSDF	42.49	42.66	42.74	43	43.15	44.46	47.17	47.74	48.24
IBk	TFIDF	38.03	37.47	37.31	37.71	38.11	38.2	38.03	37.95	37.79
	CSDF	63.5	78.5	80.23	82.61	86.4	87.04	88.27	88.52	90.4
J48	TFIDF	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2	38.2
	CSDF	35.03	37.87	36.26	39.21	39.71	39.62	39.71	40.2	40.28

The experimental results from using testing set gave better classification accuracy for CSDF data than for TF-IDF data with any classification method. For J48 and IBk approaches, CSDF data gave classification accuracy of about 96%. However, the classification accuracies of SMO, Naïve bayes, and logistic regression were below 56% with all data representation methods, including CSDF.

ontoTFIDF combines the scores from ontoSen and TF-IDF, while ontoCSDF combines the scores from ontoSen and CSDF. The combination uses alpha as a weighting factor to balance the two combined scores, with values from 0.1 to 0.9. The experimental results show that the classification accuracy of ontoCSDF data with all alpha values were lower than with CSDF data. Increasing the ontoSen scores degraded classification accuracy below that with CSDF data. Therefore, ontoSen did not improve the performance of classification. On the other hand, the experimental results with ontoTFIDF which combines ontoSen and TF-IDF in the different proportions are shown in Figure 2. It was found that the classification accuracies of logistic regression, SMO, and J48 methods were almost similar to each other with TF-IDF data. In addition, the classification accuracy of ontoTFIDF by Naïve Bayes method gradually decreased; while the classification accuracy of ontoTFIDF by IBk method gradually increased from 37.39% of original TF-IDF to the maximum of 38.2% at alpha = 0.6. We used a t-test assuming unequal variances to determine if there is a significant difference between the means of the classification accuracy of ontoTFIDF. The results are shown in Table 4: the classification accuracy with ontoTFIDF data was significantly better than that with TF-IDF data on using IBk method, at a significance level of 0.05. Therefore, ontoTFIDF helped significantly improve the performance of sentiment analysis by IBk classification method, compared to TF-IDF scoring.

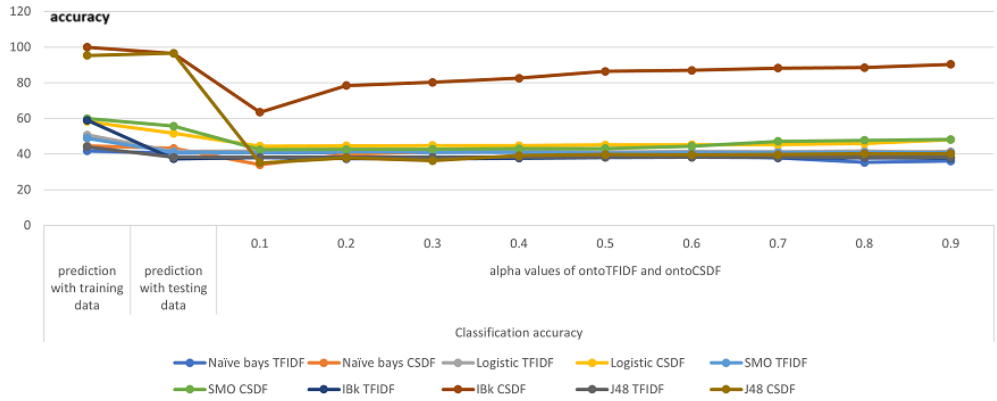

Figure 1. The classification accuracies with TF-IDF, CSDF, ontoTFIDF, and ontoCSDF data

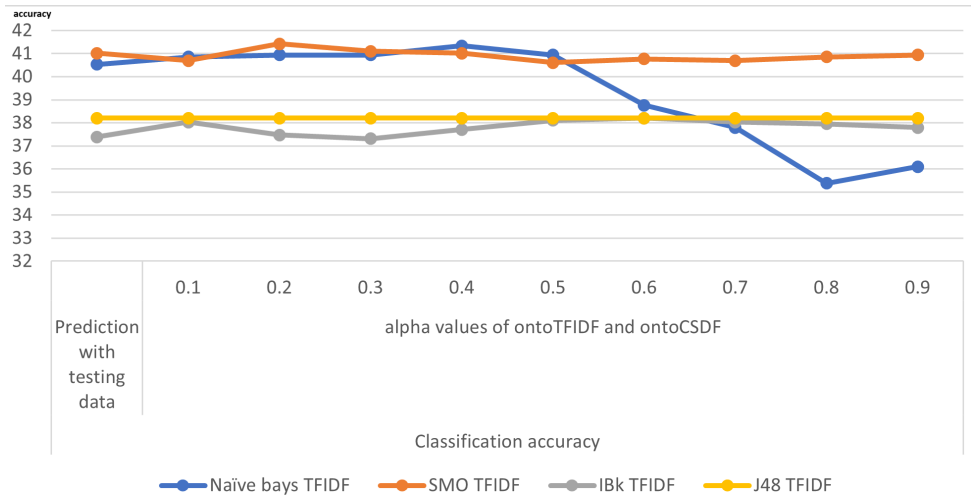

Figure 2. The classification accuracies with TF-IDF and ontoTFIDF data

Table 4. t-Test comparison of using TFIDF and ontoTFIDF (Two-sample t-test assuming unequal variances)

	TFIDF	ontoTFIDF
Mean	37.35	37.91125
Variance	0.0032	0.057498
Observations hypothesized mean difference	0	
df	8	
t Stat	-5.98727	
P(T¡=t) one-tail	0.000164	
t Critical one-tail	1.859548	
P(T¡=t) two-tail	0.000328	
t Critical two-tail	2.306004	

6. CONCLUSION

This study explored two novel approaches to improve the classification accuracy in sentiment analysis by using combinations of a word-based approach (TF-IDF or CSDF) with an ontology-based approach (ontoSen) giving two new methods, called ontoTF-IDF and ontoCSDF. The experimental results show that based on classification accuracy the CSDF was the best data representation method for sentiment analysis, using J48 or decision tree, and IBk or kNN classification methods. The best classification performance was about 96%. However, a combination of CSDF and ontology-based approach, ontoCSDF, did not improve the classification of sentiment analysis data. On the other hand, the classification of TF-IDF data was not good enough; therefore, the combination of TF-IDF and ontoSen (ontoTFIDF) improved the classification accuracy significantly ($p<0.05$) on using IBk classifier.

7. DISCUSSION AND FUTURE WORK

From the observations, CSDF method with normalization adjusts the values measured on different scales to a notionally common scale. The CSDF values are only from 0 to 1. This may have improved the performance of classification of sentiment analysis data. Furthermore, instead of using the ontoSen method, there are alternative ontology-based approaches that could improve the performance of sentiment analysis in the future.

REFERENCES

[1] J. Tao and T. Tan, "Affective computing: a review," in *International Conference on Affective computing and intelligent interaction*, J. Tao, T. Tan, and R. W. Picard, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 981–995. doi: 10.1007/11573548_125.

[2] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, New York, NY, USA: ACM, Oct. 2003, pp. 70–77, doi: 10.1145/945645.945658.

[3] R. Prabowo and M. Thelwall, "Sentiment analysis: a combined approach," *Journal of Informetrics*, vol. 3, no. 2, pp. 143–157, Apr. 2009, doi: 10.1016/j.joi.2009.01.003.

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, 2011, pp. 30–38.

[5] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: the good the bad and the OMG!," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, ICWSM 2011*, Aug. 2011, pp. 538–541, doi: 10.1609/icwsm.v5i1.14185.

[6] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2622–2629, May 2008, doi: 10.1016/j.eswa.2007.05.028.

[7] F. Benamara, C. Cesarano, A. Picariello, D. Reforgiato, and V. S. Subrahmanian, "Sentiment analysis: adjectives and adverbs are better than adjectives alone," *ICWSM 2007 - International Conference on Weblogs and Social Media*, vol. 7, pp. 203–206, 2007.

[8] Y. Zhang, R. Jin, and Z. H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1–4, pp. 43–52, Dec. 2010, doi: 10.1007/s13042-010-0001-0.

[9] G.Salton and M. J. McGill, *Introduction to modern information retrieval*, 1983.

[10] S. Plansangket and J. Q. Gan, "A new term weighting scheme based on class specific document frequency for document representation and classification," in *2015 7th Computer Science and Electronic Engineering Conference, CEEC 2015 - Conference Proceedings*, IEEE, Sep. 2015, pp. 5–8, doi: 10.1109/CEEC.2015.7332690.

[11] K. Munir and M. S. Anjum, "The use of ontologies for effective knowledge modelling and information retrieval," *Applied Computing and Informatics*, vol. 14, no. 2, pp. 116–126, Jul. 2018, doi: 10.1016/j.aci.2017.07.003.

[12] M. Dragoni, S. Poria, and E. Cambria, "OntoSenticNet: a commonsense ontology for sentiment analysis," *IEEE Intelligent Systems*, vol. 33, no. 3, pp. 77–85, May 2018, doi: 10.1109/MIS.2018.033001419.
[13] L. Floridi, *The Blackwell Guide to the Philosophy of Computing and Information*, John Wiley. 2008, doi: 10.1111/b.9780631229193.2003.00008.x.
[14] A. A. Salatino, T. Thanapalasingam, A. Mannocci, A. Birukou, F. Osborne, and E. Motta, "The computer science ontology: a comprehensive automatically-generated taxonomy of research areas," *Data Intelligence*, vol. 2, no. 3, pp. 379–416, Jul. 2020, doi: 10.1162/dint_a_00055.
[15] A. Sharma and S. Kumar, "Machine learning and ontology-based novel semantic document indexing for information retrieval," *Computers and Industrial Engineering*, vol. 176, p. 108940, Feb. 2023, doi: 10.1016/j.cie.2022.108940.
[16] D. K. Jain, S. Qamar, S. R. Sangwan, W. Ding, and A. J. Kulkarni, "Ontology-based natural language processing for sentimental knowledge analysis using deep learning architectures," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 1, pp. 1–17, 2024, doi: 10.1145/3624012.
[17] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*, vol. 463. New York, 1999.
[18] D. Jurafsky and J. H. Martin, *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 2000.
[19] K. Wójcik and J. Tuchowski, "Ontology based approach to sentiment analysis," 2014, [Online]. Available: http://www.researchgate.net/publication/267324473_Ontology_Based_Approach_to_Sentiment_Analysis.
[20] J. Blitzer, M. Dredze, and F. Pereira, "Multidomain sentiment dataset (version 2.0)," Johns Hopkins University, 2009. Accessed: Mar. 23, 2009. [Online]. Available: https://www.cs.jhu.edu/ mdredze/datasets/sentiment/
[21] R. O. Duda and P. E. Hart, *Pattern classification*. new york, 2006.
[22] W. B. Powell, *Approximate dynamic programming: solving the curses of dimensionality*, vol. 703. 2007.
[23] University of Waikato, "Class CfsSubsetEval," Weka Application Programming Interface Document. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CfsSubsetEval.html
[24] Machine Learning Group University of Waikato, "WEKA - The workbench for machine learning," University of Waikato. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/index.html
[25] M. A. Hall, "Correlation-based feature subset selection for machine learning," Univer sity of Waikato, 1998.

BIOGRAPHIES OF AUTHORS

Suthira Plansangket is a lecturer at Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand. She received the B.S. (Computer Science) degree in 2002, M.S. (Computer Science) in 2006 from Prince of Songkla University, Songkhla, Thailand and Ph.D. (Computer Science) in 2017 from University of Essex, Colchester, UK. Her research interests are in machine learning, data science, and information retrieval. She can be contacted at email: suthira.p@psu.ac.th.

Supaporn Kansomkeat received the B.S. (Mathematics) degree in 1991, M.S. (Computer Science) in 1995 from Prince of Songkla University and D. (Computer Engineering) in 2007 from Chulalongkorn University. Since 1996, she has been instructor at Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand. Her research is concerned with software testing, image processing, and machine learning. She can be contacted at email: supaporn.k@psu.ac.th.

Supasit Kajkamhaeng is a lecturer in the Information and Communication Technology Programme,Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand. He received the master's degree in the Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Thailand. His research interests include parallel/distributed computing and high performance computing. He can be contacted at email: supasit.k@psu.ac.th.