Discover Artificial Intelligence

Discover

Research

Instructor-assisted question classification system using machine learning algorithms with N-gram and weighting schemes

Delali Kwasi Dake[1] · Edward Nwiah[1] · Griffith Selorm Klogo[2] · Wisdom Xornam Ativi[3]

Received: 11 February 2023 / Accepted: 28 June 2023
Published online: 08 August 2023

Abstract
One aspect of natural language processing, text classification, has become necessary in the educational domain due to the increasing number of students and the COVID-19 outbreak. The advent of the devastating pandemic and the need to remain safe have surged the discussions around online learning and integrated modules in teaching and learning. In this study, we employed machine learning to develop an automatic instructor-assisted question classification module for learning management systems. In selecting the best classifier, the conventional and the ensemble machine learning algorithms were compared using the tenfold and the fivefold cross-validation techniques. In addition, the N-gram feature selection mechanism and three weighting schemes were evaluated for performance enhancement. The detailed analysis indicates that the ensemble algorithms outperform the conventional ones with decreasing accuracy as the N-gram size increases. For all compared algorithms, the AdaBoost (SVM) ensemble algorithm has the highest accuracy of 78.55% for Unigram (TP, TF, TF-IDF). In addition, the AdaBoost (SVM) emerged with the highest F1-score of 0.782, whiles the ensemble Bagging (RF) algorithm had the highest ROC value of 0.955 for Unigram (TP).

Keywords Supervised algorithms · Ensemble algorithms · Natural language processing · Text classification

1 Introduction

The COVID-19 pandemic remains one of the most significant tragic outbreaks that has necessitated numerous educational reforms globally [32]. The sudden closure of tertiary institutions in April 2020 affected schools in 185 countries, with 89.4% of total learners involved [40]. As countries battle to contain the spread of the virus and save lives, in-person teaching and learning was suspended indefinitely. The pandemic revealed flaws in the current educational system and created an opportunity for proactive policies in the eLearning space [32]. Even as institutions migrated hastily to online learning and distance education during the pandemic, it came with infrastructure, assessment, technological, pedagogical and financial challenges [40]. As the pandemic effect decreases in 2022 and learner enrollment increases, the eLearning adoption issues in higher education institutions remain precarious [19]. In June 2022, the UNESCO Institute of Statistics data indicated that the total number of tertiary students worldwide has doubled in the last two decades. There was a staggering 240% increase in tertiary enrollment in South and West Asia, East Asia and the Pacific from 2000 to 2020, while Central and Eastern Europe learner enrollment shot up by 84% within the same year. Sub-Saharan African countries averagely, had a 9.4% enrollment increase, with Mauritius alone having the highest gross tertiary enrollment of 40% by 2020.

✉ Delali Kwasi Dake, dkdake@uew.edu.gh; Edward Nwiah, enwiah@uew.edu.gh; Griffith Selorm Klogo, klogoselorm@yahoo.com; Wisdom Xornam Ativi, ativiw@yahoo.com | [1]University of Education, Winneba, Ghana. [2]Kwame Nkrumah University of Science and Technology, Kumasi, Ghana. [3]University of Electronics Science and Technology of China (UESTC), Chengdu, China.

Within the same year range, Ghana and Togo have an enrollment estimate of 15%, while Niger's estimate is 4.4% [41]. The rising enrollment numbers have emphasised the importance of tertiary institutions having robust infrastructure support and diverse online application modules to meet successful learning outcomes during and after the pandemic.

E-learning comprises the internet or web-based education that uses digital resources and involves interactive activities [13]. The proliferation of e-learning in tertiary institutions has increased because of internet availability, affordable computing devices, multimedia integration and well-built learning management systems (LMS) [21]. Furthermore, e-learning is adaptable to various circumstances, including the pandemic, geographical location, and time. E-learning also improves learner collaboration, personalised learning and analytics due to the ease of data generation [24]. LMS, a core aspect of e-learning, is a web-based technology that handles significant online teaching and learning elements. The open-source LMS, including Moodle, Schoology, Odoo, Sakai and Totara, are accessible at no cost to users. In contrast, proprietary LMS such as Blackboard, Saba, Litmos and Topyx are vendor lock-in with subscriptions. LMS provides basic functionality, including course administration, assessment management, video conferencing and asynchronous learning to advanced features such as gamification, data tracking, intelligent scheduling and offline learning tracking. Integrating machine intelligent modules in LMS to detect hidden patterns has become profound due to the amount of data generated online [5].

Machine Learning (ML) has seen integration in diverse areas significantly. In agriculture, ML has helped farmers to detect the best time to plant, fruit detection and classification, and prevent post-harvest negligence [22]. In a smart grid, ML has been used to predict load peak hours, sudden fluctuations in power output, network intrusion detection and future optimum scheduling of customer demands [16]. The area of healthcare has seen the application of ML algorithms in clinical research, personalised medicine, disease identification and viral outbreak predictions [4]. In manufacturing, transportation and industry, ML integration has shown major benefits in automatic machine translation, driverless vehicles, virtual assistants, product malfunction detection, predictive maintenance, decision support and responsive customer experience [31, 46, 48]. In the retail industry, ML algorithms has seen usage in customer behaviour prediction, intelligent stocking and inventory, demand forecasting, and pricing optimisation [45]. Social media internet-based platforms currently utilise ML for personalised advertisement, fraudulent account detection, emotion detection behind sentiments, and early detection of hate speech [6]. In the educational sector, ML has seen significant usage in learner performance prediction, assessment modelling, student drop-out prediction, learner groupings, instructor feedback polarity detection, and learner behaviour modelling [3, 35].

2 Problem definition

The global economy, including education, has still not recovered from the devastating COVID-19 pandemic [9, 27]. Even as countries continue to spring out novel policies to minimise the spread of the virus and rejuvenate the economy, student enrollment across tertiary institutions is still on the rise [41]. Most institutions globally have migrated to online and distance learning paradigms to limit physical contact and curtail the spread of the virus [28, 34]. Traditional education has excellent advantages, including face-to-face meetings with instructors, instant feedback from the instructor when questions are asked, learner monitoring, sociability and solidarity, effective practical sessions, and comprehensive support services [17, 36]. Even as the trade-off between online and traditional education continues to improve via blending learning [11], the instructors' ability to address all learner questions and emotions during the limited class sessions remains an issue. Even worse is the repetition of questions or similar questions that the instructor has already addressed. In constructivism and personalised learning, learners are expected to be active in the learning process and deduce their knowledge [2, 14]. As learners construct new knowledge, continuous interaction with the instructor is relevant in shaping the contents learners are exposed to in the active process.

2.1 Purpose of study

In solving the need for a responsive instructor to the seemingly increasing number of students across tertiary institutions, the study proposes an automatic instructor-assisted question classification system using machine learning algorithms. In addition, we tested the relevance of N-gram and weighting schemes in building the classification model. In line with the purpose of the study, we pose the following research questions:

RQ1. Which machine learning algorithm has the highest accuracy in developing an instructor-assisted question classification system?
RQ2. What other significant machine learning metrics contributed to selecting the best classifier for future prediction?

The rest of the paper is organised as follows: a brief introduction of natural language processing and question classification, a review of related literature, research methods, results and analysis, discussion and findings, conclusion, and future work.

3 Natural language processing and question classification

Natural Language Processing (NLP) is the basis for information retrieval in text and covers various application areas in computer science and artificial intelligence. With computational techniques, NLP produces human language content and abstraction by relying on mostly unstructured data [43]. Instead of encoding words as indices in dictionaries, NLP currently represents words as continuous vector forms with much lower dimensions and semantic similarities [38]. In online learning, students have the impersonal opportunity devoid of tension to ask diverse questions via text. The text using NLP can be segmented for the instructor to develop an automatic response system or address questions based on importance. The instructor, primarily via segmentation, has a labelled data set for context understanding and predictive modelling using machine learning or rule-based systems. A well-built model with high accuracy becomes a crucial basis for the instructor to check the course manual, recommended books, methods of teaching, and the curriculum for optimal learning.

4 Review of literature

Text categorisation and identification require question classification using machine learning algorithms [51] or rule-based systems [20]. Ehrentraut et al. [8] stated three stages as the basics of a question-answering system. In the first phase, known as question processing, students pose questions that act as input for categorisation and representation. The next step entails selecting documents and paragraphs based on text concentration. The final stage involves matching answers based on the questions' similarity with appropriate responses to the learner. Rule-based and machine learning methods are the two most common ways to construct a question classification system. The rule-based approach infers a knowledge base to detect text relations and provide answers based on a model. In contrast, the machine learning method utilises a dataset of labelled answers trained to extract features and predict the future categorisation of input questions. This review is limited to text and question classification using deep learning, ensemble and conventional machine learning techniques.

Zulqarnain et al. [51] analysed Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Networks (CNN), CNN-GRU and CNN-LSTM on the Turkish questions dataset. After applying the tenfold cross-validation technique, skip-gram, Word2vec and CBOW, CNN performed with the highest accuracy of 93.7%, using 300 features and skip-gram. Zhou et al. [50] created a hybrid neural network for a question classification system by combining CNN and BiLSTM. They implemented the fivefold cross-validation technique on Badu Knows and TREC benchmarks databases and compared results with SVM, LSTM and MaxEnt algorithms. Results show that the CNN-BiLSTM outperforms the other algorithms for Precision, Recall and F1 score metrics. Gaire et al. [10] compared their proposed RNN-LSTM with Multinomial Naïve Bayes (MNB), K-Nearest Neighbor (KNN) and Logistic Regression (LR) traditional machine learning algorithms. Using the Quora database, the study aimed to distinguish between sincere and insincere questions. The RNN-LSTM deep learning algorithms achieve a dominant Precision value of 0.654, Recall value of 0.733, and F1 score of 0.691. Zhen and Sun [49] presented a bagging-based CNN model using the Word2vec to map word features to proportional dimensions. They compared the model to Bayes and SVM conventional classification algorithms using bag-of-words, mutual information and IF-IDF methods. The results show that the W2V + B-CNN algorithm has a higher classification accuracy of 88.57% on the Chinese dataset from Harbin Institute of Technology's Information Retrieval Laboratory. Lei et al. [18] proposed an RR-CNN-based architecture for question classification by increasing the generalisation of sentence features. They used five datasets to compare their results to conventional SVM, NB, MNB, and other CNN-based approaches. The RR-CNN achieves the highest accuracy of 86.3% on average compared to tested algorithms. Upadhya

et al. [42] applied an LSTM model to the Amazon community question dataset to identify the questions as yes/no or opinion-related. They discovered a 29%, 31%, and 23% difference in accuracy between their LSTM model and CNN for the dataset components electronics, beauty, and appliances, respectively. Razzaghnoori et al. [33] utilised three feature extraction methods, including clustering, feedforward neural network (FNN) and RNN, to design a question classification system. After comparing the LSTM to other neural networks and SVM utilising the aforementioned feature extraction techniques, an improved accuracy of 81.77% was realised.

Madabushi and Lee [20] compared the NB and SVM conventional algorithms for question classification in a network-based online course. They identified 1120 keywords in building a predictive model. SVM performed better with higher values for Precision and Recall. Mohasseb et al. [23], in a similar study, implemented SVM, RF, DT and NB algorithms using a grammar-based approach instead of the bag-of-words or dictionary standard techniques. The dataset comprised 1,160 questions from a pool of Wikipedia data, TREC 2007 question answer data and yahoo non-factored questions. The DT classifier emerged as the best algorithm with an accuracy of 90.1% and superior Precision, Recall and F-score values. Hassan et al. [12] proposed a model comparatively using Gaussian Naïve Bayes (GNB), SVM, MNB, LR, RF and KNN. They used accuracy, precision, recall, and f1-score as performance metrics to determine the best model. Termed future engineering, they utilised bag-of-words and TF-IDF before building the final classifier. Results for classification show that the LR had an accuracy of 85.8% from the IMDB dataset, while the KNN performed best with an accuracy of 98.5% on the Spam dataset. Kamath et al. [15] presented a text classification model and compared the performance of the conventional NB, LR, SVM, MLP, and RF to CNN deep learning algorithm. They utilised the tobacco-3482 dataset and a health dataset from a public institution. Even though the LR algorithm had the highest accuracy of 77% among the conventional algorithms, the CNN performed with a superior accuracy of 83%. Yadav et al. [47] trained LR, SVM, NB and RF for text categorisation using conventional machine learning to ANN. The accuracy, precision, recall and F1-score analytical metrics formed the bases for comparing the algorithms after pre-processing. Results show a superior classification accuracy for ANN. Onan [25] compared conventional, ensemble and deep learning machine learning algorithms for sentiment analysis in massive open online courses (MOOCs). Simulation results show a dominant performance of deep learning methods over ensemble techniques. The ensemble methods also performed better than the conventional machine learning algorithms.

5 Research methodology

The research methodology is divided into four sub-sections: the proposed flow diagram, the dataset, the feature selection mechanisms, metrics for measuring performance and the algorithms implemented.

5.1 Proposed flow diagram

Figure 1 depicts the flow diagram, which begins with the manual categorisation of textual questions from students. In this phase, we compile questions on the various components of the course. Tokenisers are used before data pre-processing to break down the unstructured questions without categories into chunks of data. During data pre-processing, affixes are removed from the data through stemming. In addition, common words called stop words and punctuation marks are equally removed from the text. The feature selection module consists of N-gram-sized phrases, including sizes 1, 2, and 3 representing unigram, bigram and trigram, respectively. Term Presence (TP), Term Frequency (TF), and Term Frequency and Inverse Document Frequency (TF-IDF) are then used to calculate the weight of terms in the text. The text and respective class categories are trained using conventional ML algorithms and ensemble ML techniques. The optimal classifier is then determined using classification metrics from the implemented fivefold and tenfold cross-validation and feature selection mechanisms.

5.2 The dataset

The questions for the dataset were collected continuously during the Database Management Course at the University of Education, Winneba, Ghana, in 2022. The dataset contains 1096 textual questions that are labelled under seven categories. As shown in Fig. 2, the seven categories include Teaching–Learning Material (TLM), Attendance, Assessment, Course Manual (Manual), Practical, Theory, and Any Other Question (AOQ). The TLM refers to the resources and tools the instructor used during the database course. Attendance relates to the online and physical presence of learners during lesson periods. The Course Manual (Manual) is the primary document for lesson organisation, classroom policies,

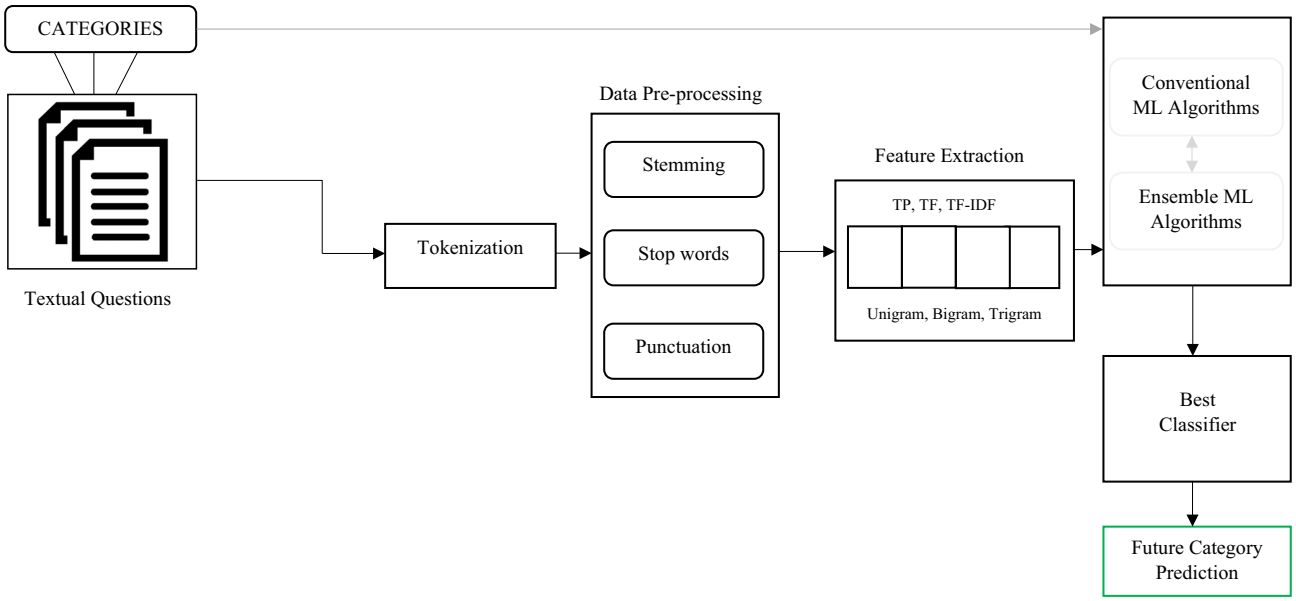


Fig. 1 Proposed flow diagram

Fig. 2 Category distribution of questions

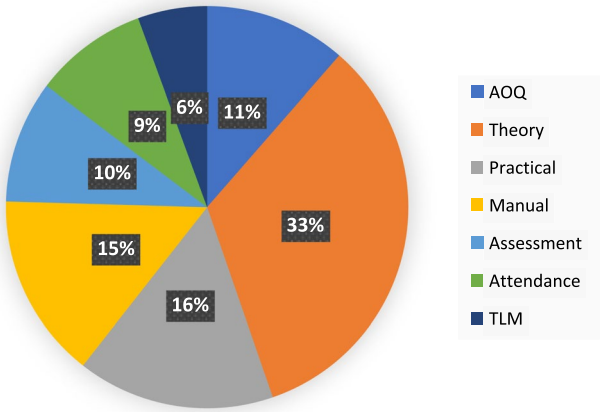


Table 1 Sample questions and respective categories

Category	Students question
Practical	It's an interesting course but more practical sessions needed
Assessment	Please, does participation online and in-class attract some points
TLM	How do I get the course-required textbooks and literature
AOQ	Will I be taking this course later in higher levels
Manual	What inspirations can I draw from this course
Theory	Can I get more materials to learn about data isolation and concurrences
Attendance	How flexible is the attendance policy, sir?

content, recommended materials and grading. The practical aspect of learner questions refers to the experimentation and implementation of concepts, while the theoretical aspects include the explanation and principles of the database course. Any Other Question (AOQ) category consists of questions unrelated to the other six categories.

Theory-based questions had the highest percentage labelling of 33. The practical questions frequency was 16%, followed by 15% for the course manual, 11% for AOQ, 10% for Assessment, and 9% for Attendance, with TLM having a minor count at 6%. As depicted in Table 1, questions are carefully matched with the correct categories.

5.3 Feature selection mechanism

In NLP, feature selection involves the selection of a subset of occurring terms in the training data. In the question classification system, feature selection primary is for two purposes: eliminating noisy words to increase classification accuracy and decreasing the training data size for efficiency. One feature mechanism utilised in the study is the N-gram model. An N-gram is a continuous sequence of tokens in a textual document. The N-gram feature technique is useful for predicting the next probable word in a sequence. The research utilised the unigram (N = 1), the bigram (N = 2) and the trigram (N = 3) as inputs during training. One aspect of the feature selection mechanism is the weight of a word in the question. After tokenisation, the Bag-of-Words (BOW) feature extraction method for word counting is implemented using TF, TP and TF-IDF weighting schemes. With TF-based weighting, the number of times a word occurs in the question are computed. In contrast, the TF-IDF scheme measures rare words, whereas the TP considers the existence of a word in the question.

5.4 Classifier performance evaluation metrics

Aside from using classification accuracy as an evaluation metric, the study examined revealing metrics including precision, recall, F-measure, and the Receiver Operating Characteristics (ROC-AUC) in comparing the performance of the algorithms. As shown in Eq. 1, classification accuracy influences the proportion of accurate predictions. It measures the proportion of accurate predictions to all predictions.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{1}$$

TN, TP, FN and FP denote the number of true negatives, true positives, false negatives and false positives, respectively.

Precision metric identifies only the true positives. As shown in Eq. 2, the precision is the ratio of the number of true positives to the number of true positives plus the number of false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall metric determines only the relevant positives. As shown in Eq. 3, the recall is the ratio of true positives to the true positives plus the false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

The F-measure is a comparative metric for the performance of classifiers. As depicted in Eq. 4, the F-measure is the harmonic mean of precision and recall.

$$\text{F} - \text{measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The ROC curve is a probability performance measurement metric that identifies classification difficulties at different thresholds. The ROC curve is a graph that plots True Positive Rate (TPR) against False Positive Rate (TPR). Area Under the Curve (AUC) determines the degree of separability between classes. The greater the AUC value, the better the classifier can differentiate between the specified classes.

5.5 Implemented machine learning algorithms

The algorithms utilised in this work fall into two primary categories: supervised learning or conventional traditional machine learning algorithms and multi-classifier ensemble algorithms.

5.5.1 Supervised learning algorithms

Supervised learning methods infer prediction based on a labelled dataset. The classifier generated from classification algorithms has input instances and matching class outputs. The primary objective of a supervised algorithm is to approximate a mapping function during training for future prediction [30].

The Random Forest (RF) algorithm constructs many decision trees and combines them to get a more precise and steady prediction. While splitting a node, the RF algorithm analyses a random subset of features for the most significant feature and generates a more accurate prediction model [26].

The Decision Tree (DT) algorithm is a tree-structured classifier in which branches represent decision rules, internal nodes represent dataset features, and leaf nodes represent class labels. In a DT algorithm, the decision nodes lead to multiple branches, while the leaf nodes are the outputs that cannot have branches [29].

The Support Vector Machine (SVM) plots each data item as a point in an n-dimensional environment where n represents the number of attributes. Each coordinate in an SVM plot matches the value of an attribute with a hyper-plane to differentiate the class labels. The hyper-plane segregates the classes and coordinates respective observations [26].

The K-nearest neighbor (KNN) algorithm is a non-parametric classifier that uses proximity to classify and cluster data points. The KNN uses similarity assumption to categorise similar data variables. The KNN is often referred to as a lazy learner algorithm because it only learns from the training data during classification [7].

5.5.2 Ensemble algorithms

Ensemble learning, primarily a multi-classifier system, combines several machine learning models in the prediction process [37]. The ensemble method uses a base-learner and an inducer algorithm to train the dataset. This meta approach of utilising a base-learner and an inducer algorithm seeks to produce an optimal predictive classifier [39]. The base-learner is referred to as homogeneous ensembles when a single base learning algorithm is employed and as heterogeneous ensembles when multiple learning algorithms are used [39].

Bagging ensemble is a parallel classifier that combines bootstrap and aggregation. The bootstrapping functions randomly select and replace training data instances, which helps reduce data variance and prevent overfitting [1].

Adaboost is a boosting-based classifier that decreases bias and variance by converting weaker learners to strong learners. Before classification, the boosting classifier assigns equal weight to all data points. After classification, the weight of instances incorrectly classified is increased while the weight of instances correctly classified is decreased [44].

Random Subspace algorithm combines predictions from various decision trees trained on distinct subsets of the training dataset. In a Random Subspace classifier, diversity is introduced by varying the columns used to train instances of the ensemble randomly [44].

6 Results and analysis

This section presents experimentation and training in Weka using supervised learning and ensemble classifiers. The tenfold and fivefold cross-validation techniques are implemented to resample and evaluate the algorithms through training, testing, and validation.

6.1 Tenfold cross-validation experiments

The tenfold cross-validation utilised a 90–10% training and testing ratio iteratively on the dataset. As shown in Table 2, sixteen machine learning algorithms were compared for accuracy using the Unigram, Bigram, and Trigram N-gram sequence under three text weighting schemes TP, TF and TF-IDF.

The results from classification show a dominant performance for AdaBoost (SVM) ensemble machine learning algorithm under all three weighting schemes using Unigram. The conventional SVM still performed with the highest accuracy under Unigram compared to the other traditional machine learning algorithms. The classification accuracy decreased to low levels under Bigram and Trigram feature selection methods. The AdaBoost (KNN) was the worst-performing classifier in terms of accuracy using the tenfold cross-validation technique, as illustrated in Table 2.

The F-measure score, the harmonic mean of precision and recall, is depicted in Table 3. Adaboost (SVM) ensemble algorithm under Unigram for the three weighting schemes has the highest F1 score of 0.782, indicating a good classification

Table 2 Classification accuracy of machine learning algorithms

	Unigram + TP	Unigram + TF	Unigram + TF-IDF	Bigram + TP	Bigram + TF	Bigram + TF-IDF	Trigram + TP	Trigram + TF	Trigram + TF-IDF
RF	72.99	72.81	73.08	61.49	60.85	61.22	50.27	51.09	50.72
DT	68.15	68.15	68.15	60.31	60.03	60.31	47.35	47.35	47.35
KNN	68.33	68.33	67.88	47.90	46.35	47.90	38.22	38.22	38.22
SVM	78.46	78.46	78.46	69.34	68.06	69.34	54.37	54.37	54.37
Bagging (RF)	72.81	72.53	72.08	59.21	57.66	58.85	47.71	47.71	47.71
Bagging (DT)	70.52	70.52	70.52	60.85	59.85	60.85	48.44	48.44	48.44
Bagging (KNN)	68.15	68.15	68.15	47.90	46.80	48.17	37.04	37.04	37.04
Bagging (SVM)	76.27	76.27	76.27	69.34	63.22	64.23	49.36	49.36	49.36
AdaBoost (RF)	68.70	70.34	69.98	59.85	59.03	59.58	48.26	47.81	48.17
AdaBoost (DT)	74.08	74.08	74.08	63.13	61.67	63.13	47.35	47.35	47.35
AdaBoost (KNN)	68.97	68.97	68.97	48.44	46.80	48.44	36.31	36.31	36.31
AdaBoost (SVM)	78.55	78.55	78.55	47.62	68.15	69.34	54.47	54.47	54.47
RandomSubSpace (RF)	73.99	73.26	72.99	61.40	61.49	61.86	45.80	46.25	46.16
RandomSubSpace (DT)	70.98	70.43	71.35	57.93	57.57	57.66	45.62	45.80	45.71
RandomSubSpace (KNN)	68.61	67.51	68.24	47.90	50.45	51.64	39.96	40.60	39.87
RandomSubSpace (SVM)	76.45	74.72	76.18	64.96	65.05	64.23	48.63	48.81	48.63

Table 3 F-measure score obtained by the machine learning algorithms

	Unigram + TP	Unigram + TF	Unigram + TF-IDF	Bigram + TP	Bigram + TF	Bigram + TF-IDF	Trigram + TP	Trigram + TF	Trigram + TF-IDF
RF	0.715	0.713	0.716	0.577	0.567	0.572	0.444	0.456	0.448
DT	0.677	0.677	0.677	0.580	0.574	0.580	0.232	0.232	0.232
KNN	0.662	0.662	0.671	0.417	0.391	0.417	0.213	0.201	0.201
SVM	0.781	0.781	0.781	0.675	0.662	0.675	0.499	0.499	0.499
Bagging (RF)	0.713	0.707	0.703	0.547	0.526	0.543	0.408	0.408	0.408
Bagging (DT)	0.695	0.695	0.695	0.580	0.568	0.580	0.406	0.406	0.406
Bagging (KNN)	0.662	0.662	0.662	0.417	0.394	0.419	0.214	0.200	0.200
Bagging (SVM)	0.757	0.757	0.757	0.675	0.598	0.606	0.428	0.428	0.428
AdaBoost (RF)	0.664	0.683	0.678	0.555	0.544	0.552	0.403	0.399	0.406
AdaBoost (DT)	0.737	0.737	0.737	0.613	0.595	0.613	0.203	0.210	0.210
AdaBoost (KNN)	0.671	0.671	0.671	0.427	0.400	0.427	0.223	0.212	0.212
AdaBoost (SVM)	0.782	0.782	0.782	0.415	0.662	0.675	0.495	0.495	0.495
RandomSubSpace (RF)	0.727	0.720	0.716	0.580	0.580	0.585	0.387	0.394	0.392
RandomSubSpace (DT)	0.699	0.696	0.705	0.532	0.527	0.528	0.210	0.233	0.233
RandomSubSpace (KNN)	0.663	0.654	0.658	0.417	0.446	0.462	0.201	0.305	0.295
RandomSubSpace (SVM)	0.758	0.739	0.755	0.622	0.622	0.614	0.425	0.427	0.422

Table 4 Receiver characteristics curve value obtained by the machine learning algorithms

	Unigram + TP	Unigram + TF	Unigram + TF-IDF	Bigram + TP	Bigram + TF	Bigram + TF-IDF	Trigram + TP	Trigram + TF	Trigram + TF-IDF
RF	0.949	0.951	0.950	0.918	0.915	0.920	0.841	0.839	0.841
DT	0.851	0.851	0.851	0.831	0.826	0.831	0.648	0.648	0.648
KNN	0.838	0.838	0.853	0.729	0.721	0.729	0.608	0.608	0.608
SVM	0.928	0.928	0.928	0.866	0.864	0.866	0.766	0.766	0.766
Bagging (RF)	0.955	0.954	0.954	0.920	0.916	0.920	0.827	0.829	0.828
Bagging (DT)	0.930	0.930	0.930	0.875	0.874	0.875	0.728	0.728	0.728
Bagging (KNN)	0.872	0.872	0.872	0.729	0.743	0.746	0.597	0.597	0.597
Bagging (SVM)	0.939	0.939	0.939	0.866	0.869	0.874	0.758	0.758	0.758
AdaBoost (RF)	0.883	0.879	0.879	0.839	0.830	0.844	0.743	0.736	0.745
AdaBoost (DT)	0.938	0.938	0.938	0.882	0.876	0.882	0.648	0.648	0.648
AdaBoost (KNN)	0.801	0.801	0.801	0.628	0.602	0.628	0.583	0.583	0.583
AdaBoost (SVM)	0.903	0.903	0.903	0.621	0.801	0.802	0.772	0.772	0.772
RandomSubSpace (RF)	0.952	0.949	0.952	0.920	0.920	0.921	0.862	0.867	0.863
RandomSubSpace (DT)	0.936	0.930	0.934	0.861	0.855	0.856	0.685	0.675	0.666
RandomSubSpace (KNN)	0.917	0.915	0.919	0.729	0.855	0.855	0.772	0.781	0.763
RandomSubSpace (SVM)	0.944	0.939	0.942	0.899	0.897	0.898	0.810	0.805	0.805

compromise between precision and recall. The traditional SVM algorithm under Unigram performed as the second best classifier with an F1 score of 0.781.

The ROC value indicates a trade-off between specificity and sensitivity. The resulting ROC value shows the classifier's ability to differentiate between the seven classes from the confusion matrix correctly. As illustrated in Table 4, Bagging (RF) under Unigram (TP) has the highest ROC value of 0.955. For Bigram and Trigram, the traditional RF algorithm and the RandomSubSpace (RF) performed equally well.

6.2 Fivefold cross-validation experiments

The same experiments were repeated using the fivefold cross-validation technique to compare and ascertain classifier performance. As depicted in Table 5, ensemble AdaBoost (SVM) still performed with the highest accuracy under Unigram (TP, TF, TF-IDF). The conventional SVM algorithm follows as the second-best classifier in terms of accuracy under the same weighting schemes in Unigram. The AdaBoost (KNN) is the worst-performing classifier under the Trigram feature selection mechanism. The performance of the fivefold cross-validation approach is consistent with that of the tenfold cross-validation technique for the best and worst performing algorithms.

As shown in Table 6, the AdaBoost (SVM) ensemble machine learning algorithm under Unigram (TP, TF, TF-IDF) performed with the highest F1 score of 0.769, whiles the conventional SVM performed with a score of 0.766.

As depicted in Table 7, the Unigram (TF, TF-IDF) of the Bagging (RF) ensemble machine learning algorithm has the highest ROC value of 0.951.

6.3 Comparative analysis, accuracy

The classification results for machine learning metrics accuracy, F score and the ROC value show a compelling trend between the conventional machine learning algorithms under the tenfold and the fivefold cross-validation techniques. This aspect of the analysis focuses on the best feature selection mechanism, Unigram, with respective weighting schemes (TP, TF, TF-IDF) under the cross-validation techniques.

RQ1: Which machine learning algorithm has the highest accuracy in developing an automatic instructor-assisted question classification system?

6.3.1 Comparative accuracy for the conventional algorithms

As depicted in Fig. 3a and b, for the conventional machine learning algorithm, the SVM from the tenfold cross-validation has the highest accuracy of 78.46 compared to 76.91 for the fivefold SVM.

6.3.2 Comparative accuracy for the ensemble algorithms

As depicted in Fig. 4a and b, for the ensemble machine learning algorithm, AdaBoost (SVM) from the tenfold cross-validation has the highest accuracy of 78.55 compared to 77.18 for the fivefold AdaBoost (SVM).

6.3.3 Summary of accuracy

In response to **RQ1**, as shown in Fig. 5, the ensemble AdaBoost (SVM) has the highest accuracy of 78.55 compared to 78.46 for the conventional SVM using the tenfold cross-validation technique.

6.4 Comparative analysis, F-measure

The F-measure is the weighted average of precision and recall. The F-score metric is essential since the text classification generated balanced data between precision and recall.

RQ2: What other significant machine learning metrics contributed to selecting the best classifier for future prediction?

Table 5 Classification accuracy of machine learning algorithms

	Unigram + TP	Unigram + TF	Unigram + TF-IDF	Bigram + TP	Bigram + TF	Bigram + TF-IDF	Trigram + TP	Trigram + TF	Trigram + TF-IDF
RF	71.71	71.89	71.35	59.03	60.12	60.94	48.26	47.99	48.81
DT	67.88	67.88	67.88	57.93	59.85	57.93	46.98	46.98	46.98
KNN	67.24	67.24	67.24	47.08	45.89	47.08	37.40	37.40	37.40
SVM	76.91	76.91	76.91	67.88	66.78	67.88	53.10	53.10	53.10
Bagging (RF)	70.80	70.34	70.89	56.47	56.38	56.66	45.80	45.62	45.43
Bagging (DT)	69.61	69.61	69.61	59.48	59.94	59.48	47.90	47.90	47.90
Bagging (KNN)	67.33	67.33	67.33	47.08	45.52	47.99	38.13	38.13	38.13
Bagging (SVM)	75.72	75.72	75.72	67.88	62.22	61.95	47.99	47.99	47.99
AdaBoost (RF)	69.06	67.97	69.25	57.48	56.84	57.29	46.53	46.16	46.53
AdaBoost (DT)	74.17	74.17	74.17	60.40	61.58	60.40	46.98	46.98	46.98
AdaBoost (KNN)	67.88	67.88	67.88	47.62	46.25	47.62	34.30	34.30	34.30
AdaBoost (SVM)	77.18	77.18	77.18	67.60	67.60	67.60	52.28	52.2	52.28
RandomSubSpace (RF)	73.08	71.44	71.98	59.30	58.66	58.12	42.97	44.98	45.07
RandomSubSpace (DT)	71.25	70.80	71.35	55.47	56.75	56.11	44.79	46.16	45.16
RandomSubSpace (KNN)	67.97	66.69	66.87	47.08	49.54	47.81	38.59	39.23	40.05
RandomSubSpace (SVM)	75.91	74.63	75.00	62.22	62.31	61.58	47.08	47.99	47.44

Table 6 F-measure score obtained by the machine learning algorithms

	Unigram + TP	Unigram + TF	Unigram + TF-IDF	Bigram + TP	Bigram + TF	Bigram + TF-IDF	Trigram + TP	Trigram + TF	Trigram + TF-IDF
RF	0.701	0.702	0.698	0.548	0.562	0.570	0.415	0.408	0.424
DT	0.671	0.754	0.671	0.542	0.571	0.542	0.396	0.396	0.396
KNN	0.651	0.651	0.651	0.406	0.386	0.406	0.212	0.210	0.210
SVM	0.766	0.766	0.766	0.660	0.647	0.660	0.481	0.481	0.481
Bagging (RF)	0.689	0.685	0.689	0.512	0.509	0.514	0.234	0.376	0.376
Bagging (DT)	0.683	0.683	0.683	0.560	0.566	0.560	0.405	0.405	0.405
Bagging (KNN)	0.654	0.654	0.654	0.406	0.377	0.418	0.210	0.209	0.209
Bagging (SVM)	0.752	0.752	0.752	0.660	0.586	0.580	0.403	0.403	0.403
AdaBoost (RF)	0.669	0.657	0.670	0.528	0.517	0.524	0.379	0.377	0.383
AdaBoost (DT)	0.737	0.737	0.737	0.579	0.595	0.579	0.396	0.396	0.396
AdaBoost (KNN)	0.659	0.659	0.659	0.415	0.392	0.415	0.233	0.210	0.210
AdaBoost (SVM)	0.769	0.769	0.769	0.656	0.656	0.656	0.469	0.469	0.469
RandomSubSpace (RF)	0.718	0.698	0.705	0.551	0.543	0.541	0.343	0.373	0.210
RandomSubSpace (DT)	0.704	0.694	0.705	0.499	0.519	0.510	0.358	0.376	0.376
RandomSubSpace (KNN)	0.654	0.642	0.644	0.406	0.433	0.413	0.273	0.210	0.210
RandomSubSpace (SVM)	0.754	0.739	0.744	0.589	0.590	0.579	0.399	0.410	0.210

Table 7 Receiver characteristics curve value obtained by the machine learning algorithms

	Unigram + TP	Unigram + TF	Unigram + TF-IDF	Bigram + TP	Bigram + TF	Bigram + TF-IDF	Trigram + TP	Trigram + TF	Trigram + TF-IDF
RF	0.946	0.943	0.947	0.909	0.914	0.912	0.821	0.821	0.821
DT	0.853	0.878	0.853	0.829	0.830	0.829	0.648	0.648	0.648
KNN	0.830	0.830	0.830	0.718	0.707	0.718	0.594	0.594	0.594
SVM	0.921	0.921	0.921	0.857	0.861	0.857	0.755	0.755	0.755
Bagging (RF)	0.950	0.951	0.951	0.912	0.911	0.912	0.817	0.820	0.817
Bagging (DT)	0.925	0.925	0.925	0.870	0.878	0.870	0.724	0.724	0.724
Bagging (KNN)	0.866	0.866	0.866	0.718	0.744	0.741	0.597	0.597	0.597
Bagging (SVM)	0.933	0.933	0.933	0.857	0.869	0.866	0.755	0.755	0.755
AdaBoost (RF)	0.873	0.875	0.877	0.832	0.813	0.828	0.723	0.725	0.717
AdaBoost (DT)	0.941	0.941	0.941	0.876	0.873	0.876	0.648	0.648	0.648
AdaBoost (KNN)	0.794	0.794	0.794	0.621	0.601	0.621	0.579	0.579	0.579
AdaBoost (SVM)	0.900	0.900	0.900	0.802	0.802	0.802	0.747	0.747	0.747
RandomSubSpace (RF)	0.948	0.947	0.948	0.911	0.914	0.916	0.854	0.858	0.855
RandomSubSpace (DT)	0.932	0.930	0.934	0.845	0.851	0.845	0.685	0.661	0.659
RandomSubSpace (KNN)	0.911	0.914	0.917	0.718	0.625	0.850	0.766	0.764	0.766
RandomSubSpace (SVM)	0.939	0.937	0.939	0.887	0.886	0.887	0.798	0.795	0.788

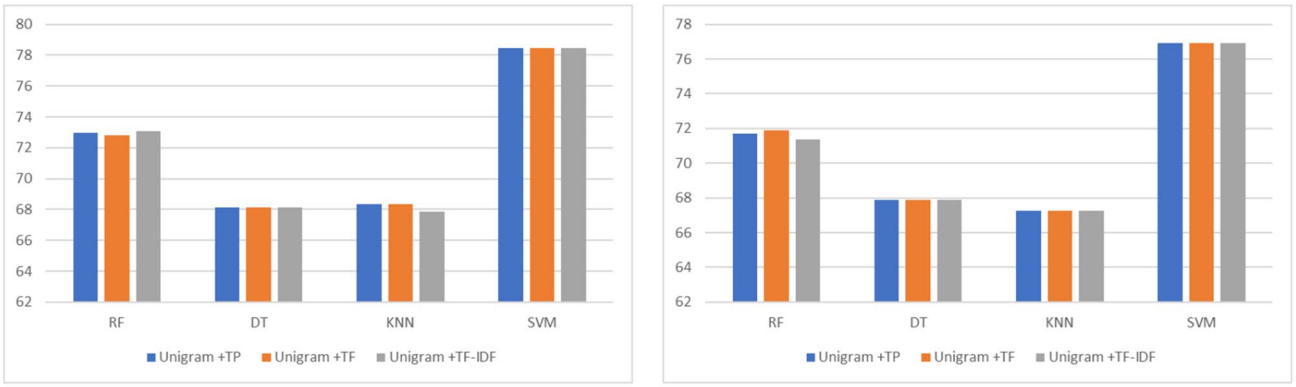
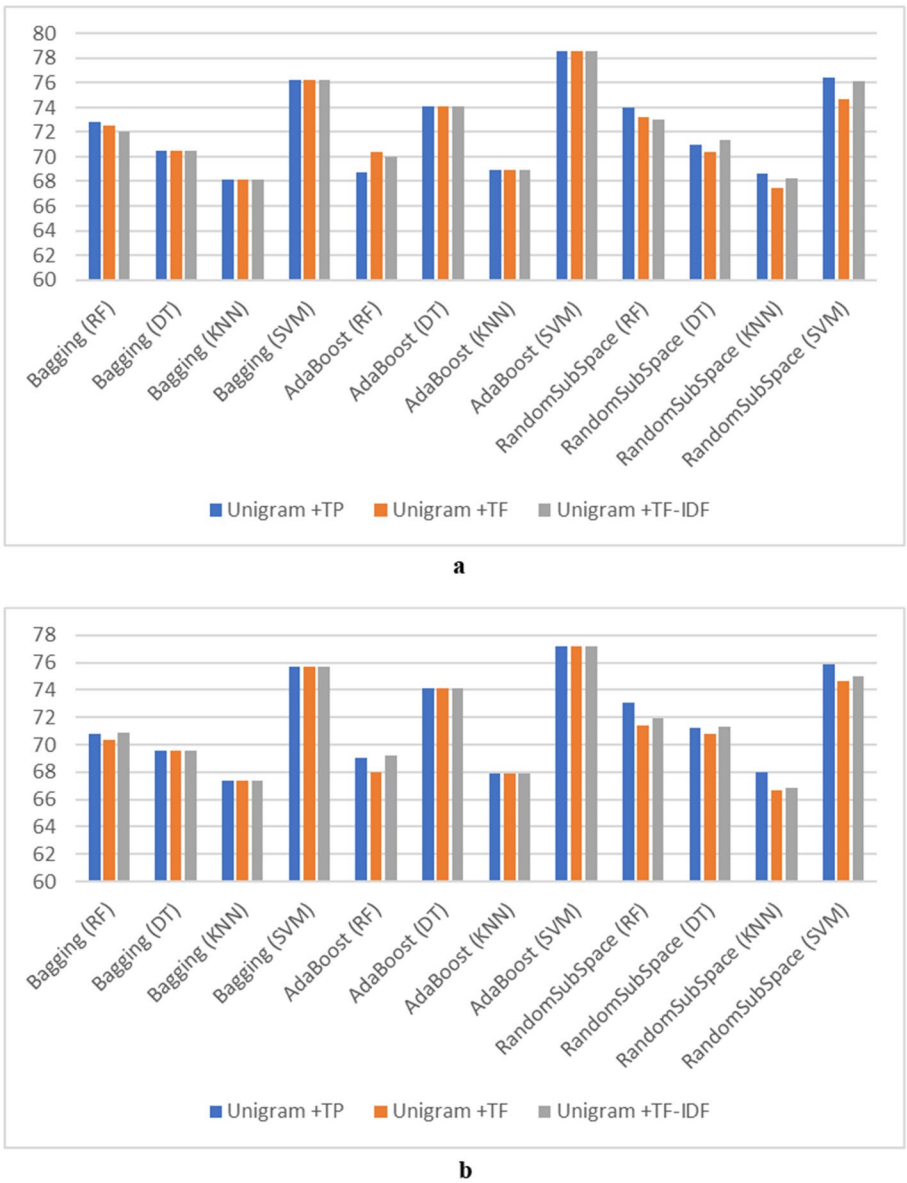


Fig. 3 **a**: tenfold cross-validation accuracy. **b**: fivefold cross-validation accuracy

Fig. 4 **a**: tenfold cross-validation accuracy. **b**: fivefold cross-validation accuracy



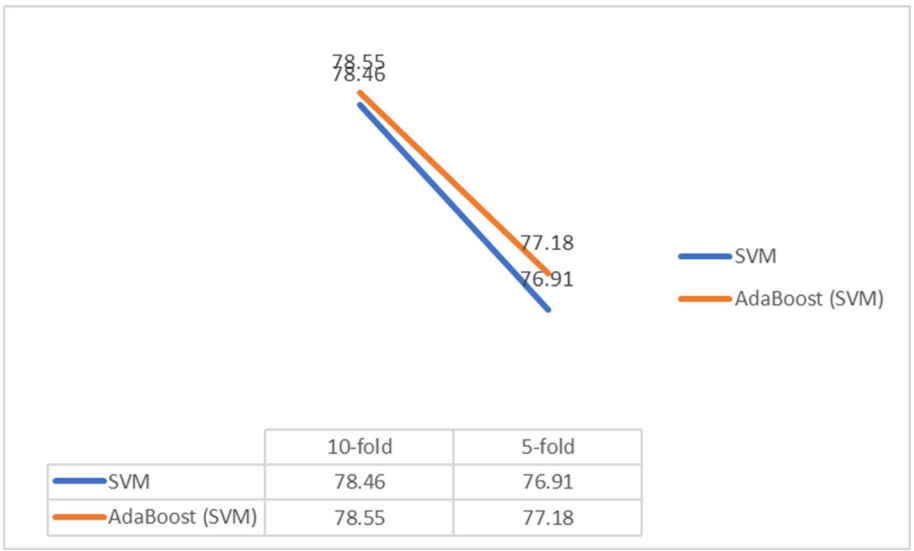


Fig. 5 Final comparative accuracy

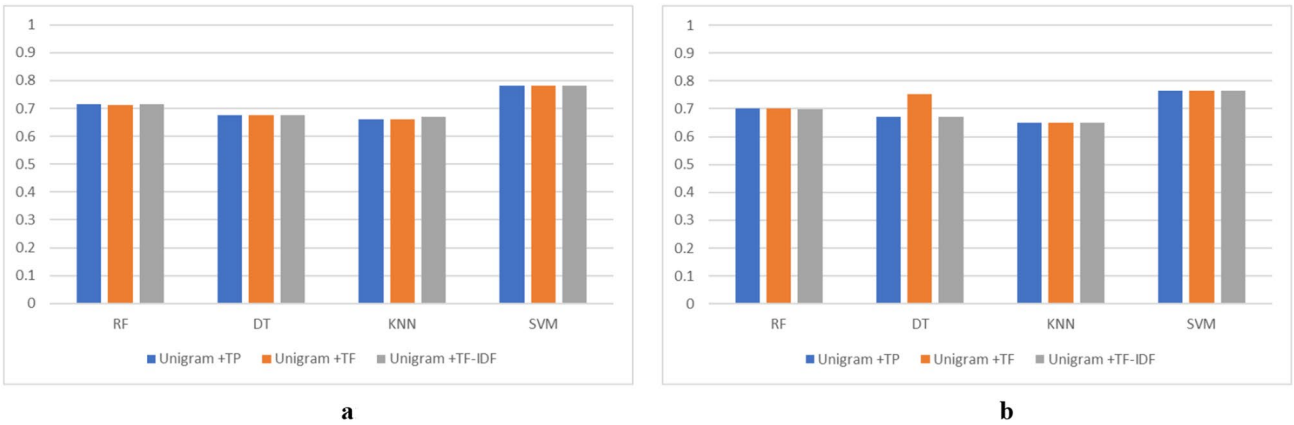


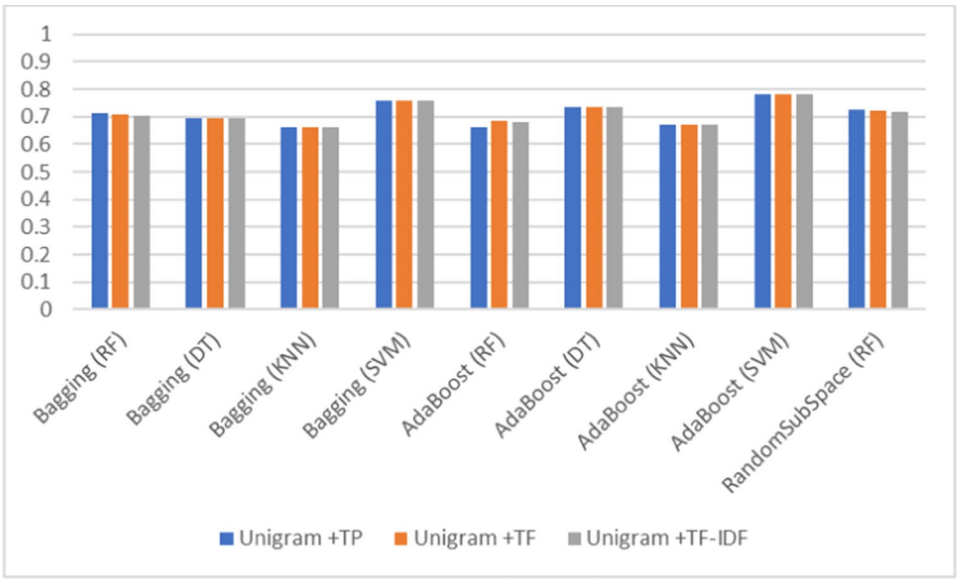
Fig. 6 **a**: tenfold cross-validation F1-score. **b**: fivefold cross-validation F1-score

6.4.1 Comparative F1-score for the conventional algorithms

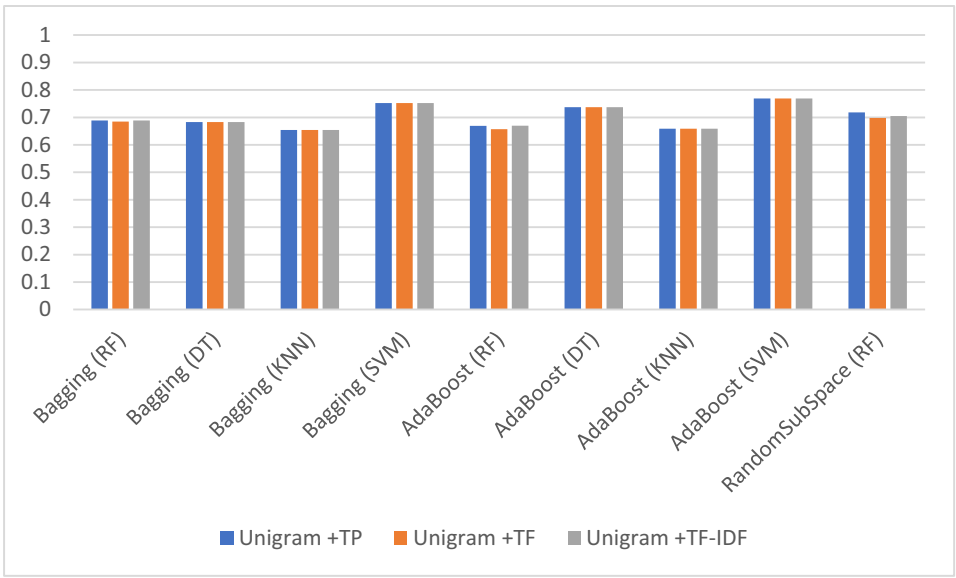
The F-measure score for the conventional machine learning algorithms, as depicted in Fig. 6a and b, shows that the SVM from the tenfold cross-validation has the highest value of 0.781. In contrast, the F-measure value for the fivefold is still SVM, with a score of 0.766.

6.4.2 Comparative F1-score for the ensemble algorithms

In Fig. 7a and b, the AdaBoost (SVM) using the tenfold cross-validation method has the highest F1-score of 0.782 compared with 0.769 for AdaBoost (SVM) using the fivefold.



a



b

Fig. 7 a: tenfold cross-validation F1-score. **b**: fivefold cross-validation F1-score

6.4.3 Summary of F-score

In response to **RQ2**, the F-measure score of AdaBoost (SVM) for the tenfold cross-validation method outperformed the traditional SVM with a 0.001 margin. This is depicted in Fig. 8.

6.5 Comparative analysis, ROC Value

The ROC curve shows how much the model can distinguish between the labelled classes. The ROC curve plots the TPR against the FPR with an AUC measure of the two-dimensional area.

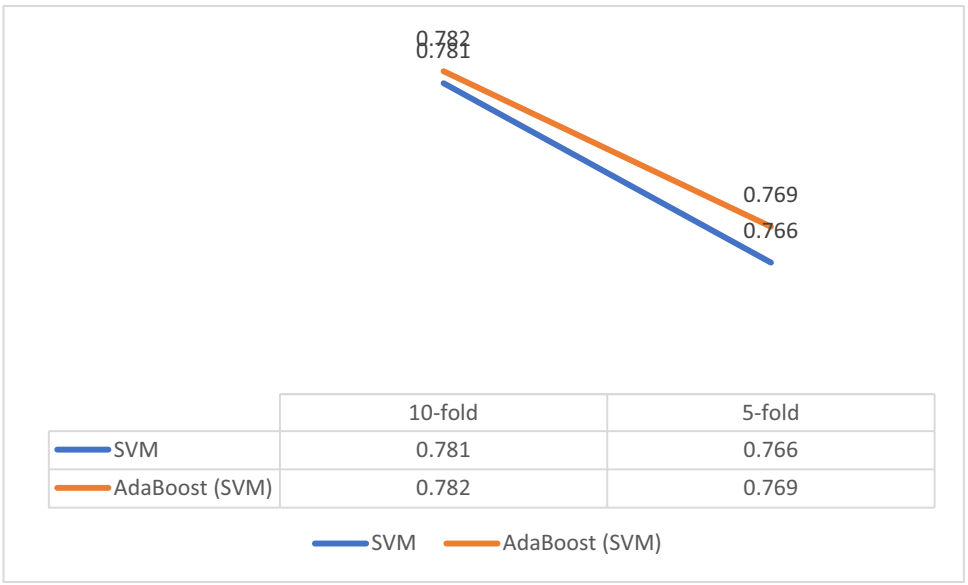


Fig. 8 Final comparative F1-score

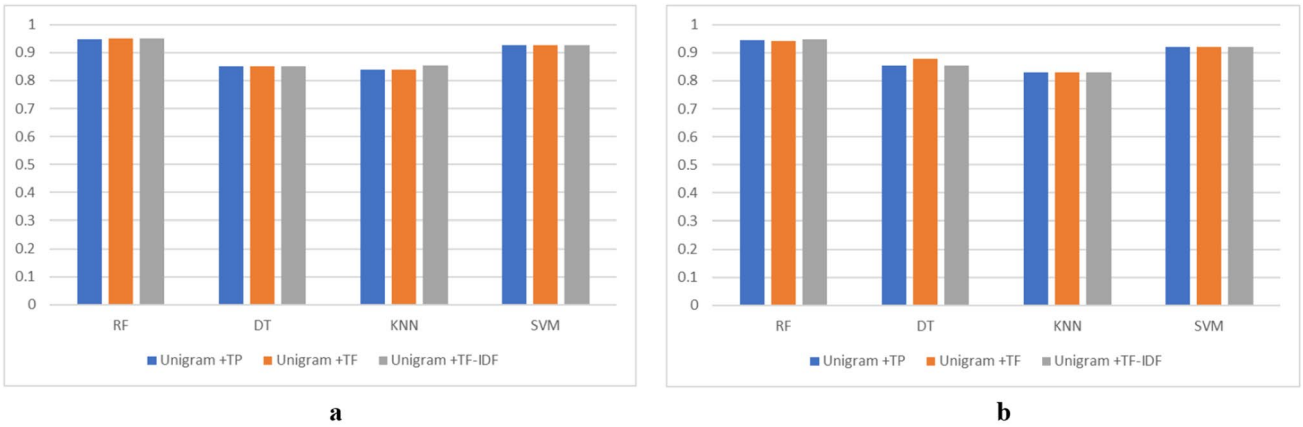


Fig. 9 **a**: tenfold cross-validation, ROC Value. **b**: fivefold cross-validation, ROC Value

6.5.1 Comparative ROC value for the conventional algorithms

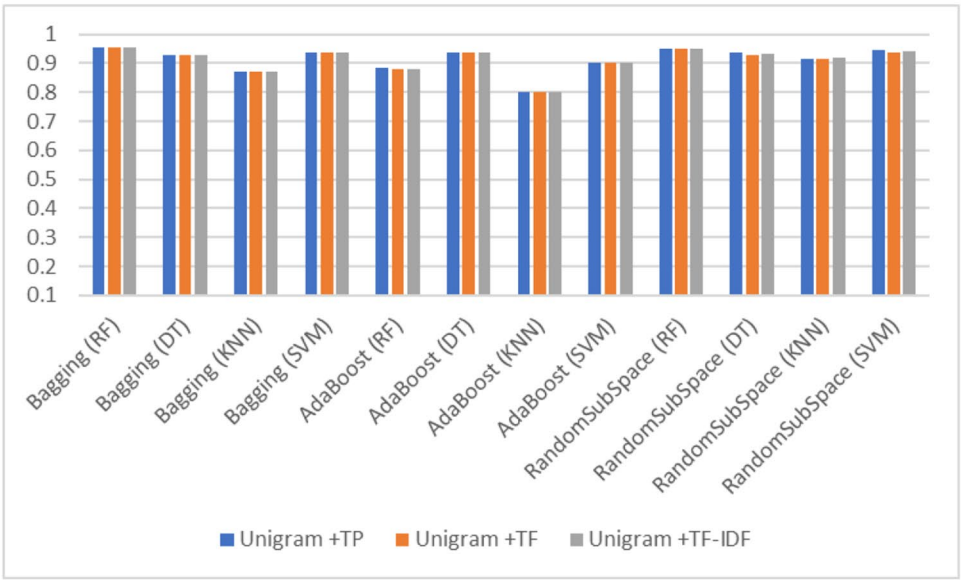
In Fig. 9a and b, the RF algorithm has the highest ROC value of 0.951 for Unigram (TF) using the tenfold cross-validation technique, while the SVM is second with a value of 0.928 across all the weighting schemes of Unigram.

6.5.2 Comparative ROC value for the ensemble algorithms

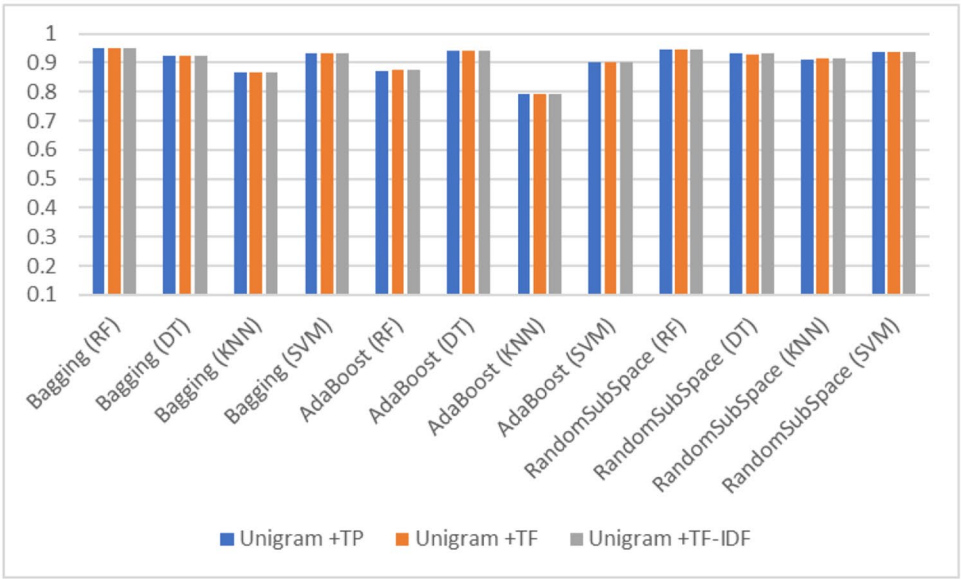
In Fig. 10a and b, the Bagging (RF) for Unigram (TP) has the highest ROC value of 0.955 using the tenfold cross-validation. The ensemble for SVM did not perform well. The highest value of 0.944 for ensemble SVM was obtained by RandomSubSpace using the Unigram (TP).

6.5.3 Summary of ROC value

In response to **RQ2,** as shown in Fig. 11, the ROC value of the Bagging (RF) ensemble algorithm using the tenfold cross-validation outperforms the conventional RF algorithm.



a



b

Fig. 10 **a**: tenfold cross-validation, ROC value. **b**: fivefold cross-validation, ROC value

7 Discussion and findings

The results demonstrate that the tenfold cross-validation outperforms the fivefold cross-validation method for all the classification metrics, accuracy, F1-score and ROC value. The Unigram, Bigram and Trigram feature selection methods implemented across the three weighting schemes (TP, TF, TF-IDF) still show better performance for the tenfold cross-validation techniques. Secondly, the ensemble machine learning algorithms outperform the conventional algorithms when implemented with Unigram and Bigram feature selection. The Trigram, however, shows a better performance for the conventional machine learning algorithms across the classification metrics when compared to the ensemble methods. The ensemble algorithm's primary drawback is the length of the training time. The ensemble classifiers took a significant time to train the dataset during simulation. The accuracy for both conventional and ensemble algorithms decreased as the N-gram increased from Unigram to Trigram. The weighting schemes also did not significantly affect accuracy, even though they adversely delayed the classifier's training. The F1-score across the tenfold and the fivefold

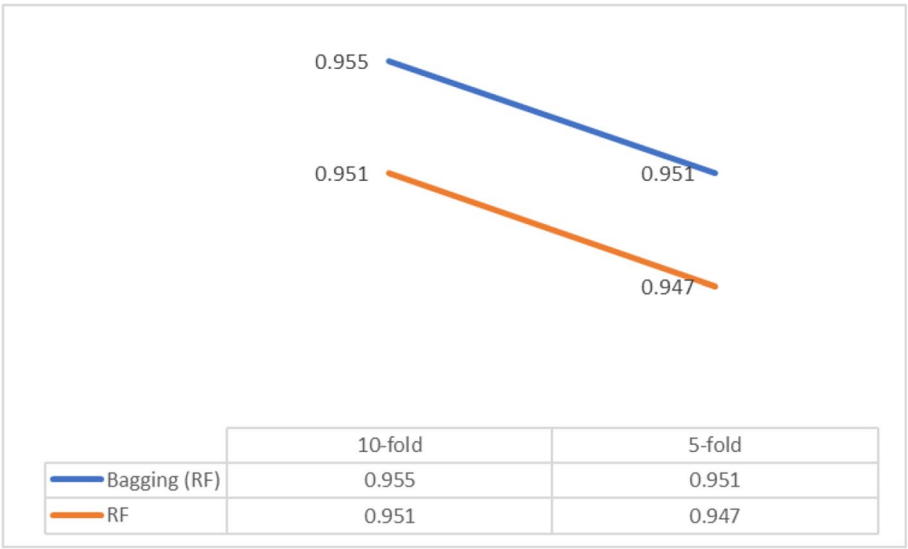


Fig. 11 Final comparative, ROC value

cross-validation techniques using the Trigram resulted in very low values. The weighting schemes did not adversely affect the ROC value even though it decreased from Unigram to Trigram.

The simulation results indicate that AdaBoost (SVM) ensemble algorithm has the highest accuracy of 78.55 using the tenfold cross-validation technique in Unigram (TP, TF, TF-IDF). In addition, the AdaBoost (SVM) has the highest F1-score of 0.757. The ROC value, however, shows a higher score of 0.955 for the Bagging (RF) ensemble algorithm using the tenfold cross-validation technique under Unigram (TP), whiles AdaBoost (SVM) has a ROC value of 0.903.

The study by Onan [25] compares ensemble and traditional machine learning algorithms for Unigram, Bigram and Trigram under the weighting schemes TP, TF, and TF-IDF. Even though the study was limited to the tenfold cross-validation method, the results indicate a superior performance for the ensemble algorithms over conventional algorithms. In addition, Onan [25] simulation shows a slightly decreased accuracy and F1-score from Unigram to Trigram across the traditional and the ensemble machine learning methods. The findings of Onan align with our study, where ensemble methods dominated conventional algorithms for accuracy, F1-score and ROC value. Even though Hassan et al. [12], whiles comparing conventional KNN, RF, and SVM, show a dominant performance for KNN, the study we conducted shows KNN as the worst performing algorithm.

8 Conclusion and future work

In this study, we demonstrated the dominance of the tenfold cross-validation technique over the fivefold methods and significantly showed the impact of the weighting schemes for question classification. Although the study utilised data from Ghana, results show a trend where ensemble algorithms outperformed conventional ones. The discussion aspect of the study is narrow since the data sources are different, and most of the reviewed literature focused on the tenfold cross-validation method without the weighting schemes and the N-gram features. Comparing ensemble methods and deep learning algorithms is an element of the research for future studies.

Author contributions DKD: conceptualization, methodology, analysis, EN: literature review, draft review, GSK: introduction, review and editing, WXA: results interpretation, methodology.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability Data available upon request.

Declarations

Ethics approval and consent to participate The paper reflects authors own original research and has not been published elsewhere not is it under consideration for publication elsewhere.

Competing interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Al-Sahaf H, Bi Y, Chen Q, Lensen A, Mei Y, Sun Y, Tran B, Xue B, Zhang M. A survey on evolutionary machine learning. J R Soc N Z. 2019;49:205–28.
2. Amineh RJ, Asl HD. Review of constructivism and social constructivism. J Soc Sci Lit Lang. 2015;1:9–16.
3. Bakhshinategh B, Zaiane OR, ElAtia S, Ipperciel D. Educational data mining applications and tasks: a survey of the last 10 years. Educ Inf Technol. 2018;23:537–53.
4. Bhardwaj R, Nambiar AR, Dutta D. A study of machine learning in healthcare. Proc Int Comput Softw Appl Conf. 2017;2:236–41.
5. Cantabella M, Martínez-España R, Ayuso B, Yáñez JA, Muñoz A. Analysis of student behavior in learning management systems through a Big Data framework. Futur Gener Comput Syst. 2019;90:262–72.
6. Drus Z, Khalid H. Sentiment analysis in social media and its application: systematic literature review. Procedia Comput Sci. 2019;161:707–14.
7. De Taunk KS, Verma S, Swetapadma A. A brief review of nearest neighbor algorithm for learning and classification. In: Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems, ICCS 2019; 2019. pp. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747.
8. Ehrentraut C, Ekholm M, Tanushi H, Tiedemann J, Dalianis H. Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. Health Inform J. 2018;24:24–42.
9. García Docampo L. [Reseña del libro] Primary and secondary education during Covid-19. Disruptions to educational opportunity during a pandemic. Rev Iberoam Educ. 2021. https://doi.org/10.35362/rie8724757.
10. Gaire B, Sharma S, Rijal B, Gautam D, Lamichhane N. Insincere question classification using deep learning. Int J Sci Eng Res. 2019;10:2001–4.
11. Geng S, Law KMY, Niu B. Investigating self-directed learning and technology readiness in blending learning environment. Int J Educ Technol High Educ. 2019. https://doi.org/10.1186/s41239-019-0147-0.
12. Hassan SU, Ahamed J, Ahmad K. Analytics of machine learning-based algorithms for text classification. Sustain Oper Comput. 2022;3:238–48.
13. Hubackova S. History and perspectives of elearning. Procedia Soc Behav Sci. 2015;191:1187–90.
14. Kara M. A systematic literature review: constructivism in multidisciplinary learning environments. Int J Acad Res Educ. 2018;4:19–26.
15. Kamath CN, Bukhari SS, Dengel A. Comparative study between traditional machine learning and deep learning approaches for text classification. Proceedings of the ACM Symposium on Document Engineering 2018, DocEng; 2018. https://doi.org/10.1145/3209280.3209526
16. Kotsiopoulos T, Sarigiannidis P, Ioannidis D, Tzovaras D. Machine learning and deep learning in smart manufacturing: the smart grid paradigm. Comput Sci Rev. 2021. https://doi.org/10.1016/j.cosrev.2020.100341.
17. Kotrikadze EV, Zharkova LI. Advantages and disadvantages of distance learning in Universities. Propósitos y Represent. 2021. https://doi.org/10.20511/pyr2021.v9nspe3.1184.
18. Lei T, Shi Z, Liu D, Yang L, Zhu F. A novel CNN-based method for question classification in intelligent question answering. ACM Int Conf Proc Ser. 2018. https://doi.org/10.1145/3302425.3302483.
19. Maatuk AM, Elberkawi EK, Aljawarneh S, Rashaideh H, Alharbi H. The COVID-19 pandemic and E-learning: challenges and opportunities from the perspective of students and instructors. J Comput High Educ. 2022;34:21–38.
20. Madabushi HT, Lee M. High accuracy rule-based question classification using question syntax and semantics. COLING 2016—26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, 2016. pp. 1220–1230.
21. Mall A, Haupt TC. A Review of COVID-19 ' s digitalisation of built environment education. 2022;1–8.
22. Meshram V, Patil K, Meshram V, Hanchate D, Ramkteke SD. Machine learning in agriculture domain: a state-of-art survey. Artif Intell Life Sci. 2021;1:100010.
23. Mohasseb A, Bader-El-Den M, Cocea M. Question categorisation and classification using grammar based approach. Inf Process Manage. 2018;54:1228–43.
24. Moharm K, Eltahan M. The role of big data in improving E-learning transition. IOP Conf Ser Mater Sci Eng. 2020. https://doi.org/10.1088/1757-899X/885/1/012003.
25. Onan A. Sentiment analysis on massive open online course evaluations: a text mining and deep learning approach. Comput Appl Eng Educ. 2021;29:572–89.
26. Osisanwo FY, Akinsola JET, Awodele O, Hinmikaiye JO, Olakanmi O, Akinjobi J. Supervised machine learning algorithms: classification and comparison. Int J Comput Trends Technol. 2017;48:128–38.

27. Ozili PK, Arun T. Spillover of COVID-19: impact on the global economy. SSRN Electron J. 2020. https://doi.org/10.2139/ssrn.3562570.
28. Patnaik S, Gachago D. Supporting departmental innovation in eLearning during COVID-19 through eLearning champions. 2020 IFEES World Engineering Education Forum—Global Engineering Deans Council, WEEF-GEDC 2020. https://doi.org/10.1109/WEEF-GEDC49885.2020.9293653
29. Priyam A, Gupta R, Rathee A, Srivastava S. Comparative analysis of decision tree classification algorithms. Int J Curr Eng Tecnol. 2013;334–337.
30. Ray S. A quick review of machine learning algorithms. proceedings of the International Conference on machine learning, big data, cloud and parallel computing: trends, Prespectives Prospect Com 2019. 2019. pp. 35–39. https://doi.org/10.1109/COMITCon.2019.8862451
31. Rai R, Tiwari MK, Ivanov D, Dolgui A. Machine learning in manufacturing and industry 4.0 applications. Int J Prod Res. 2021;59:4773–8.
32. Rashid S, Yadav SS. Impact of Covid-19 pandemic on higher education and research. Indian J Hum Dev. 2020;14:340–3.
33. Razzaghnoori M, Sajedi H, Jazani IK. Question classification in Persian using word vectors and frequencies. Cogn Syst Res. 2018;47:16–27.
34. Roman M, Plopeanu AP. The effectiveness of the emergency eLearning during COVID-19 pandemic. The case of higher education in economics in Romania. Int Rev Econ Educ. 2021;37:100218.
35. Romero C, Ventura S. Educational data mining and learning analytics: an updated survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2020;10:1–21.
36. Samsudin CM. No 主観的健康感を中心とした在宅高齢者における健康関連指標に関する共分散構造分析Title. Konstruksi Pemberitaan Stigma Anti-China Pada Kasus Covid-19 Di Kompas.Com. 2020; 68: 1–12.
37. Sagi O, Rokach L. Ensemble learning: a survey. Wiley Interdiscip Rev Data Min Knowl Discov. 2018;8:1–18.
38. Shah DS, Schwartz HA, Hovy D. Predictive biases in natural language processing models: a conceptual framework and overview. 2020; 5248–5264. https://doi.org/10.18653/v1/2020.acl-main.468.
39. Thomas Rincy N, Gupta R. Ensemble learning techniques and its efficiency in machine learning: a survey. 2nd International Conference on Data, Engineering and Applications, IDEA 2020. 2020. https://doi.org/10.1109/IDEA49133.2020.9170675
40. Taye BT, Mihret MS, Tiguh AE. Readiness and intention for adapting new normal COVID-19 prevention campaign for sustainable response among debre berhan university student's during campus re-entry: a cross-sectional study. Front Educ. 2021;6:1–13.
41. UNESCO. School enrollment, tertiary (% gross). 2020. https://data.worldbank.org/indicator/SE.TER.ENRR. Retrieved 16 Aug 2022.
42. Upadhya BA, Udupa S, Kamath SS. Deep neural network models for question classification in community question-answering forums. In: Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019; 2016. pp. 6–11. https://doi.org/10.1109/ICCCNT45670.2019.8944861
43. Vasilakes J, Zhou S, Zhang R. Natural language processing. Mach Learn Cardiovasc Med. 2020;32:123–48.
44. Wan S, Yang H. Comparison among methods of ensemble learning. Proceedings—2013 International Symposium on Biometrics and Security Technologies, ISBAST 2013, 2013. pp. 286–290. https://doi.org/10.1109/ISBAST.2013.50.
45. Weber F, Schütte R. A domain-oriented analysis of the impact of machine learning—the case of retailing. Big Data Cogn Comput. 2019;3:1–14.
46. Wuest T, Weimer D, Irgens C, Thoben KD. Machine learning in manufacturing: advantages, challenges, and applications. Prod Manuf Res. 2016;4:23–45.
47. Yadav BP, Ghate S, Harshavardhan A, Jhansi G, Kumar KS, Sudarshan E. Text categorisation performance examination using machine learning algorithms. IOP Conf Ser Mater Sci Eng. 2020. https://doi.org/10.1088/1757-899X/981/2/022044.
48. Zantalis F, Koulouras G, Karabetsos S, Kandris D. A review of machine learning and IoT in smart transportation. Future Internet. 2019;11:1–23.
49. Zhen L, Sun X. The research of convolutional neural network based on integrated classification in question classification. Sci Progr. 2021. https://doi.org/10.1155/2021/4176059.
50. Zhou Z, Zhu X, He Z, Qu Y. Question classification based on hybrid neural networks. 2016;50:44–52. https://doi.org/10.2991/iceeecs-16.2016.11.
51. Zulqarnain M, Ghazali R, Ghouse MG, Husaini NA, Alsaedi AKZ, Sharif W. A comparative analysis on question classification task based on deep learning approaches. PeerJ Comput Sci. 2021;7:1–27.