Received: 10 November 2019 | Accepted: 21 April 2020

DOI: 10.1002/cae.22253

SPECIAL ISSUE ARTICLE

WILEY

Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach

Aytuğ ONAN

Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Katip Çelebi University, İzmir, Turkey

Correspondence
Aytuğ ONAN, Department of Computer Engineering, Faculty of Engineering and Architecture, İzmir Katip Çelebi University, 35620 İzmir, Turkey.
Email: aytug.onan@ikc.edu.tr

Abstract

Massive open online courses (MOOCs) are recent innovative approaches in distance education, which provide learning content to participants without age-, gender-, race-, or geography-related barriers. The purpose of our research is to present an efficient sentiment classification scheme with high predictive performance in MOOC reviews, by pursuing the paradigms of ensemble learning and deep learning. In this contribution, we seek to answer several research questions on sentiment analysis on educational data. First, the predictive performance of conventional supervised learning methods, ensemble learning methods and deep learning methods has been evaluated. Besides, the efficiency of text representation schemes and word-embedding schemes has been evaluated for sentiment analysis on MOOC evaluations. For the evaluation task, we have analyzed a corpus containing 66,000 MOOC reviews, with the use of machine learning, ensemble learning, and deep learning methods. The empirical analysis indicate that deep learning-based architectures outperform ensemble learning methods and supervised learning methods for the task of sentiment analysis on educational data mining. For all the compared configurations, the highest predictive performance has been achieved by long short-term memory networks in conjunction with GloVe word-embedding scheme-based representation, with a classification accuracy of 95.80%.

KEYWORDS

deep learning, massive open online courses, sentiment analysis, text mining

1 | INTRODUCTION

Information and communication technologies have considerably affected many aspects, including education domain. Massive open online courses (MOOCs) are recent innovative approaches in distance education, which provide learning content to participants without age-, gender-, race-, or geography-related barriers. MOOC can be defined as an online course provided to an unlimited number of participants from geographically scattered locations [33]. The number of people enrolled in MOOCs has been steadily increasing, and there are several platforms, such as Coursera, edX, MiríadaX, or FutureLearn, to provide free open courses from prestigious institutions worldwide to thousands of participants [47]. MOOC platforms generally employ conventional learning content, such as short video lectures, slide presentations, reading texts, problem sets, live chat, and online learning assessments [31,43]. In addition, MOOCs are generally supported by discussion forums to enhance the interactions between the stakeholders of the learning process. MOOCs are typically characterized by free registration and an easy access to

educational information and resources. The use of MOOCs provides instructors the opportunity to reach many students throughout the world, and it also provides students the opportunity to select courses from a large set of courses offered by instructors from prestigious higher institutions, which may not be possible otherwise [6,34]. MOOCs support the conceptualization of ongoing professional learning and lifelong personalized learning [21]. In addition, MOOCs are means to provide a more accessible and democratized model of higher education [35]. MOOCs have been constrained by several disadvantages. The completion rates for MOOCs are very low as compared with the conventional face-to-face courses or closed online courses. For instance, Breslow et al [15] identified a completion rate of 15% for a physics MOOC offered in 2012 to 154,763 students. Similarly, Ho et al [30] examined the completion rates of the 17 HarvardX and MITx MOOCs offered between Fall 2012 and Summer 2013, identifying that only 5% of the students received the course certificate. In addition, instructor-student interaction and learning assessment are limited in MOOCs [31].

Educational data mining (EDM) is an emerging research field concerned with the application of tools and techniques from data mining, machine learning, and statistics to data obtained from educational settings, with the aim of better understanding students and the settings of learning process [65]. EDM can provide educational policy makers useful insights to improve the efficiency and quality of teaching and learning [56]. The classification, clustering, association rule mining, and statistics techniques have been employed on educational data. Students' learning performance assessment and monitoring, dropout and retention prediction, and modeling the learning behavior are several application tasks for EDM [7]. Text mining is the process of extracting useful information from unstructured text document. Compared with the other applications, such as classification, clustering, association rule mining, or regression, the number of earlier works dedicated to the text mining on education data is very limited [7]. In the domain of text mining, the main techniques may be broadly classified as exploratory analysis, concept extraction, summarization, text categorization, and sentiment analysis [17].

Sentiment analysis (also, known as, opinion mining) is the computational field concerned with contextually mining unstructured text documents so that subjective information (such as opinions, sentiments, attitudes, evaluations, or emotions) can be extracted [54]. With the use of sentiment analysis, structured and insightful knowledge can be obtained from unstructured text documents, which can be useful for decision support and individual decision makers [52]. Sentiment analysis can be employed for many different tasks, such as assessment of online hotel reviews [38], identification of opinions toward refugee crisis [74], and disaster management [64]. In education domain, sentiment analysis can be utilized to evaluate the progress of group discussion [22], to recognize and regulate e-learners' emotions [66] and to identify learning-related emotions of students on text feedbacks [9].

Sentiment analysis methods have been typically grouped into two classes, that is, lexicon-based methods and machine learning-based methods [44]. In the lexicon-based methods, sentiment orientation of a text document has been identified with the use of a dictionary with positive and negative sentiment values for each word. In this way, the semantic orientation of words and phrases has been computed to identify the sentiment of the text. In contrast, the machine learning-based methods model the sentiment analysis as a supervised learning task. In this scheme, the learning model has been obtained by training the supervised learning algorithm in conjunction with the labeled set of text documents. In machine learning-based sentiment analysis, conventional supervised learning algorithms, such as Naïve Bayes algorithm, support vector machines, and k-nearest neighbor algorithm, have been successfully employed. Ensemble learning and deep learning methods may also be employed for sentiment analysis tasks. Ensemble learning (also known as multiple classifier systems) is the process of training multiple learning algorithms and combining their predictions by regarding them as a committee of decision makers to enhance the predictive performance of the learning model. Ensemble learning has been successfully employed in a wide range of application fields, including energy demand prediction [60], medical data analysis [49], and text genre classification [61]. For instance, Rodger [60] analyzed the predictive performance of machine learning algorithms (such as artificial neural networks and k-nearest neighbor) for energy demand prediction. In this contribution, Rodger [60] presented an ensemble classification scheme based on fuzzy-rough nearest neighbor algorithm. Similarly, Rodger [49] presented an ensemble classification scheme based on k-nearest neighbor algorithm for medical data analysis.

Deep learning is an emerging field of machine learning that processes data on the basis of multiple layers/or stages of nonlinear information processing in a hierarchical way [19]. Ensemble learning and deep learning can yield promising results for sentiment analysis tasks [25,58].

In this study, we have collected 66,000 student reviews for MOOCs from coursetalk.com, a comprehensive platform for MOOC reviews. We present a text mining approach to analyze MOOC reviews, with the use of machine learning, ensemble learning, and deep learning methods. In the machine learning-based approach, we utilized three term weighting schemes (i.e., term

presence [TP], term frequency [TF], and TF-IDF). The representation schemes have been evaluated in conjunction with five supervised learners (i.e., Naïve Bayes, support vector machines, logistic regression, k-nearest neighbor, and random forest) and five ensemble learning methods (i.e., AdaBoost, Bagging, Random Subspace, voting, and Stacking). In the deep learning-based approach, we have utilized three word-embedding schemes (i.e., word2vec, fastText, and GloVe) in conjunction with five deep learning architectures (i.e., convolutional neural network, recurrent neural network, bidirectional recurrent neural network with attention mechanism, gated recurrent unit, and long short-term memory). To the best of our knowledge, this is the first comprehensive study on sentiment analysis of massive open online course reviews, in which the predictive performances of conventional classification algorithms, ensemble learning methods, and deep learning algorithms have been reported. In the empirical analysis, the following research questions have been addressed:

(1) Is there a statistically meaningful difference between the predictive performance of conventional supervised learning methods, ensemble learning algorithms, and deep learning methods?
(2) Do ensemble methods enhance the predictive performance of supervised learning methods?
(3) Which text representation scheme yields the highest predictive performance for sentiment analysis on MOOC reviews?
(4) Is there a significant difference in predictive performance between different word-embedding schemes on MOOC reviews?
(5) Is there a significant difference in predictive performance between different deep learning architectures on MOOC reviews?

The rest of this paper is structured as follows. In Section 2, related works on sentiment analysis have been presented. Section 3 presents the materials and methods of study, namely text corpus, text representation schemes, machine learning classifiers, ensemble learning methods, word-embedding schemes, and deep learning architectures. Section 4 reports the experimental procedure and the empirical results. Finally, concluding remarks have been presented in Section 5.

2 | RELATED WORKS

Sentiment analysis on educational data can be employed to obtain feedback on learning content and resources, which can provide useful insights to enhance the quality of learning content and to identify learning behavior of students. This section briefly presents the earlier works in the field. Adamopoulos [2] analyzed user-generated online reviews about MOOCs to identify the effects of factors, such as course, platform, and university on student retention. In another study, Valakunde and Patwardhan [67] employed sentiment analysis on review comments provided by students on evaluation reviews. In this scheme, frequency-inverse document frequency (TF-IDF) weighting scheme has been utilized to represent review comments in conjunction with two machine learning algorithms, that is, Naïve Bayes and support vector machines. Similarly, Wen et al [71] employed sentiment analysis on MOOC discussion forums. In this scheme, forum posts from three MOOCs have been utilized to identify dropout characteristics of students. The analysis revealed that there is a significant correlation between sentiment expressed in the course forum posts and the completion rates for MOOCs. In another study, Altrabsheh et al [9] introduced a machine learning-based sentiment analysis scheme to extract learning-related emotions of students on text feedbacks. In this scheme, student feedbacks, opinions, and feelings about different courses, such as calculus, communication skills, database, engineering, have been collected/gathered using Twitter. To represent text documents, three conventional N-gram models (namely unigram, bigram, and trigram) and their combinations have been examined. In the classification phase, Naïve Bayes algorithm, support vector machines, maximum entropy classifier, and random forest algorithm have been utilized. In another study, Adinolfi et al [3] presented a sentiment analysis framework to examine student satisfaction on different platforms, such as massive open online courses, learning diaries, and Twitter. In the presented scheme, the students' and teachers' behaviors have been also modeled. Similarly, Bogdan [12] employed sentiment analysis to improve the course content and to identify students' opinions regarding the integration of MOOCs into embedded system blended courses. In another study, machine learning-based sentiment analysis has been employed to mine opinions from students' comments about the performance of instructors [26]. In the presented scheme, support vector machines and random forest algorithm have been applied for sentiment analysis. Moreno-Marcos et al [46] presented empirical results for lexicon-based and machine-learning based approaches to sentiment classification on forum messages in MOOCs to extract patterns of learners' behavior. In this scheme, logistic regression, support vector machines, decision trees, random forest, and Naïve Bayes algorithm have been utilized as the supervised learning methods. The empirical analysis indicated that random forest

algorithm outperforms the other supervised learning methods and lexicon-based framework for sentiment analysis on MOOC reviews. Text mining and sentiment analysis have been also employed to extract information about the drivers of higher educational institutions' online success [63]. In the presented scheme, topic modeling and topic profiling analysis methods have been employed to enhance the international attractiveness of higher educational institutions. In another study, Abdi et al [1] presented a query-based, multi-document, opinion-oriented summarization approach. The presented scheme utilizes sentiment analysis to extract sentiment orientation and subjective information. In the summarization module, user's query relevant sentences have been identified. Belbachir and Boughanem [10] utilized language models used in information retrieval to represent the query and document for sentiment analysis. Moreover, Al-Smadi et al [8] employed morphological, syntactic, and semantic features for sentiment analysis.

Recently, Bustillos et al [53] presented a comprehensive analysis of machine learning and deep learning methods for opinion mining in an intelligent learning environment. In this work, several machine learning algorithms (such as Bernoulli Naïve Bayes, multinomial Naïve Bayes, support vector machines, linear support vector machine, stochastic gradient descent, and k-nearest neighbor algorithm) and several deep learning architectures (such as convolutional neural network and long short-term memory) have been employed. The highest predictive performance, with a classification accuracy of 88.26%, has been obtained with the use of a deep learning-based architecture. Similarly, Cabada et al [16] employed two deep earning-based architectures (i.e., convolutional neural network and long short-term memory) on educational reviews, obtaining a classification accuracy of 84.32%. Nguyen and Nguyen [48] introduced a convolutional N-gram bidirectional LSTM word-embedding architecture for sentiment analysis on video comments. In the presented scheme, a word with semantic and contextual information in short- and long-distance periods has been represented. Lin et al [40] examined the predictive performance of knowledge-based and machine learning-based approaches for sentiment analysis on student evaluations of teaching. In another study, López et al [42] presented a framework on the basis of opinion mining and semantic profiling on educational resource platform. Recently, Onan [50] examined the predictive performance of conventional classification algorithms, ensemble methods, and deep learning algorithms on student evaluations of teaching.

In another study, Wang et al [70] presented a hybrid deep learning-based scheme for sentiment analysis on the basis of convolutional neural networks and long short-term memory. Similarly, Wang et al [69] introduced a stacked residual long short-term memory-based architecture to identify sentiment intensity of text documents.

3 | METHODOLOGY

In this section, text corpus, machine learning-based approach for sentiment analysis, and deep-learning based approach for sentiment analysis have been presented.

3.1 | Corpus

To collect a text corpus on MOOC reviews, we have crawled coursetalk.com, a comprehensive platform for MOOC reviews. The course reviews from diverse range of fields, such as accounting, algebra, aerospace engineering, agriculture, computer science, data science, and education, have been considered. In this way, we have collected approximately 93,000 MOOC reviews. In this platform, feedbacks of courses can be provided with the use of a 5-point scale score, from which an overall quality score has been computed for a course. To obtain a labeled corpus, we have utilized the quality scores provided by the learners. In this way, evaluation reviews with the quality scores of 1 and 2 have been labeled as "negative," whereas evaluation reviews with the quality scores of 4 and 5 have been labeled as "positive." After the labeling process, we have obtained a corpus with approximately 33,000 negative reviews and approximately 37,000 positive reviews. To obtain a balanced corpus, our final corpus consists of 66,000 reviews, with 33,000 negative and 33,000 positive sentiments. In Table 1, sample MOOC reviews from the corpus have been presented.

To process text documents by machine learning algorithms, several preprocessing tasks should be conducted. For the preprocessing task, we have adopted the preprocessing stages outlined in [1,48]. First, we employed text normalization. In this stage, all letters in the text corpus have been converted into lowercase letters. All sequence and punctuation marks have been eliminated. Abbreviations have been converted into their expanded versions. In addition, URLs, stop words, irrelevant words, and sparse terms have been eliminated. Tokenization (which is the process of separation of given sentences and words of the documents into tokens or characters) has been performed. In addition, stemming has been performed on text corpus to reduce words to their word stems. For stemming task, we have utilized Snowball stemming algorithm [57]. Text preprocessing tasks have been implemented on natural language toolkit [41].

TABLE 1 Sample massive online open course reviews of students

Sentiment orientation	MOOC evaluation review
Negative	I was impressed in the beginning, but afterward I figured out that it is a total waste of time and money, as everything is made to demonstrate how muscular and strong the professor is, and how they can invent puzzling exercise and the hardest problems ever. This does not convey knowledge; also, the irrational amount of content in the given time is for a full-time student, not for a self-based student, as they claimed
Negative	Course content/quality is very average, and it is even loosely connected. What is the use of learning association rules mining without learning even all the different rules? There is no application of those things to real-world examples; the course only deals with the coding part—just complete the coding. How do we even justify 500 bucks for this? Plus, the lack of communication between staff and students is really demoralizing. Bugs in notebooks are reported in forum for days and there are still no replies or solution from staff. Audit learners cannot even get access to exams despite being informed that there is an exam at the beginning of the course. Seriously, it should be assured that the quality is of MicroMaster level before charging 500 for it. The Supply Chain MIT courses charge only 150–200, with much better content/organizations.
Negative	The lessons do not teach much and do not help at all with the "homework" assignments. I would not recommend this course.
Positive	This course covers all the important aspects of software engineering from requirement elicitation and architectural patterns to build and release management. The short videos, quizzes, and programming exercises helped to understand the topics, and the instructors answered open questions using Slack. I really enjoyed this course and want to thank all the instructors for the effort they put into this course!
Positive	Great course. The videos are very informative and subtitled. The single-graded in-course quizzes provide a frequent check-up for the learned content. The final transcribing projects are a team effort. I will certainly enroll in the next course of the Deciphering Secrets series as well.
Positive	It is the best online course available for developing basic CFD and FEA concepts, and learning Ansys side by side. I really look forward to an advanced course or may be a specialization series in CFD and/or FEA with Ansys.

Abbreviation: CFD, computational fluid dynamics; FEA, finite element analysis; MOOC, massive open online courses.

3.2 | Machine learning-based sentiment analysis

In the machine learning-based sentiment analysis, there are two main stages, namely extraction of features from the data and their representation in terms of feature vectors, and training of the supervised learning algorithms on the feature vectors to obtain the learning model. Based on the obtained learning model, the class labels for unseen instances have been determined [36]. For machine learning-based sentiment analysis, we have employed three term weighting schemes (i.e., TP, TF, and TF-IDF) and three N-gram models (namely bigram, unigram, and trigram model). These text representation schemes have been utilized to train five supervised learning algorithms (i.e., Naïve Bayes, support vector machines, logistic regression, k-nearest neighbor, and random forest) and five ensemble methods (i.e., AdaBoost algorithm, Bagging, Random Subspace, voting, and Stacking). The rest of this section presents text representation schemes, supervised learning algorithms, and ensemble learning methods utilized in the machine learning-based sentiment analysis.

3.2.1 | Feature construction

To process text documents in conjunction with supervised learning algorithms, the conversion of documents into a feature vector is a crucial task. In text mining and information retrieval task, one common scheme that has been frequently and successfully employed is bag-of-words (BOW) framework. In this framework, a text document is regarded as a bag of words and represented by a vector containing all the words encountered in the document, without taking into account syntax, word orderings, and grammar [27]. In this framework, each text document has been represented on the basis of the frequency of each word. The set of features has been utilized to train the supervised learning algorithm to obtain the learning model. Based on the bag-of-words framework, there are three types of weighting schemes that may be employed, namely TP, TF, and TF-IDF scheme. For the TP-based weighting, it has been considered whether a word occurs in a text document or not. In this scheme, a binary-valued feature vector has represented each text document, such that one has been used to represent that a word has occurred and zero has been used to indicate that

the word has not occurred in the document. For the TF-based weighting, the number of occurrences of each word encountered in a document has been computed. In this way, frequently encountered words have been assigned higher scoring values, whereas rarely encountered words have been assigned lower scoring values. This issue may be problematic, as frequent words will have dominance over the rarely encountered ones. For some tasks in natural language processing, some rarely encountered words may be domain-specific words and more informative about the context. To eliminate the problems associated with TF-based weighting, an inverse document frequency may be utilized to measure the frequency of rare words across the text documents. This scheme has been known as TF-IDF weighting. In this scheme, the frequencies of words have been rescaled on the basis of the number of occurrences in all documents. The frequently encountered words have been penalized. In text mining tasks, the N-gram model is an important representation scheme. In this model, n-character slice of a text document has been extracted. The common N-gram models utilized in sentiment analysis are unigram model (N = 1), bigram (N = 2), and trigram (N = 3). For the experimental analysis, we have modeled the text corpus using three weighting schemes (TF, TP, and TF-IDF) and three N-gram models (bigram, unigram and trigram). In this way, nine different configurations have been obtained for the corpus.

3.2.2 | Supervised learning methods

To obtain learning models based on the feature sets outlined in Section 3.2.1., we have considered five supervised learning methods. The algorithms have been briefly described:

- Naïve Bayes algorithm is a statistical supervised learning algorithm based on Bayes' theorem and conditional independence assumption [39].
- Support vector machines (SVM) are linear supervised learning algorithms that may be utilized for classification and regression tasks. SVM finds a hyperplane in the higher dimensional space to separate instances of different classes [68].
- Logistic regression is a linear classification algorithm, which provides a scheme to apply linear regression to classification problems. In this scheme, a linear regression model and transformed target variables have been employed to obtain a linear classification scheme [28].
- K-nearest neighbor (KNN) is an instance-based supervised learning algorithm that may be utilized for classification and regression tasks. In this scheme, an instance has been assigned to a class label, based on the majority voting of its neighbors. KNN involves storing all the instances at the time of classification [4].
- Random forest algorithm is an ensemble of bagging algorithm and random subspace algorithm. In this algorithm, decision trees have been employed as the base learner. Each tree has been built on the basis of bootstrap samples of the training data. The diversity among the base learners has been provided by a random feature selection. In this way, the model yields satisfactory results in the existence of noisy or irrelevant data [14].

3.2.3 | Ensemble learning methods

Ensemble learning (also known as multiple classifier systems) is the process of training multiple learning algorithms and combining their predictions by regarding them as a committee of decision makers. Ensemble learning aims to identify a learning model with a higher predictive performance [51]. The algorithms have been briefly described:

- AdaBoost is a boosting-based ensemble learning algorithm, in which a more robust classification model has been obtained by focusing on the instances that are harder to learn [24].
- Bagging (also known as Bootstrap aggregating) is another ensemble algorithm, which combines base learning algorithms trained on different training subsets obtained from the original training set by the bootstrap sampling [13].
- Random Subspace algorithm is another ensemble learning method, which achieves diversity among the members of the ensemble with feature space-based partition [29].
- Voting is a simple way to combine the predictions of individual learning algorithms of the ensemble. In general, voting schemes can be divided into two categories, as unweighted voting schemes and weighted voting schemes. The unweighted voting schemes include minimum probability, maximum probability, majority voting, product of probability, and average of probabilities [52]. In the empirical analysis, we have considered the five unweighted voting schemes to combine the supervised learning algorithms.
- Stacking (also known as stacked generalization) is another ensemble combination scheme. Stacking algorithm employs a two-staged structure to combine the predictions of multiple learning algorithms [72].

3.3 | Deep learning-based sentiment analysis

For the deep learning-based sentiment analysis, the text corpus has been represented by three word-embedding schemes (namely word2vec, fastText, and GloVe). To process text, we have utilized five deep learning architectures (i.e., convolutional neural network, recurrent neural network, bidirectional recurrent neural network with attention mechanism, gated recurrent unit, and long short-term memory).

3.3.1 | Word embedding-based representation

For text mining and sentiment analysis tasks, one common representation scheme is bag-of-words scheme. Yet, bag-of-words scheme cannot capture the semantic relations among the components of text documents. In addition, this scheme yields a sparse data representation with a high-dimensional feature space [5]. For text classification, word embedding-based representation is an effective scheme, which can be utilized in conjunction with machine learning algorithms and deep learning architectures. The use of word embedding enables to represent text documents in a compact and more expressive way. Word embedding-based representation provides learning by distributed expressions of words existing in a low-dimensional space [11]. Word embeddings have been based on the distributional hypothesis. Based on this hypothesis, words with similar meanings should be encountered in a similar context. Hence, the vector-based representation aims to capture characteristics of the neighbors of a word. In this way, the similarity between words can be captured. In this scheme, a large unsupervised set of documents has been utilized to extract semantic and syntactic meaning among the words [59]. For this study, we have considered three word-embedding schemes (namely word2vec, fastText, and GloVe) in conjunction with the deep learning architectures. The representation schemes have been briefly described:

- The word2vec is an unsupervised and computationally efficient prediction model to learn word embeddings from text documents. The word2vec model consists of two models, that is, continuous bag of words model (CBOW) and continuous skip-gram model [45]. The CBOW model predicts the target word from its context words that surround it across a window size of k. In contrast, the skip-gram model predicts the context words, given the target word.
- The fastText is a computationally efficient representation scheme to learn word embeddings from text documents. In this scheme, each word has been regarded as a bag of character n-grams [32]. As compared with word2vec, the fastText scheme can yield a higher predictive performance for morphologically rich languages and rare words [20].
- The global vectors (GloVe) is an unsupervised prediction model to obtain vector representations for words. In this scheme, the local context-based learning of word2vec model has been integrated by the global matrix factorization. Based on the global word-word co-occurrence statistics obtained from the text corpus, training has been conducted. Based on the training process, linear structures of the word vector space have been extracted [55].

3.3.2 | Deep learning architectures

Deep learning algorithms and architectures have been already employed in a wide range of applications, including computer vision, pattern recognition, and natural language processing. For conventional machine learning schemes presented in Section 3.2, a high-dimensional and sparse feature representation has been utilized in conjunction with classifiers. Deep learning architectures provide the learning of multi-level feature representations. The architectures aim to identify learning models on the basis of multiple layers/or stages of nonlinear information processing in a hierarchical way [37]. The rest of this section briefly descries the deep learning architectures utilized in the empirical analysis.

- Convolutional neural networks (CNN) are deep neural network-based architectures, which process data with the use of a grid-based topology. CNN has been characterized by a special kind of mathematical operation, referred to as convolution. The convolution operation has been handled in one or more convolutional layers. Typical convolutional neural network architecture consists of input layer, output layer, and hidden layers. The hidden layers of the architecture comprise several layers, that is, convolutional layers, pooling layers, fully connected layers, and normalization layers. In convolutional layers, convolution operation has been employed on the input data and the feature maps have been obtained. To add the nonlinearity to the architecture, the activation functions (such as rectified linear unit) have been utilized in conjunction with the feature maps. The pooling layers have been utilized to combine the outputs of neuron clusters. In this way,

the spatial size of feature spaces has been reduced and the models' ability to deal with overfitting has been enhanced. In pooling layer, maximum pooling has been employed. The fully connected layers have been utilized to obtain the final output of the architecture [23].

- Recurrent neural network (RNN) is another type of deep learning architecture to process sequential data. In RNN, connections between neurons constitute a directed graph. In this architecture, internal state has been utilized to process sequence of inputs. Hence, the architecture can be successfully employed for sequential tasks, such as speech recognition. In RNN, each output has been determined by recurrently processing the same task over the instances of the sequence. Based on all the earlier computations, the output has been determined [73].
- Long short-term memory network (LSTM) is another deep learning architecture based on recurrent neural networks. Conventional RNN architecture suffers from the exploding or vanishing gradient problem. In RNN, the arbitrarily long sequences of input cannot be properly handled. In response, LSTM utilizes forget gates to overcome the problems. In LSTM architecture, the back propagation of error has been allowed until a limited number of time steps. For a typical LSTM unit, there is a cell and three kinds of gates, namely an input gate, an output gate, and a forget gate. The open and close operations at the gates have been utilized to control which information should be preserved and when the information should be accessed [62].
- Gated recurrent unit (GRU) is another deep learning architecture based on recurrent neural networks. In a typical GRU architecture, there are two gates (namely reset gate and update gate) [18].
- Recurrent neural network with attention mechanism (RNN-AM) is another deep learning architecture based on recurrent neural networks. The conventional encoder-decoder frameworks must encode all information, which may not be relevant to the current task [50,62]. For long input sequences, it is not possible to fully capture information-rich and selective encoding. To deal properly with this problem, attention mechanisms have been employed. In the bidirectional recurrent neural network architecture with attention mechanism, each output word y_t corresponds to a weighted combination of input states. In this scheme, the weight values define the weight contribution of each input state to the output state. Based on this scheme, the decoder pays varying attentions to the states [62].

4 | EXPERIMENTS AND RESULTS

In this section, performance measures, experimental procedure and experimental results obtained from the conventional classifiers, ensemble learning algorithms, and deep learning architectures have been presented.

4.1 | Performance measures

To evaluate the performance of machine learning models, classification accuracy (ACC) and F-measure have been considered in this study.

Classification accuracy is one of the most widely employed measures for evaluation of supervised learning algorithms, which is computed as given by Equation (1):

$$\mathrm{ACC} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TP}+\mathrm{FP} + \mathrm{FN}+\mathrm{TN}} \tag{1}$$

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

F-measure is another common measure for performance evaluation on supervised learning algorithms, which is the harmonic mean of the precision and recall. Precision (PRE) is the proportion of the true positives against the true positives and false positives, as given by Equation (2). Recall (REC) is the proportion of the true positives against the true positives and false negatives, as given by Equation (3). Based on Equations (2) and (3), F-measure has been computed, as given by Equation (4):

$$\mathrm{PRE} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$

$$\mathrm{REC} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$F\text{−measure} = \frac{2 \times \mathrm{PRE} \times \mathrm{REC}}{\mathrm{PRE} + \mathrm{REC}} \tag{4}$$

4.2 | Experimental procedure

To evaluate the predictive performance of models for sentiment analysis on EDM, we have conducted two set of experiments, namely, machine learning-based sentiment analysis and deep learning-based sentiment analysis.

For evaluation task on machine learning methods, the text corpus has been represented by three traditional text weighting schemes (i.e., TP, TF, and TF-IDF schemes) and three N-gram models (i.e., unigram, bigram, and trigram model). In this way, we have obtained nine configurations on the text corpus. The feature

TABLE 2 Classification accuracy values obtained by machine learning algorithms

	Unigram +TP	Unigram +TF	Unigram +TF-IDF	Bigram +TP	Bigram +TF	Bigram +TF-IDF	Trigram +TF	Trigram +TP	Trigram +TF-IDF
KNN	75.89	76.30	76.11	74.87	75.42	75.08	72.49	74.38	73.70
SVM	78.13	78.24	78.15	77.95	78.06	77.99	77.74	77.88	77.82
LR	77.49	77.69	77.57	77.10	77.28	77.19	76.48	76.99	76.83
NB	79.14	79.23	79.17	79.04	79.10	79.08	78.91	79.01	78.94
RF	78.67	78.75	78.71	78.48	78.59	78.54	78.31	78.43	78.40
AdaBoost (KNN)	82.86	82.92	82.91	82.77	82.84	82.79	82.69	82.76	82.73
AdaBoost (SVM)	83.37	83.48	83.41	83.31	83.34	83.33	83.18	83.24	83.22
AdaBoost (LR)	83.13	83.17	83.15	83.06	83.10	83.08	82.95	83.02	82.98
AdaBoost (NB)	84.18	84.33	84.23	84.09	84.16	84.13	83.85	84.06	83.95
AdaBoost (RF)	83.72	83.81	83.74	83.60	83.64	83.62	83.51	83.57	83.53
Bagging (KNN)	79.61	79.74	79.69	79.49	79.58	79.55	79.29	79.41	79.37
Bagging (SVM)	80.53	80.59	80.55	80.39	80.45	80.40	80.30	80.37	80.34
Bagging (LR)	80.21	80.27	80.22	80.10	80.20	80.14	79.90	79.99	79.96
Bagging (NB)	81.18	81.22	81.20	81.07	81.16	81.14	81.00	81.05	81.03
Bagging (RF)	80.85	80.97	80.90	80.77	80.81	80.79	80.67	80.74	80.72
RS (KNN)	84.56	84.70	84.60	84.45	84.54	84.50	84.37	84.41	84.40
RS (SVM)	86.02	86.17	86.09	85.67	85.94	85.74	85.47	85.60	85.51
RS (LR)	85.33	85.41	85.40	85.21	85.27	85.24	85.02	85.17	85.08
RS (NB)	**89.41**	**89.62**	**89.51**	**88.24**	**88.63**	**88.47**	**87.68**	**87.96**	**87.81**
RS (RF)	***87.10***	***87.53***	***87.33***	***86.61***	***86.90***	***86.68***	***86.28***	***86.52***	***86.40***
Voting (Minimum probability)	81.40	81.47	81.43	81.35	81.38	81.37	81.25	81.28	81.27
Voting (Maximum probability)	81.69	81.75	81.72	81.59	81.67	81.66	81.50	81.57	81.55
Voting (Majority voting)	81.97	82.01	81.99	81.84	81.89	81.86	81.77	81.82	81.79
Voting (Product of probability)	82.21	82.33	82.25	82.15	82.17	82.17	82.04	82.09	82.05
Voting (Average of probabilities)	82.59	82.65	82.61	82.50	82.58	82.53	82.40	82.45	82.43
Stacking	84.95	85.00	84.97	84.82	84.88	84.85	84.72	84.78	84.74

Abbreviations: KNN, k-nearest neighbor algorithm, LR, logistic regression; NB, Naïve Bayes; RF, random forest; RS, Random subspace method; SVM, support vector machines ; TF, term frequency; TF-IDF, frequency-inverse document frequency; TP, term presence.

Note: The highest values have been indicated by bold and second highest values have been indicated by italics.

representations have been evaluated in conjunction with five supervised learning methods (namely Naïve Bayes, support vector machines, logistic regression, K-nearest neighbor, and random forest algorithm) and five ensemble learning methods (i.e., AdaBoost, Bagging, Random Subspace, voting, and Stacking). For voting ensemble, the five supervised learning algorithms have been combined by using either minimum probability, maximum probability, majority voting, product of probability, or average of probabilities' combination rule. For Stacking ensemble, the five supervised learning algorithms have been utilized as the base-level classifiers, and logistic regression has been utilized as the meta-level classifier. Ten-fold cross validation has been employed on

TABLE 3 F-measure values obtained by machine learning algorithms

	Unigram +TP	Unigram +TF	Unigram +TF-IDF	Bigram +TP	Bigram +TF	Bigram +TF-IDF	Trigram +TF	Trigram +TP	Trigram +TF-IDF
KNN	0.77	0.77	0.77	0.76	0.77	0.77	0.71	0.76	0.75
SVM	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
LR	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
NB	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
RF	0.80	0.80	0.80	0.80	0.80	0.80	0.79	0.79	0.79
AdaBoost (KNN)	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
AdaBoost (SVM)	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
AdaBoost (LR)	0.83	0.83	0.83	0.83	0.83	0.83	0.82	0.83	0.82
AdaBoost (NB)	0.84	0.84	0.84	0.83	0.84	0.84	0.83	0.83	0.83
AdaBoost (RF)	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
Bagging (KNN)	0.81	0.81	0.81	0.80	0.81	0.81	0.80	0.80	0.80
Bagging (SVM)	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Bagging (LR)	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Bagging (NB)	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
Bagging (RF)	0.82	0.82	0.82	0.81	0.82	0.81	0.81	0.81	0.81
RS (KNN)	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
RS (SVM)	0.87	0.87	0.87	***0.87***	0.87	***0.87***	0.86	***0.87***	0.86
RS (LR)	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.86
RS (NB)	**0.90**	**0.91**	**0.90**	**0.88**	**0.89**	**0.89**	**0.88**	**0.88**	**0.88**
RS (RF)	***0.88***	***0.88***	***0.88***	***0.87***	***0.88***	***0.87***	***0.87***	***0.87***	***0.87***
Voting (Minimum probability)	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Voting (Maximum probability)	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Voting (Majority voting)	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
Voting (Product of probability)	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Voting (Average of probabilities)	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
Stacking	0.86	0.86	0.86	0.86	0.86	0.86	0.85	0.86	0.86

Abbreviations: KNN, k-nearest neighbor algorithm; LR, logistic regression; NB, Naïve Bayes; RF, random forest; RS, random subspace method; SVM, support vector machines; TF, term frequency; TF-IDF, frequency-inverse document frequency; TP, term presence.
Note: The highest values have been indicated by bold and second highest values have been indicated by italics.

the experimental evaluations. For evaluation task on machine learning methods, all the experimental procedures have been conducted on WEKA 3.9, with the default parameters being considered.

For evaluation task on deep learning-based architectures, the text corpus has been represented by three word-embedding schemes (namely word2vec, fastText, and GloVe) in conjunction with five deep-learning architectures (namely convolutional neural network, recurrent neural network, bidirectional recurrent neural network with attention mechanism, gated recurrent unit, and long short-term memory). To implement and train deep learning architectures, we have utilized Tensorflow and Keras. To obtain the optimal predictive performance

from the models, we have employed hyperparameter optimization, based on Bayesian optimization using Gaussian process. For word2vec and fastText word embeddings, continuous skip-gram and continuous bag of words (CBOW) methods have been evaluated with varying vector sizes (vector size of 200 and 300) and different dimensions for projection layers (dimension size of 100 and 200). For the corpus, 80% of data have been utilized as the training set, whereas the rest of data have been utilized as the testing set.

4.3 | Experimental results

In this section, classification accuracy and F-measure values obtained by conventional supervised learning methods, ensemble learning methods, and deep-learning architectures have been presented.

Table 2 presents the classification accuracy values obtained by supervised learning algorithms and ensemble learning methods on nine configurations of text corpus. We have considered five widely utilized supervised learning algorithms (i.e., Naïve Bayes, support vector machines, logistic regression, K-nearest neighbor, and random forest algorithm) and five ensemble learning methods (i.e., AdaBoost, Bagging, Random Subspace, voting, and Stacking) in the empirical analysis. Regarding the predictive performance of supervised learning methods, the highest predictive performance in terms of classification accuracy has been obtained by Naïve Bayes algorithm. The second highest predictive performance has been obtained by random forest algorithm, and support vector machines have obtained the third highest predictive performance. The classification accuracies obtained by ensemble learning methods indicate that ensemble learners outperform the conventional supervised learners for sentiment classification task on EDM. Random Subspace algorithm outperforms the other ensemble methods analyzed in the empirical evaluations. Stacking algorithm also yields higher predictive performances, compared with the other ensemble methods, Bagging, AdaBoost, and voting. Regarding the predictive performance of conventional text representation schemes on EDM, we have evaluated nine different configurations. The unigram features with TF-based representation have obtained the highest classification accuracies. The unigram features with TF-IDF weighting have obtained the second highest predictive performances, and unigram features have obtained the third highest predictive performances with TP. As it can be seen from the results listed in Table 2, the unigram model outperforms bigram and trigram model, and TF-based representation outperforms TP and TF-IDF-based weighting schemes.

The highest predictive performance (with a classification accuracy of 89.62%) among the compared schemes has been obtained by a random subspace ensemble of Naïve Bayes, when unigram features with TF-based representation have represented text corpus. In Table 3, F-measure values obtained by the supervised learning algorithms and ensemble learning methods have been presented. As it can be observed from the results listed in Table 3, the patterns obtained by classification algorithms and ensemble learning methods in terms of classification accuracies are also valid for the results obtained in terms of F-measure. Regarding the predictive performances of supervised learning algorithms, Naïve Bayes algorithm and random forest algorithm outperform the

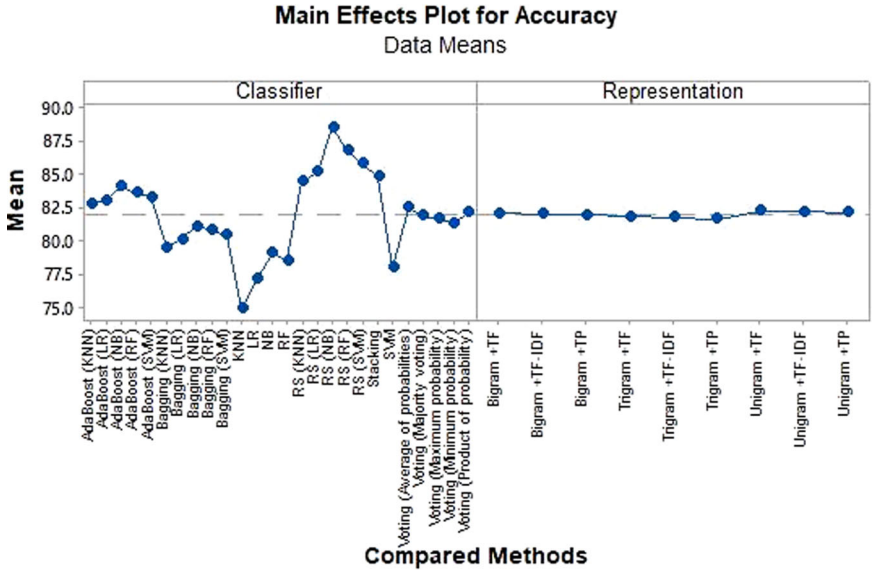

FIGURE 1 Main effects plot for classification accuracy values on machine learning-based sentiment analysis methods

TABLE 4 Classification accuracy values obtained by deep learning algorithms

Word embedding	Vector size	Dimension of projection layer	CNN	RNN	LSTM	GRU	RNN-AM
word2vec (Skip-gram)	200	100	84.14	85.52	**89.42**	86.54	***87.63***
word2vec (Skip-gram)	200	200	84.28	85.58	**89.56**	86.57	***87.66***
word2vec (Skip-gram)	300	100	84.37	85.65	**89.81**	86.63	***87.76***
word2vec (Skip-gram)	300	300	84.48	85.69	**89.98**	86.75	***87.82***
word2vec (CBOW)	200	100	84.50	85.73	**90.02**	86.78	***87.86***
word2vec (CBOW)	200	200	84.52	85.78	**90.04**	86.82	***87.94***
word2vec (CBOW)	300	100	84.63	85.86	**90.15**	86.87	***87.95***
word2vec (CBOW)	300	300	84.66	85.88	**90.32**	86.89	***88.00***
fastText (Skip-gram)	200	100	84.75	85.90	**90.38**	86.92	***88.09***
fastText (Skip-gram)	200	200	84.85	85.94	**90.59**	86.94	***88.17***
fastText (Skip-gram)	300	100	84.93	85.99	**90.83**	87.00	***88.20***
fastText (Skip-gram)	300	300	85.01	86.01	**90.96**	87.01	***88.37***
fastText (CBOW)	200	100	85.08	86.10	**91.05**	87.06	***88.46***
fastText (CBOW)	200	200	85.13	86.15	**91.24**	87.09	***88.60***
fastText (CBOW)	300	100	85.15	86.24	**91.36**	87.17	***88.77***
fastText (CBOW)	300	300	85.19	86.25	**91.64**	87.25	***88.82***
GloVe	200	100	85.22	86.29	**91.77**	87.41	***88.93***
GloVe	200	200	85.30	86.34	**92.87**	87.45	***88.98***
GloVe	300	100	85.36	86.41	**93.91**	87.53	***89.04***
GloVe	300	300	85.46	86.46	**95.80**	87.58	***89.25***

Abbreviations: CBOW, continuous bag of words; CNN, convolutional neural network, GRU, gated recurrent units; LSTM, long short-term memory networks; RNN, recurrent neural network; RNN-AM, recurrent neural network with attention mechanism.
Note: The highest values have been indicated by bold and second highest values have been indicated by italics.

other classification algorithms. The predictive performance of conventional supervised learning algorithms has been enhanced with the use of ensemble learners. The highest predictive performance in terms of F-measure has been obtained with a random subspace ensemble of Naïve Bayes with an F-measure of 0.91. In Figure 1, the main effects plot for accuracy results has been presented to summarize the main findings of the empirical results on machine learning-based sentiment analysis.

In Tables 4 and 5, classification accuracies and F-measure values obtained by five deep learning architectures (convolutional neural network, recurrent neural network, long short-term memory, gated recurrent unit, and recurrent neural network with attention mechanism) on three word-embedding schemes have been presented, respectively.

Five word embedding-based representation schemes (namely word2vec skip-gram model, word2vec continuous bag of words (CBOW) model, fastText skip-gram model, fastText continuous bag of words (CBOW) model, and GloVe) have been considered here. The predictive performance results (in terms of classification accuracy) listed in Table 4 indicate that GloVe word-embedding scheme yields a higher predictive performance, compared with the other word-embedding schemes. The second highest predictive performances have been obtained by fastText CBOW model, which is followed by fastText skip-gram model. The lowest predictive performances in terms of classification accuracies have been obtained by word2vec skip-gram model. For the different vector sizes and dimensions of projection layers considered in the empirical analysis, a vector size of 300 and a dimension projection layer of 300 yield a higher predictive performance.

Regarding the predictive performances of deep learning-based architectures for sentiment analysis, long short-term memory networks (LSTM) have obtained the highest predictive performances. The second highest predictive performances have been generally achieved

TABLE 5 F-measure values obtained by deep learning algorithms

Word Embedding	Vector size	Dimension of projection layer	CNN	RNN	LSTM	GRU	RNN-AM
word2vec (Skip-gram)	200	100	0.85	0.86	**0.88**	0.86	***0.87***
word2vec (Skip-gram)	200	200	0.85	0.86	**0.91**	0.86	***0.87***
word2vec (Skip-gram)	300	100	0.85	0.86	**0.91**	0.86	***0.87***
word2vec (Skip-gram)	300	300	0.85	0.86	**0.91**	0.86	***0.87***
word2vec (CBOW)	200	100	0.85	0.86	**0.91**	***0.87***	***0.87***
word2vec (CBOW)	200	200	0.85	0.86	**0.91**	***0.87***	***0.87***
word2vec (CBOW)	300	100	0.85	0.86	**0.91**	***0.87***	***0.87***
word2vec (CBOW)	300	300	0.85	0.86	**0.91**	***0.87***	***0.87***
fastText (Skip-gram)	200	100	0.85	0.86	**0.91**	***0.87***	***0.87***
fastText (Skip-gram)	200	200	0.85	0.86	**0.92**	***0.87***	***0.87***
fastText (Skip-gram)	300	100	0.85	0.86	**0.92**	***0.87***	***0.87***
fastText (Skip-gram)	300	300	0.85	0.86	**0.92**	***0.87***	***0.87***
fastText (CBOW)	200	100	0.85	0.86	**0.92**	0.87	***0.88***
fastText (CBOW)	200	200	0.85	0.86	**0.92**	0.87	***0.88***
fastText (CBOW)	300	100	0.85	0.86	**0.92**	0.87	***0.88***
fastText (CBOW)	300	300	0.85	0.86	**0.93**	0.87	***0.88***
GloVe	200	100	0.85	0.86	**0.93**	0.87	***0.88***
GloVe	200	200	0.86	0.86	**0.93**	0.87	***0.88***
GloVe	300	100	0.86	0.86	**0.94**	0.87	***0.88***
GloVe	300	300	0.86	0.86	**0.96**	0.87	***0.88***

Abbreviations: CBOW, continuous bag of words; CNN, convolutional neural network, GRU, gated recurrent units; LSTM, long short-term memory networks; RNN, recurrent neural network; RNN-AM, recurrent neural network with attention mechanism.
Note: The highest values have been indicated by bold and second highest values have been indicated by italics.

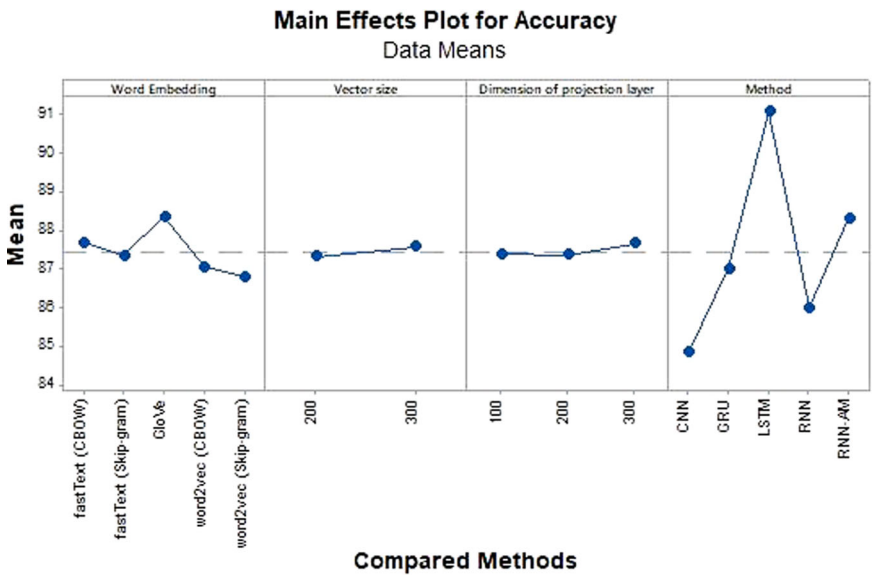

FIGURE 2 Main effects plot for classification accuracy values on deep learning-based sentiment analysis methods

by recurrent neural network with attention mechanism (RNN-AM). Gated recurrent units (GRU) have generally obtained the third highest predictive performances. The empirical results listed in Tables 4 and 5 indicate that LSTM, GRU, and RNN-AM architectures outperform the conventional recurrent neural network. For the task of sentiment analysis on EDM, the lowest predictive performance has been achieved by the convolutional neural network architecture. For all the compared configurations, long short-term memory networks (LSTM) have achieved the highest predictive performance with GloVe word-embedding scheme-based representation, with a classification accuracy of 95.80%. In the empirical results, we seek to identify whether deep learning architectures outperform the conventional supervised learning methods and ensemble learning methods. Deep learning methods outperform conventional supervised learning methods and ensemble learning methods. In addition, ensemble learning methods outperform conventional supervised learning methods. In Figure 2, the main effects plot for accuracy results has been presented to summarize the main findings of the empirical results on deep learning-based sentiment analysis.

To further evaluate the statistical significance of the results obtained in the empirical analysis, we have performed one-way analysis of variance (ANOVA) tests in Minitab statistical program. In Table 6, the statistical significance of results for conventional supervised learning methods and ensemble learning methods listed in Tables 2 and 3 has been evaluated, where *DF*, SS, MS, *F*, and *p* denote degrees of freedom, adjusted sum of squares, adjusted mean square, F-Value, and probability value, respectively. According to the one-way ANOVA test results presented in Table 6, the higher predictive performances obtained by ensemble learning methods are statistically significant ($p < .0001$). Similarly, there is a statistically meaningful difference between the results obtained by conventional text representation schemes.

In Table 7, the statistical significance of results for different word-embedding schemes and deep learning algorithms listed in Tables 4 and 5 has been evaluated. As it can be observed from the results presented in Table 7, there is a statistically meaningful difference in the predictive performance between different word-embedding schemes and deep learning architectures ($p < .0001$). In Figure 3, the confidence intervals for the mean values of supervised learning methods, ensemble learning algorithms, and deep learning architectures for a confidence level of 95% are presented. Based on the statistical significances between the results, the figure has been divided into three regions denoted by red dashed lines. The predictive performances obtained by conventional supervised learning methods, ensemble learning methods, and deep learning algorithms are statistically meaningful.

TABLE 6 One-way analysis of variance test results for compared classifiers and ensemble learning methods

Accuracy values					
Source	**DF**	**SS**	**MS**	***F*-value**	***p*-Value**
Classifier	25	2177.08	87.0833	1274.41	.000
Text representation	8	8.37	1.0467	15.32	.000
Error	200	13.67	0.0683		
Total	233	2199.12			
F-measure values					
Source	**DF**	**SS**	**MS**	***F*-value**	***p*-Value**
Classifier	25	0.209637	0.008385	467.51	.000
Text Representation	8	0.001050	0.000131	7.32	.000
Error	200	0.003587	0.000018		
Total	233	0.214275			

TABLE 7 One-way analysis of variance test results for deep learning-based sentiment analysis

Accuracy values					
Source	**DF**	**Adj SS**	**Adj MS**	***F*-value**	***p*-Value**
Word embeddings	4	29.912	7.478	24.39	.000
Deep learning architecture	4	461.210	115.302	376.04	.000
Error	91	27.903	0.307		
Lack of fit	16	17.910	1.119	8.40	.000
Pure error	75	9.993	0.133		
Total	99	519.024			
F-measure values					
Source	**DF**	**Adj SS**	**Adj MS**	***F*-value**	***p*-Value**
Word embeddings	4	0.002036	0.000509	9.83	.000
Deep learning architecture	4	0.050651	0.012663	244.62	.000
Error	91	0.004711	0.000052		
Lack of fit	16	0.002440	0.000153	5.04	0.000
Pure error	75	0.002270	0.000030		
Total	99	0.057398			

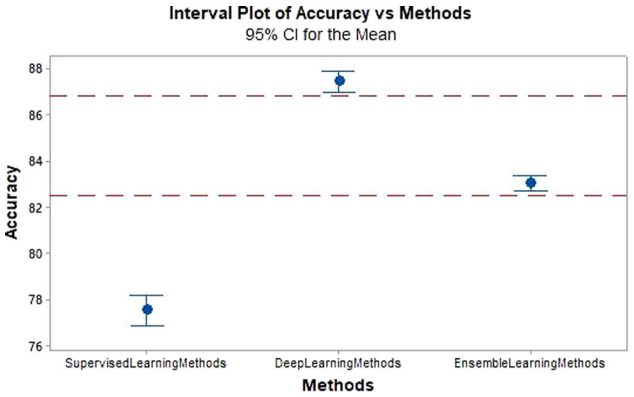

FIGURE 3 Interval plot of accuracy values for compared methods

5 | CONCLUSION

Massive open online courses (MOOCs) are recent innovative approaches in distance education, which provide learning content to participants without age-, gender-, race-, or geography-related barriers. Sentiment analysis on educational data can be employed to obtain feedback on learning content, which can help teachers to improve their teaching process and provide learners to reach the high-quality educational resources.

In this paper, we have analyzed a corpus containing 66,000 MOOC reviews, with the use of machine learning, ensemble learning, and deep learning methods. In the machine learning-based approach, we utilized three term weighting schemes (i.e., TP, TF, and TF-IDF). The representation schemes have been evaluated in conjunction with five supervised learners (i.e., Naïve Bayes, support vector machines, logistic regression, k-nearest neighbor, and random forest) and five ensemble learning methods (i.e., AdaBoost, Bagging, Random Subspace, voting, and Stacking). In the deep learning-based approach, we have utilized three word-embedding schemes (i.e., word2vec, fastText, and GloVe) in conjunction with five deep learning architectures (i.e., convolutional neural network, recurrent neural network, bidirectional recurrent neural network with attention mechanism, gated recurrent unit, and long short-term memory).

This research aims to present an efficient sentiment classification scheme with a high predictive performance in the educational domain, by pursuing the paradigms of ensemble learning and deep learning. To the best of our knowledge, this is the first comprehensive empirical study on sentiment analysis of MOOC reviews, in which the predictive performances of conventional classification algorithms, ensemble learning methods, and deep learning algorithms have been reported.

This research has some theoretical and practical implications. The results of the empirical analysis indicate that ensemble learning methods yield a higher predictive performance in educational data mining, compared with the conventional supervised learning methods. Regarding the performance of different ensemble learners, the random subspace ensemble outperforms the other ensemble methods. Regarding the empirical results obtained by conventional text representation schemes, the highest predictive performances have been obtained by unigram sets. For evaluation tasks, TF, a term weighting scheme, outperforms TF-IDF and TP-based representations. The results on MOOC reviews indicate that deep learning architectures outperform conventional supervised learning methods and ensemble learning methods. For the deep learning architectures, long short-term memory networks (LSTM) have obtained the highest predictive performances. The second highest predictive performances have been generally achieved by recurrent neural network with attention mechanism (RNN-AM). Gated recurrent units (GRU) have generally obtained the third highest predictive performances. The empirical analysis indicates that GloVe word-embedding scheme yields a higher predictive performance, compared with the other word-embedding schemes. The second highest predictive performances have been obtained by fastText CBOW model, which is followed by fastText skip-gram model.

There are several practical implications of the research. The identification of an appropriate representation scheme is a critical issue in developing machine learning-based sentiment classification schemes. In this regard, the experimental analysis presents comprehensive empirical results for different text representation schemes, supervised learning methods, ensemble learning models, and deep learning architectures for educational data mining, which may be utilized as baseline empirical results for the field. In addition, we present first corpus on massive open online course reviews, which may be helpful for further research.

ORCID

Aytuğ ONAN http://orcid.org/0000-0002-9434-5880

REFERENCES

1. A. Abdi, S. M. Shamsuddin, and R. M. Aliguliyev, *QMOS: Query-based multi-documents opinion-oriented summarization*, Inform. Process. Manag. **54** (2018), no. 2, 318–338.
2. P. Adamopoulos, 2013. *What makes a great MOOC? An interdisciplinary analysis of student retention in online courses*, in: International Conference on Information Systems, ICIS 2013.
3. P. Adinolfi et al., *Sentiment analysis to evaluate teaching performance*, Int. J. Knowl. Soc. Res. (IJKSR) **7** (2016), no. 4, 86–107.

4. D. W. Aha, D. Kibler, and M. K. Albert, *Instance-based learning algorithms*, Mach. Learn. **6** (1991), no. 1, 37–66.
5. N. Alami, M. Meknassi, and N. En-nahnahi, *Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning*, Expert Syst. Appl. **123** (2019), 195–211.
6. C. Alario-Hoyos et al., *Delving into participants' profiles and use of social tools in MOOCs*, IEEE Transact. Learn. Technol. **7** (2014), no. 3, 260–266.
7. H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, *Educational data mining and learning analytics for 21stcentury higher education: A review and synthesis*, Telematics Inform. **39** (2019), no. 4, 13–49.
8. M. Al-Smadi et al., *Enhancing aspect-based sentiment analysis of Arabic hotels' reviews using morphological, syntactic and semantic features*, Inform. Process. Manag. **56** (2019), no. 2, 308–319.
9. N. Altrabsheh, M. Cocea, and S. Fallahkhair (2015). *Predicting learning-related emotions from students' textual classroom feedback via Twitter*. Paper presented at the International Conference on Educational Data Mining (EDM) (8th, Madrid, Spain, Jun 26-29, 2015)
10. F. Belbachir and M. Boughanem, *Using language models to improve opinion detection*, Inform. Process. Manag. **54** (2018), no. 6, 958–968.
11. Y. Bengio et al., *A neural probabilistic language model*, J. Mach. Learn. Res. **3** (2003), no. Feb, 1137–1155.
12. R. Bogdan, *Sentiment analysis on embedded systems blended courses*, BRAIN. Broad Res. Artif. Intell. Neurosci. **8** (2017), no. 1, 17–23.
13. L. Breiman, *Bagging predictors*, Mach. Learn. **24** (1996), no. 2, 123–140.
14. L. Breiman, *Random forests*, Mach. Learn. **45** (2001), no. 1, 5–32.
15. L. Breslow et al., *Studying learning in the worldwide classroom research into edX's first MOOC*, Res. Pract. Assess. **8** (2013), 13–25.
16. R. Z. Cabada, M. L. B. Estrada, and R. O. Bustillos, *Mining of educational opinions with deep learning*, J. Univers. Comput. Sci. **24** (2018), no. 11, 1604–1626.
17. G. Chakraborty, M. Pagolu, and S. Garla, 2013. *Text mining and analysis*, in: Text Mining and Analysis. Practical Method, Examples and Case Studies Using SAS.
18. K. Cho et al., *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, arXiv preprint arXiv **1406** (2014), 1078–1734.
19. D. Ciregan, U. Meier, and J. Schmidhuber, 2012. *Multi-column deep neural networks for image classification*, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2012.6248110
20. W. Di, A. Bhardwaj, and J. Wei, Deep learning essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling, Packt Publishing, Birmingham, UK, 2018.
21. S. Donitsa-Schmidt and B. Topaz, *Massive open online courses as a knowledge base for teachers*, Journal of Education for Teaching **44** (2018), no. 5, 608–620.
22. L. P. Dringus and T. Ellis, *Using data mining as a strategy for assessing asynchronous discussion forums*, Comput. Educ. **45** (2005), no. 1, 141–160.
23. J. L. Elman, *Finding structure in time*, Cogn. Sci. **14** (1990), no. 2, 179–211.
24. Y. Freund, and R. E. Schapire, *Experiments with a new boosting algorithm*, Machine Learning: Proceedings of the Thirteenth International Conference, 96 (1996), 148–156.
25. X. Glorot, A. Bordes, and Y. Bengio (2011). *Domain adaptation for large-scale sentiment classification: A deep learning approach*. In Proceedings of the 28th International Conference on Machine Learning (ICML-**11**) (pp. 513-520).
26. G. Gutiérrez et al., *Mining: Students comments about teacher performance assessment using machine learning algorithms*, Int. J. Combin. Optim. Prob. Inform. **9** (2018), no. 3, 26–40.
27. G. Hackeling, Mastering machine learning with scikit-learn, Packt Publishing Ltd, Birmingham, UK, 2017.
28. T. Hastie, R. Tibsharani, and J. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*, **27**., 2nd., Springer, Berlin, 2009, pp. 83–85.
29. T. K. Ho, *The random subspace method for constructing decision forests*, IEEE. Trans. Pattern. Anal. Mach. Intell. **20** (1998), no. 8, 832–844.
30. A. Ho et al. (2014). *HarvardX and MITx: The first year of open online courses, fall 2012-summer 2013*. Ho, AD, Reich, J., Nesterko, S., Seaton, DT, Mullaney, T., Waldo, J., & Chuang, I. (2014). HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1).
31. M. Jia et al., *Who can benefit more from massive open online courses? A prospective cohort study*, Nurse Educ. Today **76** (2019), 96–102.
32. A. Joulin et al., *Fasttext. zip: Compressing text classification models*, arXiv preprint arXiv **1612** (2016), 03651.
33. A. M. Kaplan and M. Haenlein, *Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster*, Bus. Horiz. **59** (2016), no. 4, 441–450.
34. I. U. Khan et al., *Predicting the acceptance of MOOCs in a developing country: application of task-technology fit model, social motivation, and self-determination theory*, Telematics Inform. **35** (2018), no. 4, 964–978.
35. V. Kovanović et al., *Exploring communities of inquiry in massive open online courses*, Comput. Educ. **119** (2018), 44–58.
36. P. C. Lane, D. Clarke, and P. Hender, *On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data*, Decis. Support Syst. **53** (2012), no. 4, 712–718.
37. Y. LeCun, *Generalization and network design strategies*, Connectionism in perspective, **19**, Elsevier, Amsterdam, The Netherlands, 1989.
38. P. J. Lee, Y. H. Hu, and K. T. Lu, *Assessing the helpfulness of online hotel reviews: A classification-based approach*, Telematics Inform. **35** (2018), no. 2, 436–445.
39. D. D. Lewis, *Naive (Bayes) at forty: The independence assumption in information retrieval*, European conference on machine learning, Springer, Berlin, Heidelberg, 1998, pp. 4–15.
40. Q. Lin et al., *Lexical based automated teaching evaluation via students' short reviews*, Comput. Appl. Eng. Educ. **27** (2019), no. 1, 194–205.
41. E. Loper and S. Bird, 2002. *The natural language toolkit NLTK: The Natural Language Toolkit*, in: Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics, available at https://doi.org/10.3115/1225403.1225421

42. M. B. López et al., *EduRP: An educational resources platform based on opinion mining and semantic web*, J. Univers. Comput. Sci. **24** (2018), no. 11, 1515–1535.
43. W. D. Maxwell et al., *Massive open online courses in US healthcare education: Practical considerations and lessons learned from implementation*, Curr. Pharm. Teach. Learn. **10** (2018), no. 6, 736–743.
44. W. Medhat, A. Hassan, and H. Korashy, *Sentiment analysis algorithms and applications: A survey*, Ain Shams Eng. J. **5** (2014), no. 4, 1093–1113.
45. T. Mikolov et al., *Efficient estimation of word representations in vector space*, arXiv preprint arXiv **1301** (2013), 3781.
46. P. M. Moreno-Marcos et al. (2018, April). *Sentiment analysis in MOOCs: A case study*. In 2018 IEEE Global Engineering Education Conference (EDUCON) (pp. 1489-1496). IEEE.
47. P. J. Muñoz-Merino et al., *Precise effectiveness strategy for analyzing the effectiveness of students with educational resources and activities in MOOCs*, Comput. Human. Behav. **47** (2015), 108–118.
48. H. T. Nguyen and M. Le Nguyen, *Multilingual opinion mining on YouTube–A convolutional N-gram BiLSTM word embedding*, Inform. Process. Manag. **54** (2018), no. 3, 451–462.
49. A. Onan, *An ensemble scheme based on language function analysis and feature engineering for text genre classification*, J. Inform. Sci. **44** (2018), no. 1, 28–47.
50. A. Onan, *Mining opinions from instructor evaluation reviews: A deep learning approach*, Comput. Appl. Eng. Educ. **28** (2020), no. 1, 117–138.
51. A. Onan, S. Korukoğlu, and H. Bulut, *Ensemble of keyword extraction methods and classifiers in text classification*, Expert Syst. Appl. **57** (2016), 232–247.
52. A. Onan, S. Korukoğlu, and H. Bulut, *A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification*, Expert Syst. Appl. **62** (2016), 1–16.
53. R. Oramas Bustillos et al., *Opinion mining and emotion recognition in an intelligent learning environment*, Comput. Appl. Eng. Educ. **27** (2019), no. 1, 90–101.
54. B. Pang, and L. Lee, *Opinion mining and sentiment analysis*, Foundations and Trends® in Information Retrieval 2 (2008), no. 1–2, 1–135.
55. J. Pennington, R. Socher, and C. Manning (2014). *Glove: Global vectors for word representation*. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
56. A. Peña-Ayala, *Educational data mining: A survey and a data mining-based analysis of recent works*, Expert systems with applications **41** (2014), no. 4, 1432–1462.
57. M. F. Porter (2001). Snowball: A language for stemming algorithms.
58. R. Prabowo and M. Thelwall, *Sentiment analysis: A combined approach*, Journal of Informetrics **3** (2009), no. 2, 143–157.
59. S. M. Rezaeinia et al., *Sentiment analysis based on improved pre-trained word embeddings*, Expert Systems with Applications **117** (2019), 139–147.
60. J. A. Rodger, *A fuzzy nearest neighbor neural network statistical model for predicting demand for natural gas and energy cost savings in public buildings*, Expert Syst. Appl. **41** (2014), no. 4, 1813–1829.
61. J. A. Rodger, *Discovery of medical Big Data analytics: Improving the prediction of traumatic brain injury survival rates by data mining patient informatics processing software Hybrid Hadoop Hive*, Inform. Med. Unlocked **1** (2015), 17–26.
62. L. M. Rojas-Barahona, *Deep learning for sentiment analysis*, Lang. Linguist. Compass **10** (2016), no. 12, 701–719.
63. C. L. Santos, P. Rita, and J. Guerreiro, *Improving international attractiveness of higher education institutions based on text mining and sentiment analysis*, Int. J. Educ. Manag. **32** (2018), no. 3, 431–447.
64. B. Shah et al. (2019). *Twitter Analysis for Disaster Management*. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-4). IEEE.
65. G. Siemens and R. S. Baker (2012, April). *Learning analytics and educational data mining: towards communication and collaboration*. In Proceedings of the 2nd international conference on learning analytics and knowledge (pp. 252-254). ACM.
66. F. Tian et al., *Recognizing and regulating e-learners' emotions based on interactive Chinese texts in e-learning systems*, Knowl.-Based Syst. **55** (2014), 148–164.
67. N. D. Valakunde and M. S. Patwardhan (2013, November). *Multi-aspect and multi-class based document sentiment analysis of educational data catering accreditation process*. In 2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies (pp. 188-192). IEEE.
68. V. Vapnik, The nature of statistical learning theory, Springer science & business media, 2013.
69. J. Wang, B. Peng, and X. Zhang, *Using a stacked residual LSTM model for sentiment intensity prediction*, Neurocomputing **322** (2018), 93–101.
70. J. Wang et al., *Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis.* IEEE/ACM Transactions on, Audio, Speech, and Language Processing **28** (2019), 581–591.
71. M. Wen, D. Yang, and C. Rose (2014, July). Sentiment Analysis in MOOC Discussion Forums: What does it tell us? In *Educational data mining 2014*.
72. D. H. Wolpert, *Stacked generalization*, Neural Netw. **5** (1992), no. 2, 241–259.
73. L. Zhang, S. Wang, and B. Liu, *Deep learning for sentiment analysis: A survey*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8** (2018), no. 4, e1253.
74. N. Öztürk and S. Ayvaz, *Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis*, Telematics and Informatics **35** (2018), no. 1, 136–147.

AUTHOR BIOGRAPHY

Aytuğ Onan received the BS degree in computer engineering from the Izmir University of Economics, İzmir, Turkey, in 2010, and the MS degree in computer engineering and the PhD degree in computer engineering from Ege University, Turkey, in 2013 and 2016, respectively. He has been an Associate Professor with the Department of Computer Engineering, Izmir Katip Celebi University, Izmir,

Turkey, since April 2019. He has published several journal articles on machine learning and computational linguistics. Dr. Onan has been an editor for the KSII Transactions on Internet and Information Systems and an Associate Editor for the Journal of King Saud University Computer and Information Sciences.

How to cite this article: Onan A. Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. *Comput Appl Eng Educ*. 2020;1–18. https://doi.org/10.1002/cae.22253