

Multi-Class Text Classification on Khmer News Using Ensemble Method in Machine Learning Algorithms

Raksmei Phann, Chitsutha Soomlek, Pusadee Seresangtakul 

Department of Computer Science, College of Computing, Khon Kaen University, Khon Kaen, Kingdom of Thailand

Corresponding author: Pusadee Seresangtakul (pusadee@kku.ac.th)

Abstract

The research herein applies text classification with which to categorize Khmer news articles. News articles were collected from three online websites through web scraping and grouped into nine categories. After text preprocessing, the dataset was split into training and testing sets. We then evaluated the performance of the ensemble learning method via machine learning classifiers with k-fold validation. Various machine learning classifiers were employed, namely logistic regression, Complement Naive Bayes, Bernoulli Naive Bayes, k-nearest neighbours, perceptron, support vector machines, stochastic gradient descent, AdaBoost, decision tree, and random forest were employed. Accuracy was improved for the categorization of Khmer news articles, in which Grid Search CV was used to find the optimal hyperparameters for each machine learning classifier with feature extraction TF-IDF and Delta TF-IDF. The results determined that the highest accuracy was achieved through the ensemble learning method in the support vector machine with the optimal hyperparameters ($C = 10$, kernel = rbf), using feature extraction TF-IDF and Delta TF-IDF, at 83.47% and 83.40%, respectively. The model establishes that Khmer news articles can be accurately categorized.

Keywords

Text classification; Khmer news; Machine learning; Feature extraction; Optimal hyperparameters; News categorization; Ensemble learning method.

Citation: Phann, R., Soomlek, C., & Seresangtakul, P. (2023). Multi-Class Text Classification on Khmer News Using Ensemble Method in Machine Learning Algorithms. *Acta Informatica Pragensia*, 12(2), 243–259. <https://doi.org/10.18267/j.aip.210>

Academic Editor: Stanislav Vojir, Prague University of Economics and Business, Czech Republic

Copyright: © 2023 by the author(s). Licensee Prague University of Economics and Business, Czech Republic.

This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (CC BY 4.0).

1 Introduction

The availability of the internet, smartphones, computers and online news websites enables readers to find and read news more easily via news websites or social media platforms. News can assist readers in gaining a broad understanding of various fields. More and more people get updates on the latest news articles from the internet rather than the traditional method of reading a newspaper, as websites become a more reliable source of articles over time. News websites remain an essential source of articles for education, the economy, sports, technology, entertainment and politics. Categories of articles in news portals or websites are necessary for online readers to find articles that they are curious about. Consequently, online news websites need to shift their news from broad categories to specific categories to make it easy for readers to rapidly search for news articles.

Several Cambodian news websites lack sufficient news categorization necessary to assist newsreaders in quickly finding news articles by filtering specific news categories on their online news websites. For instance, the popular news website *freshnewsasia.com*, launched in 2012, still publishes news today, yet lacks news categorization. As a result, political news, health news and other types of news are all grouped in only one local category, making it difficult and time-consuming for people to find news of interest. Websites must, therefore, categorize new articles by spending time reading and analysing existing news on their websites. To resolve this issue, text classification was applied to categorize Khmer news articles. Text classification has been used in various languages, such as English (Luo, 2021), Arabic (Alsaleh & Larabi-Marie-Sainte, 2021), Chinese (Li, 2022), Khmer (Buoy et al., 2021; Jiang et al., 2022), Korean (Yang et al., 2022) and Spanish (Trigueros et al., 2022).

To evaluate the effectiveness of our approach, we scraped a new dataset of Khmer news articles from three online websites (*news.sabay.com.kh*, *rfa.org*, and *khmer.voanews.com*) and divided them into nine major categories: environment, politics, social, technology, land issues, human rights, health, law, and sports. We then evaluated the ensemble method in machine learning techniques with optimal hyperparameters to determine the most effective approach for news categorization in the Khmer language.

The main contribution of this research is a collection of Khmer news articles for text classification and corpus construction. This research also applies ensemble methods in machine learning techniques with optimal hyperparameters for Khmer news classification.

2 Literature Review

Text classification is a significant task in natural language processing that has been widely studied in the literature. Researchers have compared the performance accuracies of various machine learning algorithms for text classification in order to identify the best approach for a given dataset and task. In the study conducted by Singh et al. (2019), the performance of the Bernoulli Naive Bayes (NB) classifier was compared to the Multinomial NB classifier using a dataset with both a news and polarity column. The results showed that the Multinomial NB classifier had greater accuracy than the Bernoulli NB classifier. One potential advantage of the study is that it directly compares the performance of two specific algorithms, allowing a more detailed analysis of their relative strengths and limitations. However, it should be noted that the dataset used in the study may not be representative of all text classification tasks, and the results may not necessarily generalize to other datasets or languages.

In another study, Azam et al. (2018) used text classification for a dataset owned by Elsevier with five different categories: finance, medicine, agricultural & biological sciences, mathematics, and engineering. They found that the k-nearest neighbour (KNN) algorithm achieved greater accuracy than the Naive Bayes. The study is notable for its use of a larger and more diverse dataset, which may make the results more generalizable to other tasks and languages.

Dlamini et al. (2021) demonstrated that the support vector machine was the best classifier when compared to the Naive Bayes and k-nearest neighbour. The study is notable for its comparison of multiple algorithms and its use of a cross-validation approach, which can help reduce the risk of overfitting and improve the robustness of the results. However, it is worth noting that the support vector machine can be sensitive to the kernel choice and may not always perform well on smaller or less structured datasets.

Wongso et al. (2017) used text classification to categorize Indonesian news articles. They collected a dataset of 5,000 documents from a specific website and used 1,000 documents in five categories: health, economy, politics, sports, and technology. Through the use of the feature TF-IDF and the Bernoulli Naive Bayes algorithm, they achieved an accuracy rate of 98.2%. A slightly higher accuracy (98.4%) was obtained using a combination of TF-IDF and the Multinomial Naive Bayes (MNB) algorithm. The study is notable for its use of a large dataset and its comparison of multiple algorithms, but it is worth noting that the results may not generalize to other languages or tasks due to the specific dataset and pre-processing methods used.

Barua et al. (2021) developed a k-nearest neighbour classifier for a multiclass dataset, which was organized by Fahmi (Nurfikri & Mubarak, 2018) to identify categories of news articles. Their pre-processing steps consisted of case folding, data cleansing, stop-word removal and tokenization. After pre-processing, they used feature TF-IDF with 10-fold cross-validation to separate the test set and training set on the k-nearest neighbour classification model. Their greatest accuracy was 96.71%.

Dhar & Abedin (2021) collected data from well-known online news portals in three categories: sports (158 documents), health (158 documents) and technology (158 documents). In their data pre-processing, they removed digits, punctuation marks, symbols and stop words from documents and tokenized data. To categorize news documents, they used various ML algorithms, such as the support vector machine, k-nearest neighbour and NB. Through the use of a feature extraction vectorizer, TF-IDF, and optimized machine learning, they achieved an average accuracy of 81% for news categorization.

Buoy et al. (2021) performed research on text classification for the Khmer language using neural networks with word embedding. They collected a dataset that contained 13,902 articles, which were labelled by the article authors. In their experiment, they achieved the highest F1-score of 85.3% for the multi-label classification task.

Jiang et al. (2022) conducted news categorization task using a dataset of 7,166 Khmer news articles scraped from khmer.voanews.com. In their experiments, they utilized pre-trained models and fine-tuned on the Khmer news dataset to classify the articles into the 8 different categories. The experiment results showed the highest news classification accuracy of 70.61% after utilizing EasyEnsemble.

3 Proposed Methodology

In this study, we illustrate how we categorized Khmer news articles using the ensemble learning method in machine learning models and feature extraction techniques with optimal hyperparameters. Our methodology includes four main modules: data collection, text preprocessing, hyperparameter optimization and evaluation of the machine learning module. An overview of the proposed methodology is shown in Figure 1. The limitation of the study is that it assumes that each news article can only belong to one category. In total, there are nine categories of the news. These assumptions may lead to an underestimation of the performance of the proposed approach and a less accurate representation of the news articles and their contents.

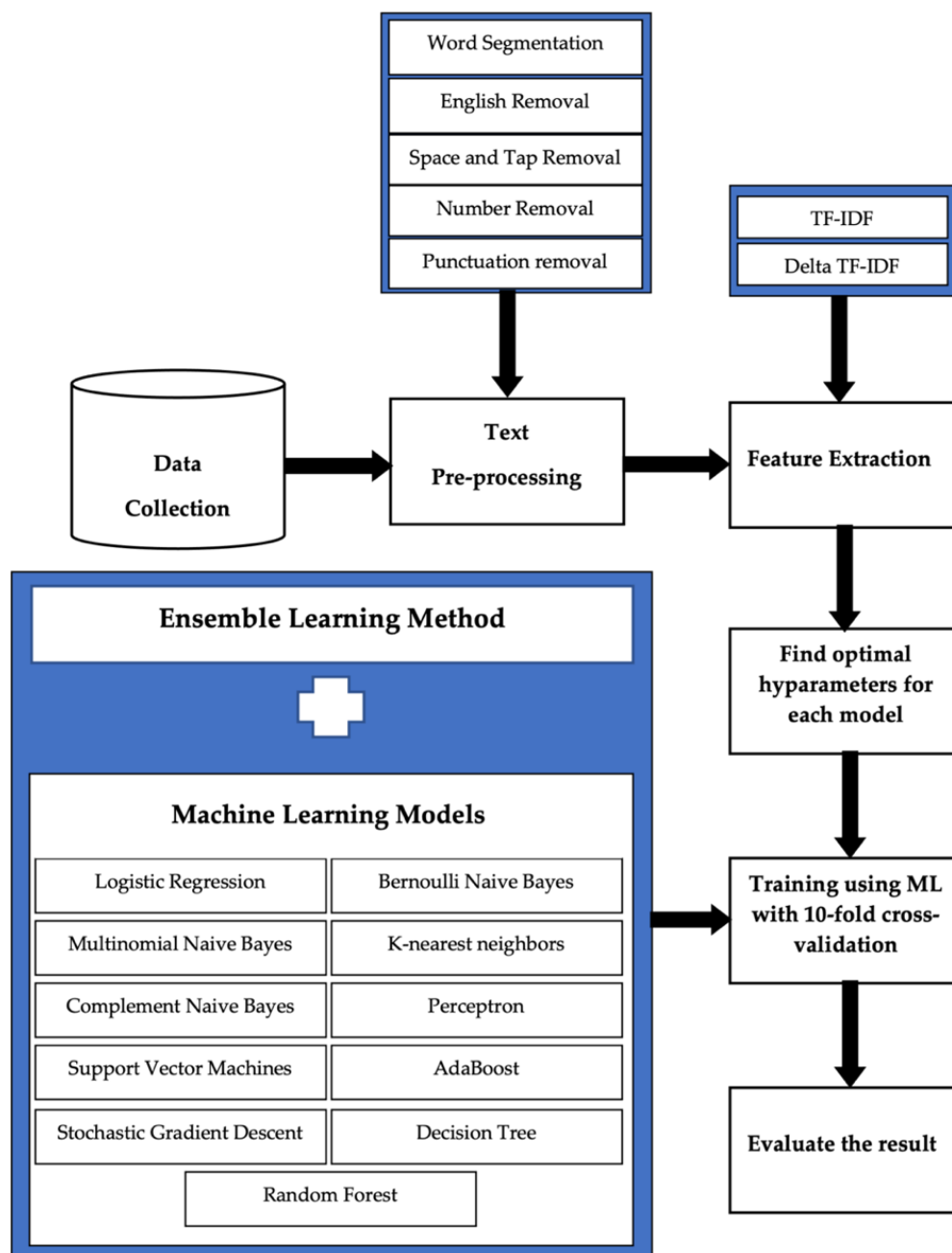


Figure 1. Overall system architecture.

3.1 Data collection

We scraped Khmer news articles from news.sabay.com.kh, a popular Khmer website with over 20 million page views; rfa.org, a website that publishes independent online news in Asia; and khmer.voanews.com, a Khmer news website headquartered in Washington, D.C. We gathered headlines and content from Khmer news articles in nine major categories: environment, politics, social, technology, health, law, human rights, land issues, and sports. Each article that we collected was categorized by the news article's authors and was published from January 2013 to September 2022. We utilized the HTTP library of Python® (Sovietov & Gorchakov, 2022) to send http requests to a specified URL that we needed to access to get the HTML content of the webpage from the server. We then used BeautifulSoup (Zheng et al., 2015), which is a Python® library for retrieving data from HTML content. To parse raw HTML content into a tree structure or nested data, we used html5lib (Uzun et al., 2018) which is a pure-Python® library. The following pseudo-code explains the implementation of data collection.

Algorithm 1: Pseudo-code to collect URLs of news articles from categories*Input:* URL of news category (e.g., <https://news.sabay.com.kh/ajax/topics/sport/>)*Output:* URL list of news articles which contain in each news category

```

1  Initialize URL_list
2  For i In first_page_number To last_page_number // pagination iteration
4      send a HTTP request to url (e.g., https://news.sabay.com.kh/ajax/topics/sport/)
5      parse HTML content to tree structure of HTML data using Python® library
        BeautifulSoup with html5lib
6      get URL of news article by HTML href Attribute
7      append URL of a news article to URL_list
8       $i \leftarrow i + 1$ 
9  End For

```

Algorithm 2: Pseudo-code to collect headlines and contents of news articles*Input:* URL list of news articles*Output:* headlines and content of news articles

```

1  For i In URL_list (we got URL_list from Algorithm 1)
2      send a HTTP request to i
3      parse HTML content to tree structure of HTML data using Python® library
        BeautifulSoup with html5lib
4      access headline by HTML class name ("title detail") with HTML tag p for website
        news.sabay.com.kh, HTML h1 tag for website rfa.org, and HTML title tag for
        website khmer.voanews.com
5      access content by HTML class name ("detail content-detail") with HTML tag p
        for website news.sabay.com.kh, HTML id name ("storytext") for website rfa.org and
        HTML tag p for khmer.voanews.com
6  End For

```

We scraped a total of 21,685 Khmer news articles from the websites news.sabay.com.kh, rfa.org and khmer.voanews.com. The average number of words in each news article was 554, and the number of Khmer news articles in each category is displayed in Figure 2. Our dataset contained the headline, content and class columns, which were labelled by the news article authors of each Khmer news website. Table 1 shows examples within the dataset for Khmer news articles.

Table 1. Examples of dataset for Khmer news articles.

Headline	Content	Class
អ្នកឃ្លាំមើលបរិស្ថានថា.... (Environmental monitors say...)	ក្រុមអ្នកឃ្លាំមើលបរិស្ថានប្រតិកម្មនឹងក្រសួងបរិស្ថាន.... (Environmental monitors react to the Ministry of Environment...)	0 (Environment)
មន្ត្រីសង្គមស៊ីវិល ទាមទារឱ្យអាជ្ញាធរ... (Civil society officials demand law enforcement...)	មន្ត្រីអង្គការសង្គមស៊ីវិល ជំរុញដល់អាជ្ញាធរខេត្ត... (Civil society officials urge provincial authorities...)	1 (Social)
ក្រុមអ្នកវិភាគថា វិបត្តិនយោបាយឆ្នាំ២០២១ នឹងមិនប្រសើរ... (Analysts say the 2021 political crisis will not improve...)	អ្នកឃ្លាំមើលនយោបាយសង្កេតឃើញថា... (Political observers observe that...)	2 (Politics)
ក្រសួងប្រៃសណីយ៍បង្កើត ប្រព័ន្ធសេវាសាធារណៈឌីជីថល... (Ministry of Posts creates digital public service system...)	ក្រសួងប្រៃសណីយ៍និងទូរគមនាគមន៍ បានបង្កើត... (The Ministry of Posts and Telecommunications established...)	3 (Technology)
កីឡាករវ័យក្មេងឆ្នើមៗ នឹងចូលរួម... (The best young players will attend...)	ខាងក្រោមនេះជាកីឡាករវ័យក្មេងឆ្នើមៗទាំង១០រូប... (Here are the top 10 young players...)	4 (Sports)
អង្គការយូនីសេហ្វបារម្ភពីបញ្ហាសុខភាពផ្លូវចិត្តរបស់... (UNICEF concerned about the mental health of Cambodian youth...)	មូលនិធិសហប្រជាជាតិដើម្បីកុមារ... (United Nations Children's Fund...)	5 (Health)
ប្រព័ន្ធយុត្តិធម៌នៅកម្ពុជាគួរ... (The Cambodian justice system should...)	អ្នកវិភាគនយោបាយថា... (Political analysts say...)	6 (Law)
ពលរដ្ឋមានជម្លោះដីជាមួយ... (Citizens have land disputes with...)	សហគមន៍ខេត្តកោះកុង មានទំនាស់ដីធ្លី... (Koh Kong Communities have land disputes...)	7 (Land issues)
អាជ្ញាធរវៀតណាមមិនទាន់... (Vietnamese authorities have not yet...)	ស្ត្រីខ្មែរក្រមម្នាក់ដែល... (A Khmer Krom woman who...)	8 (Human rights)

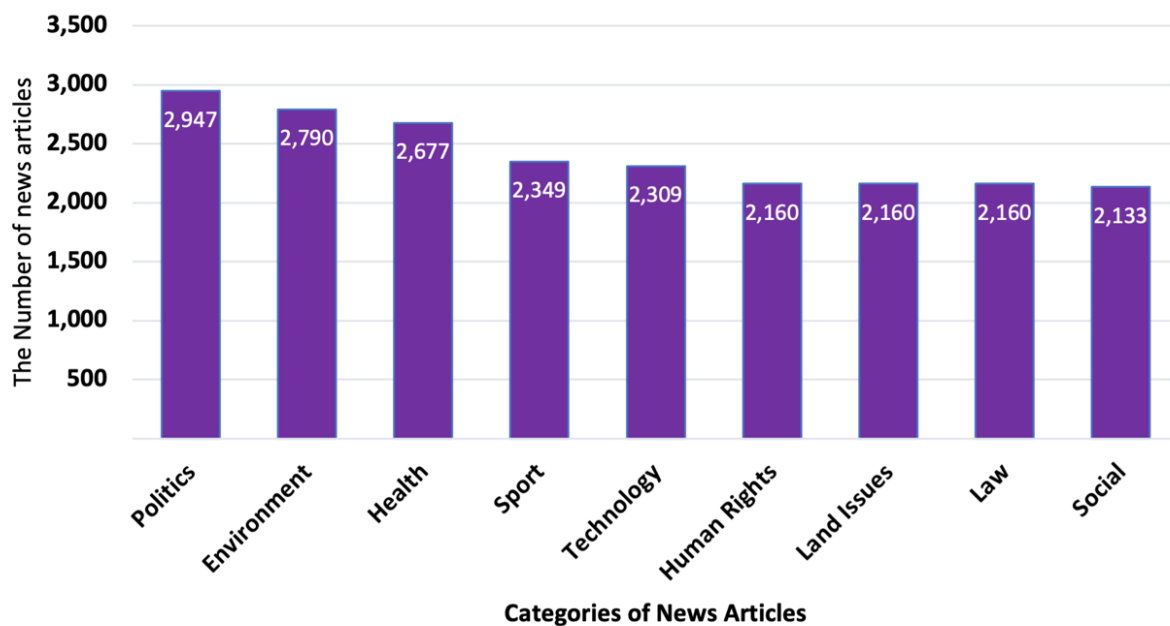


Figure 2. Numbers of categorized Khmer news articles.

3.2 Text pre-processing

Text cleaning or text pre-processing is a technique that has been utilized to remove outliers and meaningless data to improve the validity of a dataset. Moreover, text pre-processing, such as removing numbers or lemmatization, is a task necessary to improve accuracy (Symeonidis et al., 2018). The Khmer language is written without spaces between words and, as a low-resource language, text preparation for Khmer news articles is challenging. We used a Khmer word segmentation using conditional random fields (CRF) created by (Chea et al., 2015) to tokenize words in the Khmer language. For text classification, a variety of text pre-processing techniques were used extensively to reduce the text length and increase the validity of the dataset. In Khmer news articles, the content does not contain emoticons or emojis, and word spelling errors are also checked by the journalist or news editor before the news is published. The research herein therefore utilizes five steps in data preprocessing: space and tab removal, punctuation removal, number removal, English removal, and Khmer word segmentation.

Space and tab removal

Khmer news articles were collected by scraping HTML tags containing blank spaces, whitespaces or tabs. We used the “replace” method of Python® programming to replace all whitespaces, blank spaces and tabs with an empty character.

Punctuation removal

Journalists utilize punctuation such as exclamation marks, question marks, periods, parentheses and commas to separate sentences in news articles. We replaced punctuation with empty characters using the “replace” method within the built-in libraries in Python®.

Number removal

Numbers reveal nothing about text analysis and harm accuracy when computing similarity (Khamphakdee & Seresangtakul, 2021a). In this step, we used a regular expression module, which is a built-in package in the Python® programming language, to remove numbers utilized in the news.

English removal

Khmer news articles also contain the English language, which has a negative impact on Khmer word segmentation. We utilized the regular expression to remove all English language in our dataset.

Word segmentation

In written Khmer, there is a lack of standardization for the use of spaces or dividers between words. To overcome this challenge, we applied Khmer word segmentation using conditional random fields (CRF) created by Chea et al. (2015). By using CRF, we were able to accurately tokenize the Khmer words in our dataset, which is an important step in preparing the data for text classification.

3.3 Feature extraction

Feature extraction can improve the accuracy and execution time of a machine learning classification model (Liang et al., 2017). The major purpose of feature extraction is to minimize dimensionality and reduce unnecessary attributes to improve the text classification technique (Vidhya et al., 2016). Khamphakdee & Seresangtakul (2021b), who investigated text classification for the Thai language, found that, among the feature extraction techniques applied (TF-IDF, N-Gram, Delta TF-IDF and Word2Vec), the Delta TF-IDF and TF-IDF feature extraction achieved higher accuracy than the bigram and trigram feature extraction. As a result, in this study, we used both Delta TF-IDF and TF-IDF feature extraction techniques to convert Khmer news articles into a feature matrix, as Khmer and Thai are languages that are closely related and share many similarities.

TF-IDF

TF-IDF is a method of extracting keywords from documents depending on their contents. The mathematical formula of word significance within their contents was included in the TF-IDF (Yao et al., 2019). TF and IDF are two essential parts of the TF-IDF, given in Equation 1. Equation 2 represents the IDF as a constant per corpus and accounts for the ratio of documents that features that specific term. TF is the frequency of any term in a given document, expressed in Equation 3.

$$TFIDF = TF(t, d) \times IDF(t, d) \quad (1)$$

$$IDF(t, d) = \log \frac{\text{Total number of documents } d}{\text{Number of documents with term } t \text{ in it}} \quad (2)$$

$$TF(t, d) = \frac{\text{Number of times term } t \text{ appears in document } t}{\text{Total number of terms in document } d} \quad (3)$$

Delta TF-IDF

For classification purposes, the Delta TF-IDF method was used to convert a term into a numeric matrix, which is simple to calculate, utilize and comprehend (Martineau & Finin, 2009). Delta TF-IDF is generally associated with text classification tasks when a dataset belongs to more than one class. In our dataset, we used the feature Delta TF-IDF to extract Khmer news articles that belong to more than one class. Delta TF-IDF computed the various TF-IDF scores of words in the training corpus negative and positive classes via Equations 4, 5 and 6.

$$V_{t,d} = C_{t,d} * \log_2 \left(\frac{|P|}{P_t} \right) - C_{t,d} * \log_2 \left(\frac{|N|}{N_t} \right) \quad (4)$$

$$V_{t,d} = C_{t,d} * \log_2 \left(\frac{|P|N_t}{P_t|N|} \right) \quad (5)$$

$$V_{t,d} = C_{t,d} * \log_2 \left(\frac{|N_t|}{P_t} \right) \quad (6)$$

where $V_{t,d}$ denotes the feature value of the word in the context inside the document d , $C_{t,d}$ indicates how frequently the word t occurs in the document d , P_t denotes the quantity of the training set documents that have positive labeling for the term t , $|P|$ denotes the number of training set documents that do not have positive labelling, N_t represents the number of negative documents containing the term t , and $|N|$ is the amount of training set documents that have not been negatively labelled.

3.4 Ensemble learning method

Ensemble learning is a method that boosts classification problem accuracy and reliability (Mousavi & Eftekhari, 2015). The ensemble learning method can be used to increase the accuracy of any machine learning model by combining classifications or predictions (Seni & Elder, 2010). Bagging is a classifier of the ensemble learning method, which can reduce overfitting and improve the accuracy of the machine learning algorithm (Ponnaganti & Anitha, 2022; Sahoo et al., 2021). In our work, we utilized a bagging classifier with ten estimators in eleven machine-learning models. Figure 3 displays the bagging classifier of the ensemble learning method flow.

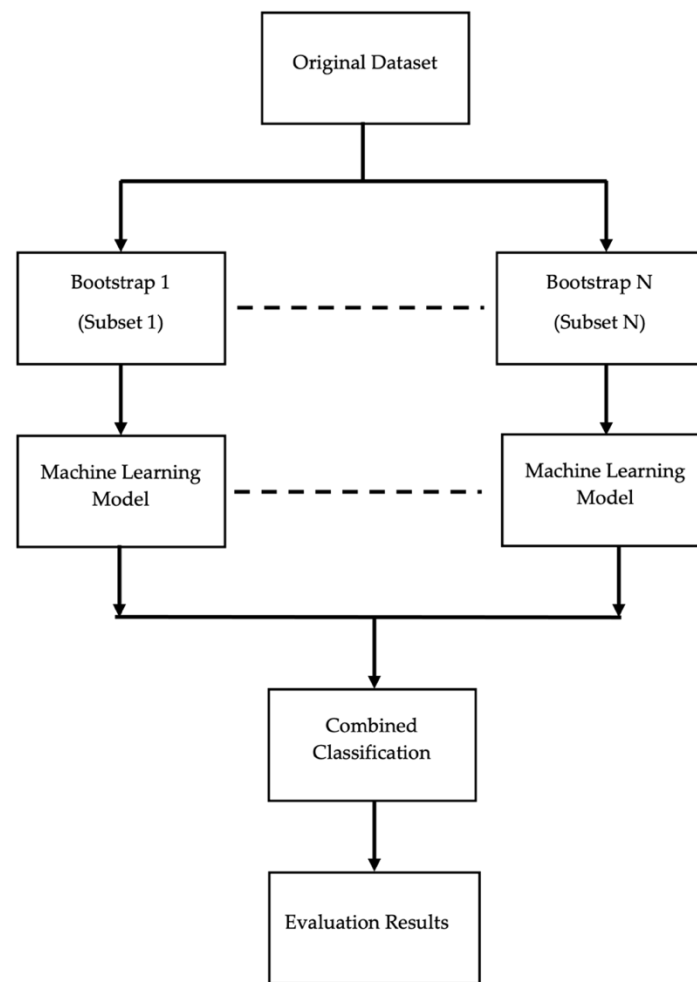


Figure 3. Bagging classifier of ensemble learning method flow.

3.5 Machine learning classification model

We chose machine learning for our dataset of 21,685 Khmer news articles, as deep learning requires large datasets and powerful computing resources (Kowsari et al., 2019). Various machine learning techniques have undergone scrutiny in performance accuracy comparisons. Singh et al. (2019) found that the accuracy of the Multinomial NB classifier was superior to the Bernoulli NB classifier. Furthermore, Azam et al. (2018) demonstrated that the k-nearest neighbour (KNN) algorithm outperformed the Naive Bayes algorithm in terms of accuracy. Dlamini et al. (2021) proved that the support vector machine (SVM) was superior to both the Naive Bayes and k-nearest neighbour classifiers. Maldonado et al. (2014) found that the SVM performs significantly better when handling numerical features and high-dimensional datasets. While researchers have confirmed the better performance of the SVM, we still must implement a variety of machine learning classifiers to determine which model provides the best performance for the Khmer language dataset. In our research, there are 11 machine learning classifiers, including logistic regression (Hu et al., 2022), Complement Naive Bayes (Umar & Nur, 2022), Bernoulli Naive Bayes (Verma et al., 2020), k-nearest neighbours (Babič et al., 2020), perceptron (Sagheer et al., 2019), support vector machines (Petridis et al., 2022), stochastic gradient descent (Bianchi et al., 2022), AdaBoost (Lee et al., 2022), decision tree (Arabameri et al., 2022) and random forest (Khan et al., 2022).

4 Experiment and Results

In our experiment, the dataset was split into training and testing sets at an 80:20 ratio, with 17,348 and 4,337 news articles in the training and testing sets, respectively. We combined headlines and content of Khmer news articles and began text pre-processing (space and tab removal, punctuation removal, number

removal, English removal and word segmentation). All news articles were converted into vectors of features through Delta TF-IDF and TF-IDF vectorization. Optimal hyperparameters improved the accuracy of the supervised learning algorithms. We used Grid Search cross-validation to find the optimal hyperparameters for each machine learning classifier to improve the accuracy of the categorization of Khmer news articles. Grid Search cross-validation, a library of the scikit-learn framework (Buitinck et al., 2013), provided methods that we utilized in the optimization hyperparameters. The optimal hyperparameters for each machine learning classification model are given in Table 2.

After finding optimal hyperparameters for each model, we evaluated the performance of the bagging classifier in machine learning classifiers with k-fold validation (k=10), where k denotes the number of groups or folds in a given data sample. We utilized the open-source framework scikit-learn (Buitinck et al., 2013) to investigate the performance of each classification model. The following formulas (7, 8, 9, and 10) were used to evaluate each machine learning classifier.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (7)$$

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$F1\ score = \frac{2(Precision \times Recall)}{Precision + Recall} \quad (10)$$

Where FP is the number of false positives, TP is the number of true positives, TN is the number of true negatives and FN is the number of false negatives (Gamal et al., 2019). The experiments were run on Google Colab's virtual machines utilizing the specifications, which include 2 vCPUs, 13 GB of RAM, an Nvidia Tesla K80 GPU with 12 GB of GPU memory, 30 GB of persistent disk storage, and a 1 Gbps network connection. Table 3 shows the results of the evaluation for ensemble learning in each machine learning classifier using the optimal hyperparameters. The eleven machine learning algorithms produced significantly different values of accuracy at $p < 0.05$. Figures 4 and 5 display the evaluation results of all models using the feature extractions TF-IDF and Delta TF-IDF, respectively. Based on descriptive statistics, the highest accuracy was achieved by the support vector machines using feature extractions TF-IDF and Delta TF-IDF, at 83.47% and 83.40%, respectively. Table 3 illustrates that feature extractions TF-IDF and Delta TF-IDF with the SVM models that achieved the highest F1 scores, precision and recall.

As a result, the best-performing of the machine learning algorithms trained on our dataset was the support vector machine classification model with optimal hyperparameters ($C = 10$, kernel = rbf). Figure 6 shows a comparison of the Delta TF-IDF and TF-IDF features in terms of accuracy. Feature extraction Delta TF-IDF and TF-IDF produced different values of accuracy, but not significantly at $p > 0.05$.

Table 2. Optimal hyperparameters.

Classification model	Hyperparameter
Multinomial Naive Bayes (MNB)	fit_prior = False, alpha = 0.5
Logistic regression (LR)	penalty = l2, C= 10, solver = liblinear
Complement Naive Bayes (CNB)	norm = False, alpha= 0.5, fit_prior = True

Classification model	Hyperparameter
Bernoulli Naive Bayes (BNB)	fit_prior = True, alpha= 0.5
K-nearest neighbour (KNN)	n_neighbors = 4, weights = distance, algorithm = auto, leaf_size = 1, p = 2
Perceptron	penalty = l2, eta0 = 0.01, max_iter = 50
Support vector machines (SVM)	C = 10, kernel = rbf
Stochastic gradient descent (SGD)	loss = squared_hinge, penalty = l2, alpha = 0.01
AdaBoost (ADA)	n_estimators = 500, learning_rate = 1.0
Decision tree (DT)	criterion = gini, splitter = best, max_depth = None, min_samples_split = 2, min_samples_leaf= 1
Random forest (RF)	n_estimators = 40, max_features = auto, criterion = gini, max_depth = None, min_samples_split = 2, min_samples_leaf = 2

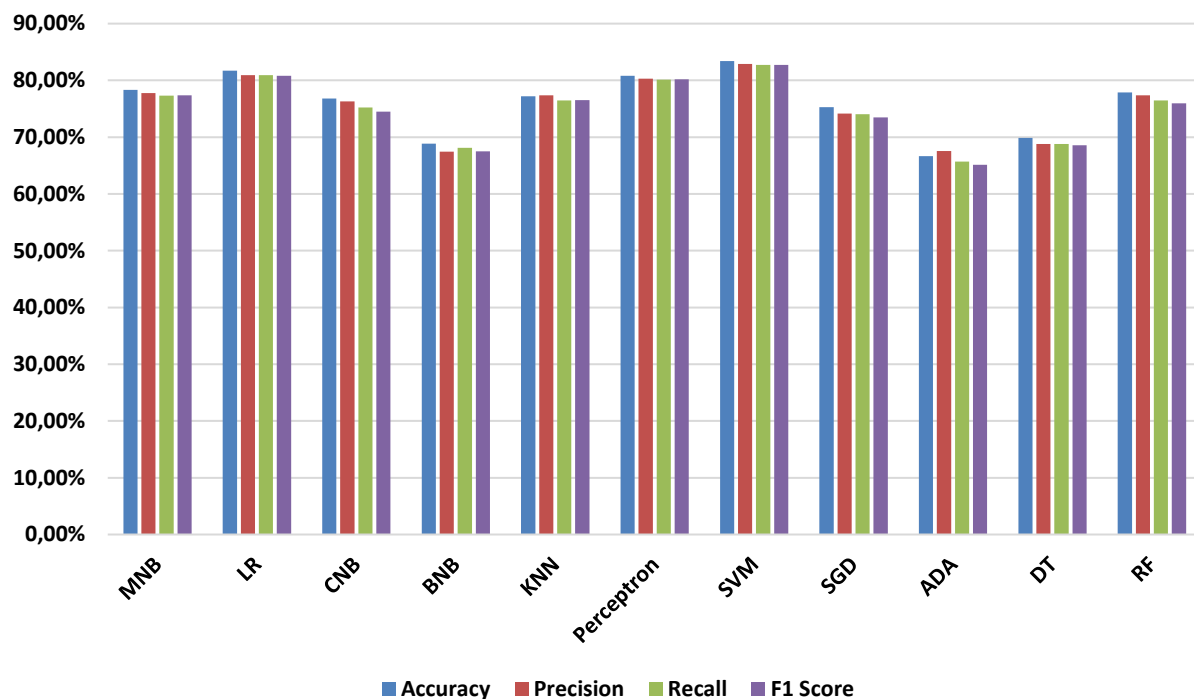


Figure 4. Evaluation results of all models using feature extraction TF-IDF.

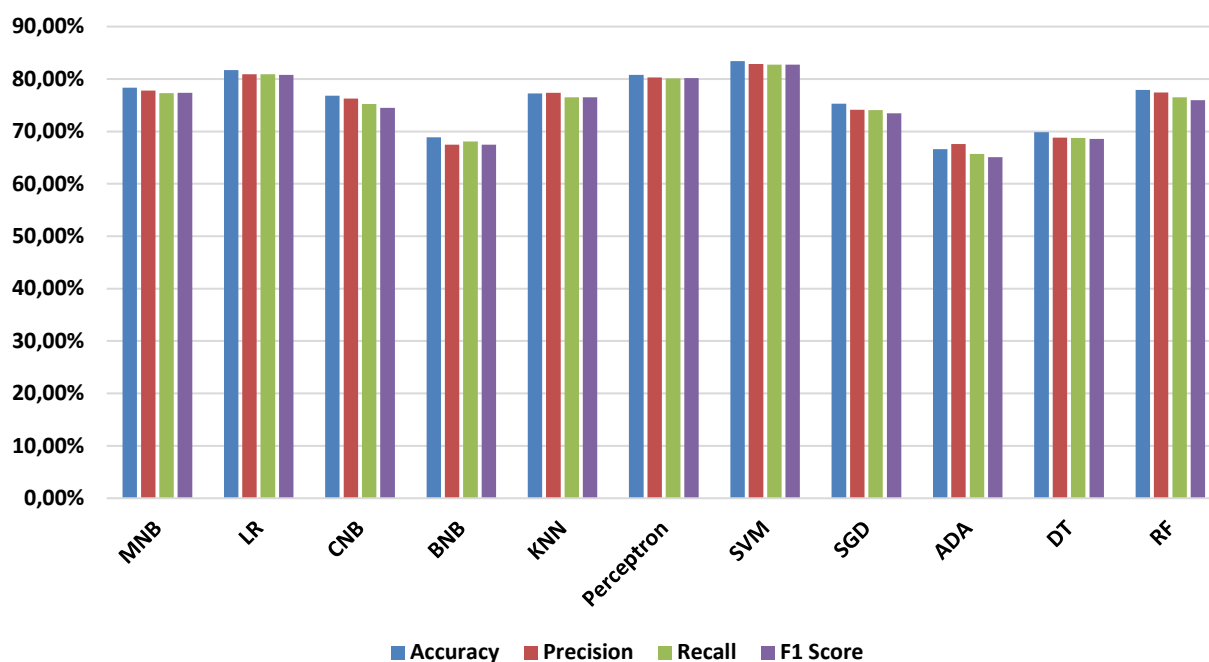


Figure 5. Evaluation results of all models using feature extraction Delta TF-IDF.

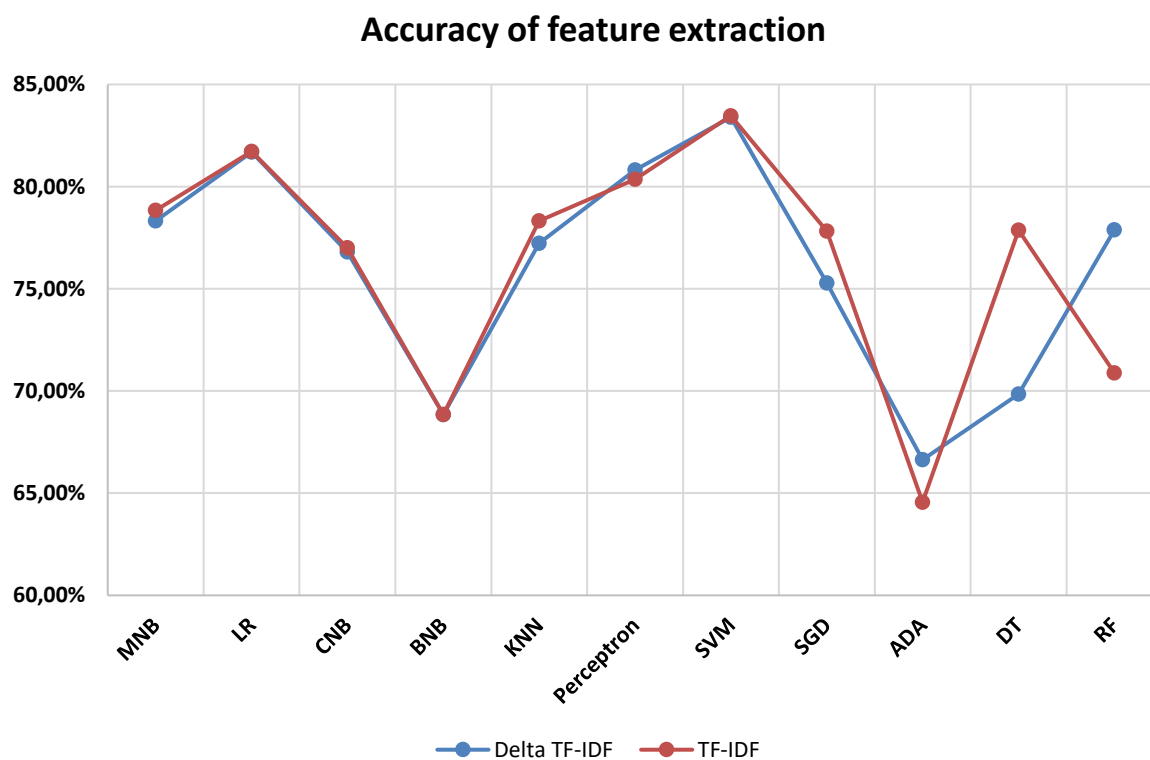


Figure 6. Performance comparison of feature extraction Delta TF-IDF and TF-IDF.

Table 3. Evaluation results for each machine learning classifier using optimal hyperparameters.

Machine learning classifier	Feature extraction	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Training time (sec)	Testing time (sec)
Multinomial Naive Bayes (MNB)	TF-IDF	78.83	78.22	77.98	77.99	12. 53	1.23
	Delta TF-IDF	78.33	77.78	77.33	77.36	16. 21	1.31
Logistic regression (LR)	TF-IDF	81.72	81.02	80.97	80.90	18. 61	1.22
	Delta TF-IDF	81.69	80.90	80.94	80.81	18. 92	1.27
Complement Naive Bayes (CNB)	TF-IDF	77.01	76.38	75.37	74.74	12.32	1.22
	Delta TF-IDF	76.80	76.28	75.21	74.47	15. 43	1.27
Bernoulli Naive Bayes (BNB)	TF-IDF	68.85	67.45	68.10	67.49	6. 19	1.26
	Delta TF-IDF	68.85	67.45	68.10	67.49	15. 31	1.29
K-nearest neighbour (KNN)	TF-IDF	78.33	78.07	77.58	77.56	18. 87	10.19
	Delta TF-IDF	77.22	77.36	76.49	76.50	30. 63	11.55
Perceptron	TF-IDF	80.36	79.70	79.74	79.61	2.98	1.24
	Delta TF-IDF	80.82	80.30	80.14	80.17	15.11	1.28
Support vector machines (SVM)	TF-IDF	83.47	82.95	82.83	82.84	168. 31	55.72
	Delta TF-IDF	83.40	82.87	82.73	82.75	179. 73	56.06
Stochastic gradient descent (SGD)	TF-IDF	77.82	76.85	76.61	76.20	2. 88	1.22
	Delta TF-IDF	75.28	74.14	74.07	73.46	16. 12	1.28
AdaBoost (ADA)	TF-IDF	64.56	65.54	64.01	63.07	90. 82	2.79
	Delta TF-IDF	66.64	67.58	65.72	65.11	93. 81	2.70
Decision tree (DT)	TF-IDF	77.86	77.23	76.42	75.95	4.91	1.24
	Delta TF-IDF	69.84	68.79	68.77	68.60	21.11	1.30
Random forest (RF)	TF-IDF	70.88	69.77	69.77	69.54	4.87	1.34
	Delta TF-IDF	77.89	77.40	76.49	75.98	21.14	1.36

5 Conclusion and Future Work

To categorize Khmer news articles, we scraped data from three websites (news.sabay.com.kh, khmer.voanews.com and rfa.org) gathering both the headlines and contents of Khmer news articles. We then developed categories (environment, politics, social, technology, health, law, human rights, land issues, and sports) using BeautifulSoup; which is a web scraping library and Python® programming language. BeautifulSoup works as a tool to write scripts extracting only the desired information from a particular web page. We optimized the hyperparameters for each machine learning classifier with preprocessing data, namely space and tab removal, punctuation removal, number removal, English removal and word segmentation. To define the most suitable supervised learning technique for the categorization of Khmer news articles, we implemented the ensemble learning method in various machine learning classification models; such as logistic regression, Complement Naive Bayes, perceptron, Bernoulli Naive Bayes, support vector machine, k-nearest neighbours, stochastic gradient descent, AdaBoost, decision tree, and random forest. The optimal hyperparameters were determined using k-fold cross-validation and the TF-IDF and Delta TF-IDF features. Grid Search CV was also used to improve the accuracy of the categorization of news articles in Khmer.

The highest accuracies were achieved by the support vector machine using the TF-IDF and Delta TF-IDF feature extraction methods, at 83.47 % and 83.40 %, respectively. These results demonstrate the effectiveness of the SVM algorithm for text classification in the Khmer language, and suggest that feature extraction methods such as TF-IDF and Delta TF-IDF can be useful for improving the performance of text classification models.

This study investigated the ensemble learning method in various machine learning models for Khmer text classification, which is useful for low-resource languages such as Khmer, while deep learning requires large datasets and powerful computing resources (Kowsari et al., 2019). Additionally, pre-training has several limitations, such as the inability to fully comprehend the diverse meanings of a word in a particular context, the high memory usage required for storage, and the incapacity to identify out-of-vocabulary words from the corpus (Kowsari et al., 2019).

This study also contributed to the collection of text classification data for Khmer news articles, which can be used in future research. In the future, we plan to collect more Khmer news articles and news categories in order to further expand the scope of our study and explore the use of ensemble learning in deep learning models with optimal hyperparameters to improve the performance of multi-class text classification on Khmer news articles.

Additional Information and Declarations

Acknowledgments: The authors would like to thank Assoc. Prof. Dr. Wichuda Chaisiwamongkol for her suggestion on statistical analysis.

Funding: This work was funded by the Royal Scholarship under Her Royal Highness Princess Maha Chakri Sirindhorn Education Project to the Kingdom of Cambodia and Computer Science Department, College of Computing, Khon Kaen University, Thailand.

Conflict of Interests: The authors declare no conflict of interest.


Author Contributions: R.P.: Conceptualization; Methodology; Data curation; Software; Validation; Formal analysis; Investigation; Writing – Original draft preparation. C.S.: Supervision; Conceptualization; Methodology; Validation; Writing – Reviewing and Editing. P.S.: Supervision; Conceptualization; Methodology; Resources; Validation; Formal Analysis; Writing – Reviewing and Editing.

Data Availability: The data that support the findings of this study are available from the corresponding author.

References

- Alsaleh, D., & Marie-Sainte, S. L. (2021). Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms. *IEEE Access*, 9, 91670–91685. <https://doi.org/10.1109/access.2021.3091376>
- Arabameri, A., Pal, S. C., Rezaie, F., Chakraborty, R., Saha, A. K., Blaschke, T., Di Napoli, M., Ghorbanzadeh, O., & Ngo, P. T. T. (2021). Decision tree based ensemble machine learning approaches for landslide susceptibility mapping. *Geocarto International*, 37(16), 4594–4627. <https://doi.org/10.1080/10106049.2021.1892210>
- Azam, M., Ahmed, T., Sabah, F., & Hussain, M. I. (2018). Feature Extraction based Text Classification using K-Nearest Neighbor Algorithm. *International Journal of Computer Science and Network Security*, 18(12), 95–101.
- Babič, F., Pustková, L., & Majnarić, L. T. (2020). Mild Cognitive Impairment Detection Using Association Rules Mining. *Acta Informatica Pragensia*, 9(2), 92–107. <https://doi.org/10.18267/j.aip.135>
- Barua, A., Sharif, O., & Hoque, M. M. (2021). Multi-class Sports News Categorization using Machine Learning Techniques: Resource Creation and Evaluation. *Procedia Computer Science*, 193, 112–121. <https://doi.org/10.1016/j.procs.2021.11.002>
- Bianchi, P., Hachem, W., & Schechtman, S. (2022). Convergence of Constant Step Stochastic Gradient Descent for Non-Smooth Non-Convex Functions. *Set-valued and Variational Analysis*, 30(3), 1117–1147. <https://doi.org/10.1007/s11228-022-00638-z>
- Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., & Grobler, J. (2013). API design for machine learning software: experiences from the scikit-learn project. *ArXiv Preprint ArXiv:1309.0238*. <https://doi.org/10.48550/arXiv.1309.0238>
- Buoy, R., Taing, N., & Chenda, S. (2021). Khmer Text Classification Using Word Embedding and Neural Networks. *ArXiv Preprint ArXiv:2112.06748*. <https://doi.org/10.48550/arXiv.2112.06748>
- Chea, V., Thu, Y. K., Ding, C., Utiyama, M., Finch, A., & Sumita, E. (2015). Khmer word segmentation using conditional random fields. <https://att-astrec.nict.go.jp/member/ding/KhNLP2015-SEG.pdf>
- Dhar, P., & Abedin, M. Z. (2021). Bengali News Headline Categorization Using Optimized Machine Learning Pipeline. *International Journal of Information Engineering and Electronic Business*, 13(1), 15–24. <https://doi.org/10.5815/ijieeb.2021.01.02>
- Dlamini, G., Kholmatova, Z., Kruglov, A., Succi, G., Tarasau, H., & Valeev, A. (2021). Meta-analytical Comparison of SVM and KNN for Text Classification. In *2021 International Conference Nonlinearity, Information and Robotics*. IEEE. <https://doi.org/10.1109/nir52917.2021.9666133>
- Gamal, D., Alfonse, M., El-Horbaty, E. M., & Salem, A. M. (2019). Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features. *Procedia Computer Science*, 154, 332–340. <https://doi.org/10.1016/j.procs.2019.06.048>
- Hu, X., Luo, H., Guo, M., & Wang, W. (2022). Ecological technology evaluation model and its application based on Logistic Regression. *Ecological Indicators*, 136, 108641. <https://doi.org/10.1016/j.ecolind.2022.108641>
- Jiang, S., Fu, S., Lin, N., & Fu, Y. (2022). Pretrained models and evaluation data for the Khmer language. *Tsinghua Science & Technology*, 27(4), 709–718. <https://doi.org/10.26599/tst.2021.9010060>
- Khamphakdee, N., & Seresangtakul, P. (2021a). A Framework for Constructing Thai Sentiment Corpus using the Cosine Similarity Technique. In *2021 13th International Conference on Knowledge and Smart Technology (KST)*, (pp. 202–207). IEEE. <https://doi.org/10.1109/KST51265.2021.9415802>
- Khamphakdee, N., & Seresangtakul, P. (2021b). Sentiment Analysis for Thai Language in Hotel Domain Using Machine Learning Algorithms. *Acta Informatica Pragensia*, 10(2), 155–171. <https://doi.org/10.18267/j.aip.155>
- Khan, M. S., Shah, M. A., Javed, M. S., Khan, M. I., Rasheed, S., El-Shorbagy, M. A., El-Zahar, E. R., & Malik, M. (2021). Application of random forest for modelling of surface water salinity. *Ain Shams Engineering Journal*, 13(4), 101635. <https://doi.org/10.1016/j.asej.2021.11.004>
- Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L. E., & Brown, D. E. (2019). Text Classification Algorithms: A Survey. *Information*, 10(4), 150. <https://doi.org/10.3390/info10040150>
- Lee, S., Tseng, C., Yang, H., Jin, X., Jiang, Q., Pu, B., Hu, W., Liu, D., Huang, Y., & Zhao, N. (2022). Random RotBoost: An Ensemble Classification Method Based on Rotation Forest and AdaBoost in Random Subsets and Its Application to Clinical Decision Support. *Entropy*, 24(5), 617. <https://doi.org/10.3390/e24050617>
- Li, X. (2022). Chinese Language and Literature Online Resource Classification Algorithm Based on Improved SVM. *Scientific Programming*, 2022, Article ID 4373548. <https://doi.org/10.1155/2022/4373548>
- Liang, H., Sun, X. W., Sun, Y., & Gao, Y. (2017). Text feature extraction based on deep learning: a review. *Eurasip Journal on Wireless Communications and Networking*, 2017(1). <https://doi.org/10.1186/s13638-017-0993-1>
- Luo, X. (2021). Efficient English text classification using selected Machine Learning Techniques. *Alexandria Engineering Journal*, 60(3), 3401–3409. <https://doi.org/10.1016/j.aej.2021.02.009>

- Maldonado, S., Weber, R., & Famili, F.** (2014). Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines. *Information Sciences*, 286, 228–246. <https://doi.org/10.1016/j.ins.2014.07.015>
- Martineau, J., & Finin, T.** (2009). Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In *Third AAAI International Conference on Weblogs and Social Media* (pp. 1–4). AAAI.
- Mousavi, R., & Eftekhari, M.** (2015). A new ensemble learning methodology based on hybridization of classifier ensemble selection approaches. *Applied Soft Computing*, 37, 652–666. <https://doi.org/10.1016/j.asoc.2015.09.009>
- Nurfikri, F. S., & Mubarak, M. S.** (2018). News topic classification using mutual information and bayesian network. In *2018 6th International Conference on Information and Communication Technology (ICICT)*, (pp. 162–166). IEEE. <https://doi.org/10.1109/ICICT.2018.8528806>
- Petridis, K., Tampakoudis, I. A., Drogas, G., & Kiosses, N.** (2022). A Support Vector Machine model for classification of efficiency: An application to M&A. *Research in International Business and Finance*, 61, 101633. <https://doi.org/10.1016/j.ribaf.2022.101633>
- Ponnaganti, N. D., & Anitha, R.** (2022). A Novel Ensemble Bagging Classification Method for Breast Cancer Classification Using Machine Learning Techniques. *Traitement Du Signal*, 39(1), 229–237. <https://doi.org/10.18280/ts.390123>
- Sagheer, A., Zidan, M. A., & Abdelsamea, M. M.** (2019). A Novel Autonomous Perceptron Model for Pattern Classification Applications. *Entropy*, 21(8), 763. <https://doi.org/10.3390/e21080763>
- Sahoo, R., Pasayat, A. K., Bhowmick, B., Fernandes, K. J., & Tiwari, M. K.** (2021). A hybrid ensemble learning-based prediction model to minimise delay in air cargo transport using bagging and stacking. *International Journal of Production Research*, 60(2), 644–660. <https://doi.org/10.1080/00207543.2021.2013563>
- Seni, G., & Elder, J. F.** (2010). *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool.
- Singh, G., Kumar, B., Gaur, L., & Tyagi, A.** (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. In *2019 International Conference on Automation, Computational and Technology Management*, (pp. 593–596). IEEE. <https://doi.org/10.1109/ICACTM.2019.8776800>
- Sovietov, P., & Gorchakov, A. V.** (2022). Digital Teaching Assistant for the Python Programming Course. In *2022 2nd International Conference on Technology Enhanced Learning in Higher Education (TELE)*, (pp. 272–276). IEEE. <https://doi.org/10.1109/tele55498.2022.9801060>
- Symeonidis, S., Effrosynidis, D., & Arampatzis, A.** (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems With Applications*, 110, 298–310. <https://doi.org/10.1016/j.eswa.2018.06.022>
- Trigueros, O., Blanco, A. G., Lebeña, N., Casillas, A., & Pérez, A. G.** (2022). Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. *International Journal of Medical Informatics*, 157, 104615. <https://doi.org/10.1016/j.ijmedinf.2021.104615>
- Umar, N., & Nur, N. M.** (2022). Application of Naïve Bayes Algorithm Variations On Indonesian General Analysis Dataset for Sentiment Analysis. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 6(4), 585–590. <https://doi.org/10.29207/resti.v6i4.4179>
- Uzun, E., Yerlikaya, T., & Kirat, O.** (2018). Comparison of python libraries used for web data extraction. *Journal of the Technical University at Plovdiv*, 24, 87–92.
- Verma, A., Pal, S., & Kumar, S.** (2020). Prediction of Skin Disease Using Ensemble Data Mining Techniques and Feature Selection Method—a Comparative Study. *Applied Biochemistry and Biotechnology*, 190(2), 341–359. <https://doi.org/10.1007/s12010-019-03093-z>
- Vidhya S., Singh, D. A. A., & Leavline, E. J.** (2016). Feature Extraction for Document Classification. *International Journal of Innovative Research in Science, Engineering and Technology*, 4(6), 50–56.
- Wongso, R., Luwinda, F. A., Trisnajaya, B. C., & Rusli, O.** (2017). News Article Text Classification in Indonesian Language. *Procedia Computer Science*, 116, 137–143. <https://doi.org/10.1016/j.procs.2017.10.039>
- Yang, D., Kim, B., Lee, S., Ahn, Y. C., & Kim, H.** (2022). AutoDefect: Defect text classification in residential buildings using a multi-task channel attention network. *Sustainable Cities and Society*, 80, 103803. <https://doi.org/10.1016/j.scs.2022.103803>
- Yao, L., Pengzhou, Z., & Chi, Z.** (2019). Research on News Keyword Extraction Technology Based on TF-IDF and TextRank. In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, (pp. 452–455). IEEE. <https://doi.org/10.1109/icis46139.2019.8940293>

Editorial record: The article has been peer-reviewed. First submission received on 28 October 2022. Revisions received on 26 January 2023 and 20 February 2023. Accepted for publication on 26 February 2023. The editor in charge of coordinating the peer-review of this manuscript and approving it for publication was Stanislav Vojir .

Acta Informatica Pragensia is published by Prague University of Economics and Business, Czech Republic.

ISSN: 1805-4951
