


RESEARCH ARTICLE

WILEY

Mining opinions from instructor evaluation reviews: A 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@ikcu.edu.tr

Abstract

Student evaluations of teaching (SET) provides potentially essential source of information to achieve educational quality objectives of higher educational institutions. The findings can be utilized as a measure of teaching effectiveness and they may aid the administrative decision-making process. The purpose of our research is to establish an efficient sentiment classification scheme on instructor evaluation reviews by pursuing the paradigm of deep learning. Deep learning is a recent research direction of machine learning, which seeks to identify a classification scheme with higher predictive performance based on multiple layers of nonlinear information processing. In this study, we present a recurrent neural network (RNN) based model for opinion mining on instructor evaluation reviews. We analyze a corpus containing 154,000 such reviews, with the use of conventional machine learning algorithms, ensemble learning methods, and deep learning architectures. In the empirical analysis, three conventional text representation schemes (namely, term-presence, term-frequency [TF], and TF-inverse document frequency schemes) and four word embedding schemes (namely, word2vec, global vector [GloVe], fastText, and LDA2Vec) have been taken into consideration. The predictive performance of supervised machine learning methods (such as, Naïve Bayes, support vector machines, logistic regression, K-nearest neighbor, and random forest) and three ensemble learning methods have been examined on word embedding schemes. The extensive empirical analysis indicates that deep learning-based architectures outperform the conventional machine learning classifiers for the task of sentiment classification on instructor reviews. For the RNN with attention mechanism in conjunction with GloVe word embedding scheme-based representation a classification accuracy of 98.29% has been obtained.

KEYWORDS

deep learning, machine learning, sentiment analysis, student evaluations of teaching

1 | INTRODUCTION

Student evaluations of teaching (SET) is a widely utilized method on higher educational institutions to gather information about the professional competence of

instructors. Student evaluations may contain questions regarding teaching approaches, assessment, resources, and administration, such as, the sufficiency of the course content, competence and qualifications of instructor, and adequacy of handouts. Evaluation forms can provide

potentially essential source of information to achieve quality objectives of educational institutions. The university administrations can utilize student evaluations as a measure of teaching effectiveness. In addition, SET can be utilized in the administrative decision-making process, such as promotion and hiring decisions [9,13]. SET instruments typically consist of several fixed-ended questions, followed by some open-ended questions. The fixed-ended questions of evaluation questionnaires have been answered by means of a rating-scale. In this way, statistical values can be computed to summarize the main findings of the evaluation. However, fixed-ended questions have been constrained by stereotypical questions provided by the evaluation forms. These questions cannot reveal all aspects of the teaching process. Lin et al [49], In this regard, open-ended questions can be viewed as an easier way of expressing the ideas, opinions or attitudes towards an entity or issue. Brockx et al [16], The evaluation comments provide a greater freedom of expression, regarding the courses and instructors and can provide more useful insights to improve the instructional content.

With the progress in information and communication technologies, the immense quantity of user-generated information has been shared on the web. These user-generated text documents include hotel reviews [82], product reviews [23], movie reviews [75], and doctor reviews [80], among the others. Students can also share anonymous evaluations of instructors on online platforms, such as, Ratemyprofessors.com and mysupervisor.com. Ratemyprofessors.com is one of the largest online platforms for SET, with over 19 million ratings, 1.7 million professors, and approximately 7,500 schools from several countries, such as, the United States, Canada, and the United Kingdom [65]. On that platform, students can evaluate instructors with the use of 5-point rating-scale, from which an overall rating has been computed. In addition, students can share their opinions at the open-ended part of the platform. The platform can be regarded as an important means of communication among students to obtain information about instructors, during the course enrollment process [34]. From the viewpoint of instructors, the platform provides valuable information related to the self-efficacy and interpersonal communication skills of instructors [12]. Hence, the information can serve as a feedback to improve those skills.

Sentiment analysis (also, known as, opinion mining) is the computational field of study to identify people's opinions, sentiments, attitudes, evaluations, or emotions towards an entity or subject, such as, products, services, organizations, individuals, issues, and events [61]. Sentiment analysis can be utilized to obtain a structured and insightful knowledge from unstructured contents, and

this knowledge can serve as an essential source of information for decision support systems and individual decision makers [29]. Sentiment analysis can be applied in a wide range of application fields, including education. In the field of education, sentiment analysis may be utilized to improve the international attractiveness of higher educational institutions [68], may be utilized to recognize and regulate e-learners' emotions [77], visual perception and learning performance of students.[41]

Sentiment analysis methods can be mainly divided into two categories, as lexicon-based approaches and machine learning-based approaches [51]. The lexicon-based approaches to sentiment analysis identify sentiment orientation of a text document by computing semantic orientation of words and phrases. In this regard, these schemes require a dictionary with positive and negative sentiment values for each word. The lexicon-based sentiment analysis methods vary according to the context in which they were constructed. The lexicon-based methods do not involve labeled data. However, it is hard to construct a unique lexical-based dictionary for different contexts [33]. Sentiment words included in lexicon-based analysis methods are not generally specific to a particular topic [32].

The machine learning-based approaches model the process of sentiment classification as a supervised learning task, where a set of labeled text documents have been utilized to construct the learning model. These schemes utilize conventional supervised learning algorithms, such as Naïve Bayes (NB) algorithm, support vector machines (SVMs), and K-nearest neighbor (KNN) algorithm, on the task of sentiment classification [58]. The earlier research studies on sentiment analysis indicate that machine learning-based schemes generally yield higher predictive performance [11,20,60]. For instance, Bermingham and Smeaton [11] obtained higher predictive performance on sentiment analysis on micro-blogging platforms with the use of machine learning-based methods, such as SVMs and multinomial NB algorithm compared with SentiWordNet.

In addition to conventional classification algorithms, ensemble learning methods and deep learning architectures may be utilized for sentiment analysis. Ensemble learning is a field of machine learning, which attempts to obtain a classification scheme with higher predictive performance by combining the predictions of multiple learning algorithms. Deep learning is a field of machine learning, which attempt to obtain a classification model with higher predictive performance based on multiple layers or stages of nonlinear information processing and supervised or unsupervised learning of feature representations in a hierarchical way [24]. Recent literature on the machine learning-based sentiment analysis indicate that

the predictive performance of sentiment classification schemes can be improved with the use of ensemble learning and deep learning architectures [31,63].

In this paper, we present a text mining approach for opinion mining on instructor evaluation reviews. We analyze a corpus containing 154,000 such reviews, with the use of conventional machine learning algorithms, ensemble learning methods, and deep learning architectures. In the empirical analysis, three conventional text representation schemes (namely, term-presence [TP], term-frequency [TF], and TF-inverse document frequency [TF-IDF] schemes) and four word embedding schemes (namely, word2vec, global vector [GloVe], fastText, and LDA2Vec) have been taken into consideration. The predictive performance of supervised machine learning methods (such as, NB, SVMs, logistic regression [LR], KNN, and Random Forest) and three ensemble learning methods (namely, AdaBoost, Bagging, and Random Subspace) have been evaluated on conventional text representation schemes. In addition, five deep learning architectures (namely, convolutional neural network [CNN], recurrent neural network (RNN), bidirectional RNN with attention mechanism (RNN-AM), gated recurrent unit, and long short-term memory) have been examined on word embedding schemes. The extensive empirical analysis indicates that deep learning-based architectures outperform the conventional machine learning classifiers for the task of sentiment classification on instructor reviews.

The main organization of this paper consists of five sections. In Section 2, the related work on machine learning-based sentiment analysis and deep learning-based sentiment analysis have been presented. Section 3 describes the methods utilized in the paper. Namely, conventional text representation schemes, word embedding based text representation schemes, machine learning algorithms, ensemble learning methods, deep learning architectures, and corpus have been presented. Section 4 introduces the experimental procedure and experimental results on machine learning-based and deep learning-based architectures. Finally, Section 5 presents the concluding remarks of the study.

2 | RELATED WORK

Machine learning methods and deep learning architectures have been successfully employed on data mining and sentiment analysis tasks on several fields, including education domain. The subsections briefly discuss related work on the field, with emphasize on sentiment analysis on educational data.

2.1 | Machine learning-based approaches to sentiment analysis

Due to their high predictive performance, machine learning classifiers have been frequently employed for the task of sentiment classification. For instance, Adinolfi et al [3] introduced a sentiment analysis-based model to evaluate student satisfaction on different learning platforms, such as massive open online courses, learning diaries, and Twitter. In another study, Altrabsheh et al [7] presented a machine learning approach to identify learning-related emotions of students on text feedbacks. In the presented scheme, Twitter has been utilized to collect student feedbacks, opinions, and feeling about different courses, such as calculus, communication skills, database, engineering, molecular biology, chemistry, physics, and science. Twitter messages have been modeled with different text representation schemes (namely, unigram, bigram, trigram-based representation, and their combinations are considered). In the classification phase, NB algorithm, SVMs, maximum entropy classifier, and random forest (RF) algorithm have been evaluated. The empirical results indicated that SVMs with radial basis kernel outperform the other compared schemes. Similarly, Gutierrez et al [35] introduced a text mining approach to evaluate the student comments about teacher performance. In the presented scheme, SVMs, and RF algorithm have been employed for the task of sentiment analysis on student reviews. In another study, Deng and Que [25] introduced a factor analysis and association rule mining-based model to analyze SET. In the presented model, the variance analysis has been performed to examine the effects of professional titles and teaching semesters on the assessment process. Based on the factor analysis, three essential factors of evaluation have been identified. In addition to extract correlations between, essential factors of assessment, the association rule mining method has been employed. In a similar way, Lin et al [49] presented two lexical-based schemes (namely, knowledge-based and machine learning-based approaches) to identify opinions from short reviews about teaching evaluations. In the knowledge-based approach, word co-occurrence method and word embedding similarity method have been utilized to automatically construct and expand the sentiment dictionary. In the machine learning-based approach, lexical, syntactic, and semantic features have been utilized to represent short reviews. In addition, TextRank algorithm has been employed to extract keywords. The lexical feature sets have been utilized in conjunction with conventional supervised learning methods, such as NB, LR, SVMs, and gradient boost decision algorithm to empirically evaluate the predictive performance on evaluating teaching

performance. In another study, Rani and Kumar [64] presented a lexicon-based approach for sentiment analysis on student feedback. In the presented scheme, natural language processing techniques have been employed in conjunction with National Research Council (NRC) Emotion lexicon to classify sentiments and emotions on course reviews. Similarly, Anderson et al [8] employed machine learning techniques on evaluation reviews of first-year engineering students. Recently, Jena [42] presented a sentiment analysis scheme on collaborative learning environments. In the presented scheme, NB SVMs; maximum entropy classifiers have been employed in conjunction with unigram, bigram, and trigram models for sentiment polarity identification from students' data.

In addition to sentiment analysis on SET, machine learning methods have been employed in several other tasks in education data mining. For instance, Santos et al [69] employed text mining and sentiment analysis techniques to extract information regarding the drivers of higher educational institutions success online. The presented scheme based on topic modeling and topic profiling analysis aims to improve the international attractiveness of higher educational institutions with the use of sentiment analysis. In another study, Adekitan and Noma-Osaghae [1] introduced a machine learning approach to identify the performance of first-year students in university based on their admission requirements. To predict the performance of students, six supervised learning algorithms, specifically, RF algorithm, tree ensemble, decision tree, NB algorithm, LR classifier, and resilient backpropagation based multi-layer perceptron have been employed. To further validate the predictive results obtained by supervised learning algorithms, linear and quadratic regression models have been employed. In a similar way, Adekitan and Salau [2] presented a prediction model for students' performance after their graduation with the use of machine learning algorithms. In the presented model, six classification algorithms (namely, probabilistic neural network, RF, decision tree, NB, tree ensemble, and LR) have been utilized. In another study, Ullmann [78] introduced a machine learning-based approach to identify reflection in writings. In the empirical analysis, RF algorithm, belief neural networks, NB, and SVMs have been considered.

Ensemble learning has been also employed on data mining tasks to obtain more robust classification schemes with higher predictive performance [59]. For instance, Beemer et al [10] presented an ensemble classification model to identify individualized treatment effects to characterize at-risk students and to evaluate student success and retention. In another study, Almasri et al [6] introduced an ensemble tree-based model to predict performance of students.

2.2 | Deep learning-based approaches to sentiment analysis

Deep learning is a recent research direction in machine learning. Deep learning-based approaches seek to obtain classification schemes with higher predictive performance based on multiple layers or stages of nonlinear information processing [32]. Deep learning has been successfully employed for sentiment analysis tasks. For instance, Glorot et al [31] and Pang et al [60] introduced a deep learning architecture based on stacked denoising auto-encoder with sparse rectifier units for domain adaptation task of sentiment analysis. In another study, dos Santos and Gatti [70] presented a CNN-based architecture for sentiment analysis on Twitter messages. In the presented scheme, character-level, word-level and sentiment-level embeddings have been extracted. In a similar way, Tang et al [74] presented a sentiment analysis model on Twitter messages based on two feature representations. In this scheme, linguistic features (such as, the number of words with all characters in uppercase, emoticons, elongated units, sentiment lexicon, negation, punctuation, and N-grams) and sentiment-specific word embedding based feature representations have been examined. Similarly, Severyn and Moschitti [71] presented a CNN-based architecture for extracting phrase-level and message-level sentiment. In another study, Hu et al [40] introduced a deep learning architecture for sentiment analysis, where linguistic and domain knowledge have been utilized to represent features. The presented architecture has been employed on three different sentiment analysis tasks, namely, sentiment analysis on electronic product reviews, movie reviews, and hotel reviews.

Recently, deep learning has been also employed for educational data mining tasks. For instance, Bustillos et al [18] presented a deep learning-based opinion mining module for opinion mining and emotion recognition in an intelligent learning environment. In the presented scheme, several supervised learning algorithms, such as Bernoulli NB, multinomial NB, SVMs, linear SVM, stochastic gradient descent, and KNN algorithm have been examined. For deep learning-based opinion mining, CNN and long short-term memory architectures have been considered. In the empirical analysis, deep learning-based architecture reached a classification accuracy of 88.26%. In a similar way, Cabada et al [19] presented several deep learning architectures for sentiment analysis on education. The CNN architecture with long short-term memory reached a classification accuracy of 84.32%. In another study, Sultana et al [73] presented a comparative analysis of machine learning classifiers for sentiment analysis on educational data. In this scheme, SVMs, RF,

C4.5 decision tree, K-star, Bayes Net, and multilayer perceptron have been taken into account. In the empirical analysis, the highest predictive performances have been achieved by SVMs and multilayer perceptron with classification accuracies of 78.75% and 78.33%, respectively. Similarly, Nguyen et al [54] employed machine learning and deep learning techniques on the Vietnamese students' feedback corpus. For machine learning-based analysis, NB, maximum entropy have been considered. In addition, two deep learning architectures (namely, long short-term memory and bidirectional long short-term memory) have been evaluated. For machine learning classifiers, unigram and bigram feature sets have been utilized to represent text corpus. For deep learning architectures, word2vec word embedding scheme has been employed. The empirical results indicate that deep learning-based architectures yield higher predictive performance compared with the conventional machine learning classifiers. The highest predictive performance has been achieved by bidirectional long short-term memory with a classification accuracy of 89.3%. Similarly, Kandhro et al [45] employed long short-term memory deep learning architecture for sentiment analysis on students' comments.

Though the use of conventional machine learning classifiers and deep learning architectures on educational data mining take great research attention in the literature, the number of works that comprehensively examine the predictive performance of machine learning classifiers, ensemble learning methods and deep learning architectures on students' evaluation of teaching is very limited. To fill this gap, this paper presents a comprehensive empirical analysis by taking three conventional text representation schemes (namely, TP, TF, and TF-IDF schemes) and four word embedding schemes (namely, word2vec, GloVe, fastText, and LDA2vec) in conjunction with five supervised learning algorithms (i.e., NB, SVMs, LR, KNN, and RF), three ensemble learner (i.e., AdaBoost, Bagging, and Random Subspace) and five deep learning architectures (i.e., CNN, RNN, bidirectional RNN-AM, gated recurrent unit, and long short-term memory).

3 | METHODOLOGY

This section presents the methods utilized in this paper. Conventional text representation schemes, word embedding based text representation schemes, machine learning algorithms, ensemble learning methods, deep learning architectures, corpus, and evaluation measures have been briefly discussed.

3.1 | Conventional text representation schemes

Bag-of-words scheme is a widely utilized model to represent text documents. In this scheme, a text document is represented as a multiset of words encountered in the document without taking syntax, word orderings, and grammar into account [36]. Bag-of-words scheme can be successfully employed for text classification and information retrieval tasks. In the bag-of-words model, each text document has been represented based on the frequency of each word, and this representation has been utilized as a set of features to construct the learning model by the classifier. All the unique words (terms) encountered in the text document compromise the vocabulary. Based on the bag-of-words scheme, three types of weights have been frequently utilized, i.e., TP, TF, and TF-IDF.

In TP-based representation, the occurrence of words in text document has been considered. In this scheme, each document has been represented by a binary-valued feature vector, where 1 has been utilized to indicate the occurrence of a word, whereas 0 has been utilized to indicate the nonoccurrence of a term.

In TF-based representation, the number of occurrences each word encountered in a document has been counted. In term-frequency-based representation, high scoring frequency values have been assigned to the frequent words and lower scoring frequency values have been assigned for rare words. In response, frequent words have dominance over rarely encountered words. However, some domain specific words encountered rarely may be more informative about the context compared with the frequently encountered terms [17].

To handle properly with this problem, a common weighting scheme has been utilized in text classification, referred as, TF-IDF model. In this scheme, inverse document frequency has been employed to measure how rare a word across the text documents. In this way, the frequencies of words have been rescaled based on the number of occurrences in all documents and the frequently encountered terms have been penalized.

In text classification, N-gram model is also an important representation scheme. In this scheme, n-character slice of a text document has been extracted. Typical n-gram models utilized in text classification are unigram model (n-gram of size 1), bigram (n-gram of size 2), and trigram (n-gram of size 3). In the empirical analysis, nine different configurations have been obtained on the text corpus based on three different weighting schemes and three different N-gram models.

3.2 | Word embedding based text representation schemes

Bag-of-words scheme is a common and simple representation scheme for natural language processing tasks. However, this scheme suffers from two main problems. First, the semantic relations among the components of text document cannot be fully revealed in this representation. In addition, bag-of-words scheme results in a sparse data representation with high dimensional feature space. Alami et al [5], word embedding based representation is an important scheme for language modeling and feature learning, which has been successfully employed in text categorization and sentiment analysis tasks. The word embedding based representation is a convenient representation scheme to construct learning models by machine learning algorithms and deep learning architectures. Word embeddings provide a compact and more expressive representation model for text documents. In this scheme, semantic and syntactic meaning among the words have been extracted with the use of a large unsupervised sets of documents [66]. Hence, word embedding based schemes outperform conventional text representation schemes on many different natural language processing tasks. In this study, four word embedding schemes have been considered, i.e., word2vec, fastText, GloVe, and LDA2vec.

3.2.1 | word2vec model

The word2vec is a commonly utilized neural network-based scheme to learn word embeddings from text documents. It is an unsupervised and efficient method to extract semantic relationships among the words based on their co-occurrence in text documents in a corpus. To obtain word embeddings, word2vec provides two different models, namely, continuous bag of words model (CBOW) and continuous skip-gram model [54]. The CBOW model identifies the target word by taking the context of each word as the input, while the skip-gram model predicts the context words based on the target word. The skip-gram model can work properly with small amount of data and can yield promising results in representation of rare words. In contrast, the CBOW model is a faster scheme and can effectively represent more frequently encountered words. Let we denote a sequence of training words w_1, w_2, \dots, w_T with length T , the objective of skip-gram model is determined based on (1) [52,56]:

$$\arg\max_{\theta} \frac{1}{T} \sum_{t=1}^T \sum_{-C \leq j \leq C, j \neq 0} \log P_{\theta}(w_{t+j}/w_t) \quad (1)$$

where C represent the size of training context, $P(w_{t+j}/w_t)$ represents a neural network with a set of parameters denoted by θ .

3.2.2 | fastText model

The fastText is another efficient representation scheme to learn word embeddings from text documents. In this scheme, each word has been represented by breaking words into several character n-grams [44]. In this model, the internal meanings of words have been considered. In this way, the model provides a more efficient word embedding scheme for morphologically rich languages and rare words [26].

3.2.3 | Global vector model

The GloVe is an extension of word2vec to efficiently learn word embeddings from text documents. The model combines the local context-based learning of word2vec model with global matrix factorization. The model is a global log-bilinear regression model and the objective function of the model has been formulated as given by (2) [62]:

$$J = \sum_{i,j=1}^V f(X_{ij})(w_i^T \omega_j + b_i + b_j - \log X_{ij})^2. \quad (2)$$

where V denotes the vocabulary size, $w \in R^d$ represent word vectors, $\omega \in R^d$ represent context word vectors, X denote co-occurrence matrix, and X_{ij} denotes the number of times word j occurs in the context of word i . $f(X_{ij})$ denotes a weighting function and b_i, b_j are bias parameters [62].

3.2.4 | LDA2vec model

The LDA2vec model [53] is another word embedding scheme based on word2vec model and the latent Dirichlet allocation method. In this scheme, dense word vectors from the latent document-level mixture vectors have been jointly extracted based on the Dirichlet-distribution. The model provides to identify topics of text collections and topic-adjusted word vectors. In this regard, the model provides a topic-enriched word embedding scheme by linking each word to a topic. In the model, the skip-gram negative sampling has been utilized as the objective function [53,56].

3.3 | Machine learning algorithms

In the empirical analysis, the corpus has been represented by three different N-gram models (i.e., unigram, bigram, and trigram models) and three different weighting schemes (i.e., TP, TF, and TF-IDF based weighting). In this way, nine different feature sets have been constructed. To build learning models based on these

feature sets, five conventional supervised learning methods (namely, NB algorithm, SVMs, LR, KNN, and RF algorithm) have been employed. The rest of this section briefly describes the learning algorithms:

- NB [47] is a statistical learning algorithm based on Bayes' rule and conditional independence assumption. The conditional independence assumption regards as if the attributes are conditionally independent given the class. The assumption simplifies the required computations involved in the algorithm. In response, the algorithm is efficient and can scale well [50].
- SVMs [79] are linear algorithms that may be employed on classification and regression tasks. The algorithm can build suitable learning models in case of large amount of data. SVMs can be successfully employed in wide range of classification tasks, including text mining [43].
- LR [37] is a linear learning algorithm. The algorithm provides a scheme to apply linear regression to classification tasks. It employs a linear regression model and transformed target variables have been utilized to construct a linear classification scheme.
- KNN is an instance-based learning algorithm for supervised learning tasks, including classification and regression tasks [4]. In this scheme, the class label for an instance has been determined based on the similarity of the instance to its nearest neighbors in the training set.
- RF [15] is a supervised learning algorithm, which combines bagging algorithm and random subspace method. In this scheme, decision trees have been utilized as the base learning algorithm. Each tree has been built based on bootstrap samples of the training data. To provide the diversity among the base learners, a random feature selection has been employed. In response, the model can yield promising learning models on datasets with noisy or irrelevant data.

3.4 | Ensemble learning methods

Ensemble learning refers to the process of combining the predictions of multiple supervised learning algorithms and treating the algorithms as a committee of decision makers. Ensemble learning schemes seek to identify a more robust classification scheme with higher predictive performance. The earlier work on the sentiment analysis indicated that ensemble learning can yield promising results [58]. In the empirical analysis, ensembles of five supervised learning algorithms with three well-known ensemble learning methods (namely, AdaBoost, Bagging, and Random Subspace) have been considered. The rest of

this section briefly describes the ensemble learning methods:

- AdaBoost is a boosting based ensemble learning algorithm [30]. In this scheme, the base learning algorithms have been trained sequentially and a new learning model has been constructed at each round. In response, the base learning algorithm seeks to dedicate more rounds on instances that are harder to learn and to compensate classification errors made in earlier models.
- Bagging (Bootstrap aggregating) [14] is another ensemble learning method. In this scheme, different training subsets have been obtained from the original training set by bootstrap sampling. The predictions made by the base learning algorithms have been combined by the majority voting scheme.
- Random Subspace [38] is another ensemble learning method. In this scheme, the diversity among the members of the ensemble have been achieved in terms of feature space based partition.

3.5 | Deep learning architectures

Deep learning is a research direction of machine learning that seeks to identify a classification scheme with higher predictive performance based on multiple layers/or stages of nonlinear information processing in a hierarchical way [46]. In the hierarchies of levels, higher levels have more distributed and compact representations regarding the data. In this way, complex relationships among the lower-levels can be revealed. In machine learning, feature extraction has been followed by the learning model construction by a classification algorithm. In contrast, a deep learning architecture learns features from large datasets without any feature extraction process. Deep learning can be employed in a wide range of application fields, including sentiment analysis. The rest of this section briefly describes the deep learning architectures utilized in the empirical analysis:

3.5.1 | Convolutional neural networks

CNN are a type of deep neural networks that process data by grid-based topology [22]. In CNN, convolution, a specialized kind of mathematical operation, has been employed in one or more convolutional layers, instead of general matrix multiplication employed in conventional neural network architectures. The CNN architecture comprises input layer, output layer, and hidden layers. The hidden layers of the architecture substitute several layers, referred as, convolutional layers, pooling layers,

fully connected layers, and normalization layers. In convolutional layers, convolution operation has been employed on input data. As a result, the feature maps have been extracted. The activation functions (such as, rectified linear unit) has been employed in conjunction with the feature maps to add the nonlinearity to the architecture. After convolution, pooling layers combine the outputs of neuron clusters so that the spatial size of the feature space has been progressively reduced. In response, the number of parameters has been reduced and the model's ability to control overfitting has been improved. In pooling layer, maximum pooling scheme is a typical function. In this scheme, the maximum value from each cluster has been taken. After convolutional and pooling layers, the fully connected layers obtain the final output of architecture [28].

3.5.2 | Recurrent neural network

RNN [48] is a type of deep neural networks for processing sequential data. In RNN, connections between neurons form a directed graph. RNN can use its internal state to process sequences of inputs, which makes it an appropriate technique on sequential tasks, such as, speech recognition [81]. In RNN, each output has been determined by recurrently processing the same task over each instance of the sequence. In this way, the output has been determined based on all the earlier computations. In RNN architecture, the length of the time steps has been determined based on the length of input. Let x_t denote the input to the architecture at time step t and let s_t denote the hidden state at time step t . The current hidden state (s_t) has been computed as given by (3), by taking the current input and the hidden state for the former time stamp [39]:

$$s_t = f(Ux_t + Ws_{t-1}) \quad (3)$$

In (3), f denotes the activation function, which is usually taken as, \tanh function or $ReLU$ function. U and W corresponds to the weights that are shared across the time.

RNN suffers from the vanishing gradient problem. Hence, RNN architectures cannot properly handle with arbitrarily long sequences of input. To handle properly with this problem, some other RNN-based architectures, such as long short-term memory (LSTM), gated recurrent units (GRUs) and bidirectional RNN have been introduced.[39]

3.5.3 | Long Short-Term memory networks

LSTM is a type of RNN which avoids exploding or vanishing gradient problem with the use of forget gates.

In contrast to conventional RNN architectures, LSTM allows the backpropagation of error through the limited number of time steps. Typical LSTM unit comprise a cell and three kinds of gates, i.e., an input gate, an output gate and a forget gate. Based on the open and close operations on gates, the cell determines which information should be preserved and when the information should be accessed by the units. The LSTM transition has been carried on based on the equations given below [67]:

$$i_t = \sigma(W^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)}) \quad (4)$$

$$f_t = \sigma(W^{(f)}x_t + U^{(f)}h_{t-1} + b^{(f)}) \quad (5)$$

$$o_t = \sigma(W^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)}) \quad (6)$$

$$u_t = \tanh(W^{(u)}x_t + U^{(u)}h_{t-1} + b^{(u)}) \quad (7)$$

$$c_t = i_t \circ u_t + f_t \circ c_{t-1} \quad (8)$$

$$h_t = o_t \circ \tanh(c_t) \quad (9)$$

where x_t denotes the input vector to the LSTM unit, f_t denotes the activation vector for the forget gate, i_t denotes the activation vector for the input gate, o_t corresponds to the activation vector for output gate, h_t denotes the hidden state vector and c_t denotes the cell state vector. In this model, W corresponds to the weight matrices and b corresponds to bias vector parameters.

3.5.4 | Gated recurrent units

GRU [21] is a type of RNN which exhibits similar empirical results to LSTM with a less complicated architecture. Typical GRU architecture consists of two gates, i.e., reset gate and update gate. The architecture has fewer parameter processing. In GRU architecture, the transition has been carried on based on the equations given below [21]:

$$z_t = \sigma(U_z x_t + W_z h_{t-1}) \quad (10)$$

$$r_t = \sigma(U_r x_t + W_r h_{t-1}) \quad (11)$$

$$s_t = \tanh(U_s x_t + W_s \cdot (h_{t-1} \circ r_t)) \quad (12)$$

$$h_t = (1 - z_t) \circ s_t + z_t \circ h_{t-1} \quad (13)$$

where x_t denotes the input vector, h_t denotes the output vector, r_t corresponds to the reset gate vector, z_t corresponds to the update gate vector and W , U , and b corresponds to parameter matrices and vector.

3.5.5 | RNN with attention mechanism

The conventional encoder-decoder framework encountered in RNN-based architectures, such as bidirectional RNN or LSTM, must encode all information which may be irrelevant to the current task. In case of long input sequence, it is not possible to fully capture information rich and selective encoding. To handle properly with this problem, a technique, referred as, attention mechanism has been introduced. In the bidirectional RNN 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, decoder pay varying attentions to the states [72].

3.6 | Corpus

To collect a text corpus on SET, we crawled Rate-myprofessors.com [65], a popular instructor review website. In this way, 286,000 instructor evaluation reviews have been collected. In this website, students can evaluate instructors with the use of a 5-point scale, from which an overall quality score has been computed. To obtain a labeled corpus, we have utilized the overall quality scores given by the users. The evaluation reviews with the overall quality score of 1 or 2 have been labeled as “negative,” whereas the reviews with the overall quality of 5 have been labeled as “positive”. In addition to the overall quality scores provided by the website, the raw review messages have been also annotated by an expert. To examine the inter-annotator agreement, Cohen’s κ metric has been computed. For the corpus, this has resulted $\kappa = .82$. Since $\kappa = 1$ indicates a perfect agreement between different annotators, the agreement levels of labels generated by overall scoring and human annotations are good. Range of 0.61–0.80 for Cohen’s κ corresponds to substantial agreement among the annotators and 0.81–1.00 range for Cohen’s κ corresponds to perfect agreement [55,72]. Hence, the labels provided by the overall quality scores provide suitable labels for text sentiment annotation. In this way, we obtained a corpus with 89,000 negative reviews and 77,000 positive reviews. To obtain a balanced corpus, our final corpus comprises a total of 154,000 reviews with 77,000 positive and 77,000 negative reviews.

In addition to the overall quality scores, Ratemyprofessors.com has also course difficulty score. To analyze the correlation between the overall quality score and course difficulty score, we have performed correlation analysis in Minitab statistical software. To perform analysis, the overall quality scores and course difficulty scores for 154,000 reviews have been taken into consideration.

Pearson’s correlation measure was employed to determine whether there is a statistically significant relationship between overall quality scores and course difficulty scores. Pearson’s correlation coefficient is -0.973 , which indicates a strong negative correlation between the two variables, overall quality score and course difficulty score. Pearson’s correlation coefficient is statistically significant as ($p = 0.000 < .0005$).

In Table 1, sample student evaluations of instructors from the corpus have been presented.

To build learning models on a text corpus, several preprocessing tasks should be carried out. For the preprocessing the corpus, we have adopted the framework utilized in Bustillos et al [18,71]. Namely, we performed sequence and punctuation marks elimination, URL removal, tokenization (separation of the sentences and words of the document into tokens or characters), stemming, and removal of stop words and irrelevant words.

3.7 | Evaluation measures

To evaluate the predictive performance of machine learning methods and deep learning algorithms, classification accuracy (ACC) and F-measure have been utilized as the evaluation measures.

Classification accuracy is one of the most widely employed measures for evaluation supervised learning algorithms, which is computed as given by (14):

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \quad (14)$$

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 (PRE) and recall (REC). PRE is the proportion of the true positives against the true positives and false positives as given by (15). REC is the proportion of the true positives against the true positives and false negatives as given by (16). Based on the (15) and (16), F -measure has been computed as given by (17):

$$PRE = \frac{TP}{TP + FP} \quad (15)$$

$$REC = \frac{TP}{TP + FN} \quad (16)$$

$$F - measure = \frac{2*PRE*REC}{PRE + REC} \quad (17)$$

TABLE 1 Sample student evaluations of instructors

Sentiment orientation	Evaluation review
Positive	Professor Ross is one of those individuals that you meet in college that truly care about their success of each and every one of his students. This class alone taught me more about preparing for the real world than any of my classes these past two years of college. Definitely a must-take for anyone no matter their major!
Positive	The class is fantastic, and the professor is knowledgeable and willing to help students. His lecture is valuable and precise. I've gained a wide range of information in class through his experience and real-life examples. I appreciate all of your hard work and thank you for teaching me valuable lessons!
Positive	Steve really cares about his student and reach out to each of them personally. It was easy for me to talk to Steve any question I had in future career choice and in life. He is also an marvelous lecturer that he is able to keep his 7-hr lecture entertaining. I learned a lot from his experience and this class.
Negative	She goes way too fast and does not really care about her students learning abilities. All she wants to do is give you the information. She needs to learn how to teach a difficult subject to students. I do not recommend her at all. UCLA Extension needs to bring better Chemistry professor who can teach and importantly care about the students.
Negative	His lectures relate very little to the textbook and homework problems. He only teaches things that he finds interesting. He started off the General Chemistry course talking about quantum mechanics rather than actual general chemistry. He is lazy and makes his TAs do all the work for him.
Negative	Worst "teacher" I've ever had, period. He cannot teach to save his life. Studied my a** off and it did nothing for me. His test questions were completely from out of left field, and his lectures make no sense whatsoever. The little that I did learn was all thanks to my TA and myself. Take a different professor and save yourself the stress.

Abbreviations: TA, teaching assistant; UCLA, University of California, Los Angeles.

4 | EXPERIMENTS AND RESULTS

This section presents the empirical results obtained from the conventional classification algorithms, ensemble learning methods, and deep learning architectures on text corpus described in Section 3.6.

4.1 | Results on machine learning-based sentiment analysis

For evaluation tasks, we have conducted two set of experiments, i.e., machine learning-based sentiment analysis and deep learning-based sentiment analysis. For the machine learning-based sentiment analysis, text corpus described in Section 3.6 has been represented by taking three conventional text representation schemes (namely, TP, TF, and TF-IDF schemes) and three different N-gram models (namely, unigram, bigram, and trigram model). In this way, nine different configurations have been obtained on the text corpus based on three different weighting schemes and three different N-gram models. The predictive performance of supervised machine learning methods (such as, NB, SVMs, logistic regression, KNN, and RF) and three ensemble learning methods (namely, AdaBoost, Bagging, and Random Subspace) have been evaluated on conventional text representation schemes. In the validation of machine learning-based algorithms, 10-fold cross-validation scheme has been employed. In this scheme, the

original data set has been randomly split into 10 equal-sized folds. In each iteration, one of the folds has been utilized as validation fold, whereas the other partitions have been utilized as training folds. The process has been repeated ten times and the average results have been presented. The supervised learning algorithms and ensemble learning methods have been implemented with WEKA 3.9. For the supervised learning and ensemble methods, the default parameters of WEKA have been employed. The general structure of machine learning-based sentiment analysis has been illustrated in Figure 1.

In this section, classification accuracy and F -measure values obtained by conventional supervised learning methods and ensemble learning algorithms as outlined in Figure 1 have been presented. In Tables 2 and 3, the classification accuracy values and F -measure values on machine learning-based schemes for the text corpus have been presented, respectively.

Table 2 presents the classification accuracy values obtained by supervised learning algorithms and ensemble learning methods on nine different configurations of text corpus based on conventional weighting schemes and N-gram models. In the empirical analysis, five supervised learning algorithms (namely, NB, SVMs, logistic regression, KNN, and RF) have been considered. Regarding the predictive performance of supervised learning methods on the text corpus, the highest predictive performances in terms of classification

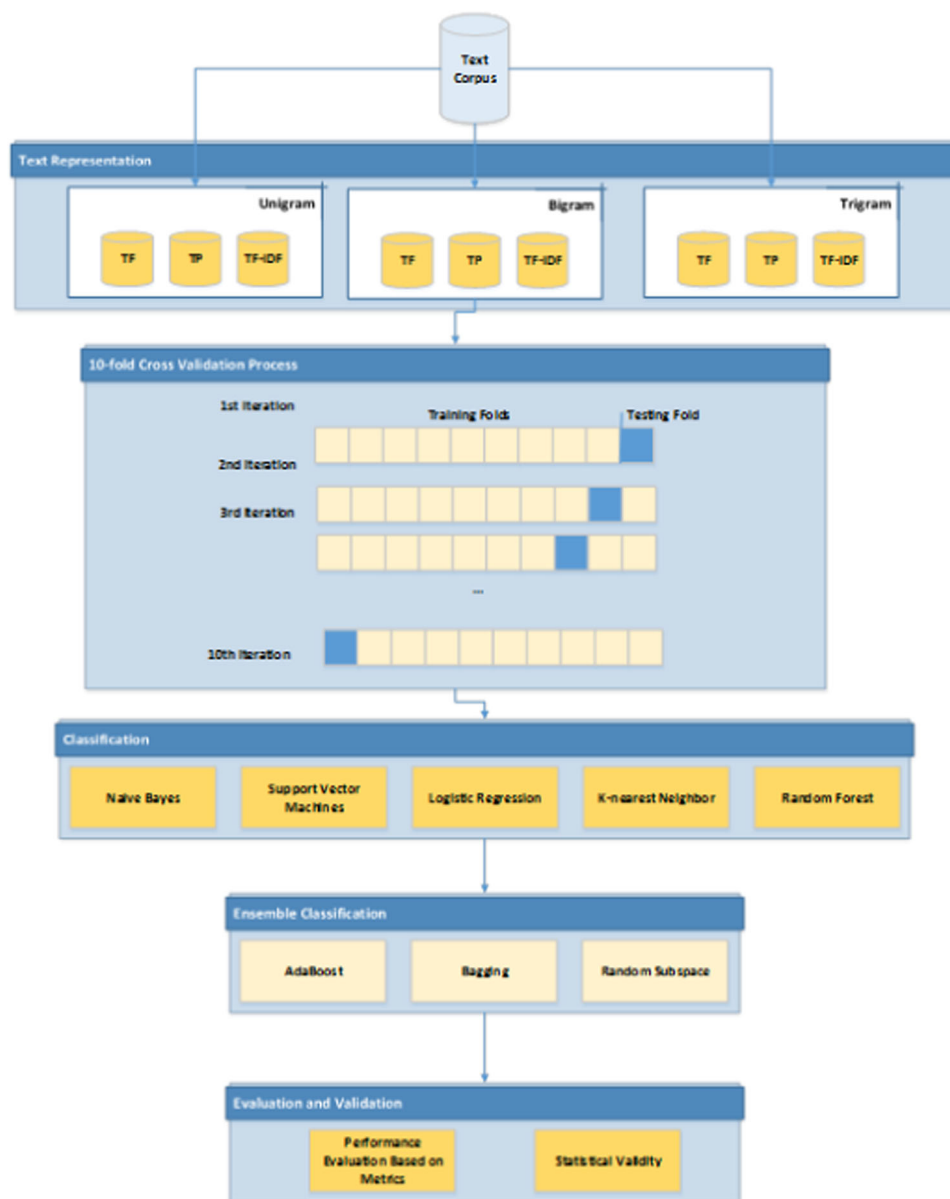


FIGURE 1 Machine learning-based sentiment analysis framework

accuracy have been achieved by RF algorithm. The second highest classification accuracies have been achieved by NB algorithm and the third highest classification accuracies have been obtained by SVMs.

Regarding the predictive performance of conventional text representation schemes, the highest classification accuracies have been obtained by unigram features with TF-based representation. The second highest predictive performances have been obtained by unigram features with term-presence and the third highest predictive performances have been achieved by unigram features with TF-IDF weighting. Empirical results presented in Tables 2 and 3 indicate that unigram model outperforms the other N-gram models, i.e., bigram and trigram models. In addition, TF-based representation yields

higher predictive performance compared with TP-based representation and TF-IDF based weighting.

As mentioned in advance, ensemble learning methods can be utilized in conjunction with supervised learning algorithms to enhance the predictive performance. In the empirical analysis, three different ensemble learning methods (AdaBoost, Bagging, and Random Subspace) have been examined. The empirical analysis results listed in Tables 2 and 3 indicate that the predictive performance of supervised learning methods can be enhanced with the use of ensemble learning methods. For ensemble methods, the highest predictive performance among the compared configurations has been obtained by random subspace ensemble of RF, with a classification accuracy of 84.25%. For this configuration, text corpus has been

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	70.37	71.35	70.12	69.52	69.94	68.85	65.62	67.05	64.22
SVM	73.33	73.50	73.13	72.50	72.98	71.84	71.43	71.61	71.39
LR	74.48	74.61	74.39	74.13	74.24	74.09	73.78	73.89	73.75
NB	75.57	75.81	75.47	75.23	75.27	75.18	74.98	75.07	74.81
RF	76.49	76.51	76.41	76.22	76.30	76.12	75.92	76.07	75.86
AdaBoost (KNN)	77.10	77.14	77.04	76.93	76.99	76.87	76.76	76.83	76.67
AdaBoost (SVM)	77.79	77.90	77.73	77.47	77.57	77.35	77.24	77.30	77.21
AdaBoost (LR)	78.35	78.37	78.31	78.12	78.22	78.10	78.01	78.02	77.95
AdaBoost (NB)	79.03	79.10	79.00	78.85	78.96	78.72	78.61	78.69	78.43
AdaBoost (RF)	79.62	79.71	79.55	79.46	79.50	79.35	79.22	79.29	79.17
Bagging (KNN)	80.13	80.17	80.06	79.97	80.03	79.93	79.81	79.85	79.74
Bagging (SVM)	80.58	80.67	80.52	80.38	80.45	80.34	80.29	80.30	80.23
Bagging (LR)	81.29	81.35	81.23	81.03	81.19	80.99	80.85	80.91	80.74
Bagging (NB)	81.66	81.71	81.64	81.59	81.64	81.56	81.48	81.52	81.40
Bagging (RF)	82.12	82.15	82.06	81.96	81.97	81.94	81.82	81.86	81.73
RS (KNN)	82.50	82.52	82.43	82.36	82.40	82.28	82.18	82.25	82.16
RS (SVM)	82.89	82.91	82.87	82.78	82.81	82.70	82.64	82.65	82.57
RS (LR)	83.31	83.33	83.23	83.18	83.21	83.13	83.05	83.08	83.02
RS (NB)	83.71	83.74	83.68	83.64	83.67	83.59	83.49	83.53	83.45
RS (RF)	84.24	84.25	84.23	84.13	84.20	84.08	83.90	83.97	83.83

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

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 + TP		Trigram + TF		Trigram + TF-IDF	
KNN	0.74	0.74	0.75	0.74	0.74	0.73	0.73	0.74	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73
SVM	0.76	0.76	0.76	0.76	0.76	0.75	0.75	0.76	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
LR	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76	0.76
NB	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77	0.77
RF	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
AdaBoost (KNN)	0.79	0.79	0.79	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78	0.78
AdaBoost (SVM)	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79	0.79
AdaBoost (LR)	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
AdaBoost (NB)	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
AdaBoost (RF)	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Bagging (KNN)	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Bagging (SVM)	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81
Bagging (LR)	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
Bagging (NB)	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
Bagging (RF)	0.82	0.82	0.83	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
RS (KNN)	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
RS (SVM)	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83
RS (LR)	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
RS (NB)	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
RS (RF)	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.84

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

represented by unigram features with TF-based representation. To summarize the main findings of the empirical analysis on machine learning-based sentiment analysis, Figure 2 depicts the main effects plot for accuracy values.

4.2 | Results on deep Learning-based sentiment analysis

For the deep learning-based sentiment analysis, text corpus has been represented by four word embedding schemes (namely, word2vec, GloVe, fastText, and LDA2Vec) described in Section 3.2.

To process text, five deep learning architectures (namely, CNN, RNN, bidirectional RNN-AM, GRU, and long short-term memory) described in Section 3.5 have been considered. We have utilized TensorFlow and Keras to implement and train the deep learning-based architectures utilized in the empirical analysis. For each model, we employed hyper parameter searching algorithm to obtain optimal predictive performance from each deep learning model. To do so, hyper parameter optimization based on Bayesian optimization using Gaussian process has been utilized. For the corpus, 80% of data has been utilized as the training set, whereas the rest of data has been utilized as the testing set. For word2vec and fastText schemes, continuous skip-gram and CBOW schemes have been taken into consideration with varying vector sizes (vector size of 200 and 300) and dimensions of projection layers (dimension size of 100 and 200). For LDA2vec scheme, a set of parameters (including, the number of topics and the negative sampling exponent) have been considered. The experimental results listed in Section 4.4 are the results for the

number of topics ($N=25$) and the negative sampling exponent ($\beta \in .75$). The general structure of deep learning-based sentiment analysis has been summarized in Figure 3.

In Tables 4 and 5, classification accuracies and F -measure values obtained by five deep learning architectures (CNN, RNN, long short-term memory, GRU, and RNN-AM) have been presented, respectively.

In the empirical results listed in Table 4, six word embedding models (word2vec skip-gram model, word2vec CBOW model, fastText skip-gram model, fastText CBOW model, Glove, and LDA2vec) have been examined. As it can be observed from the results listed in Table 4, GloVe word embedding scheme outperforms the other word embedding schemes for the text corpus. The second highest predictive performance has been achieved by LDA2vec word embedding scheme, 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 word embedding schemes, different vector sizes and dimensions of projection layers have been also considered. The experimental results indicate that word embedding schemes yield better predictive performance for the vector size of 300 and dimension projection layer of 300.

Regarding the predictive performances of deep learning architectures utilized in the empirical analysis, the highest predictive performances have been achieved by RNN-AM. The second highest predictive performances have been achieved by GRU. The third highest predictive performances have been achieved by LSTM networks. The empirical analysis indicates that RNN-AM, gated recurrent units and long short-term memory networks outperform conventional recurrent neural networks.

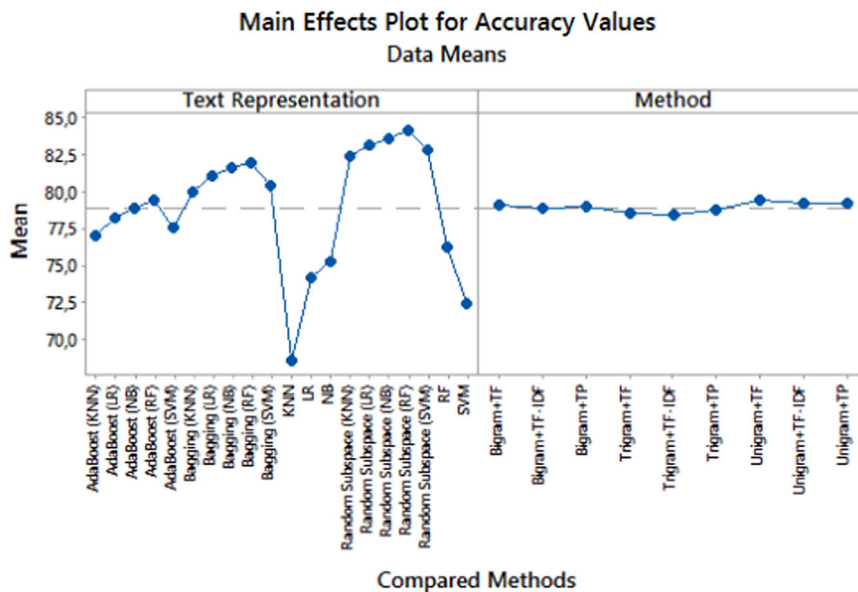


FIGURE 2 Main effects plot for classification accuracy values

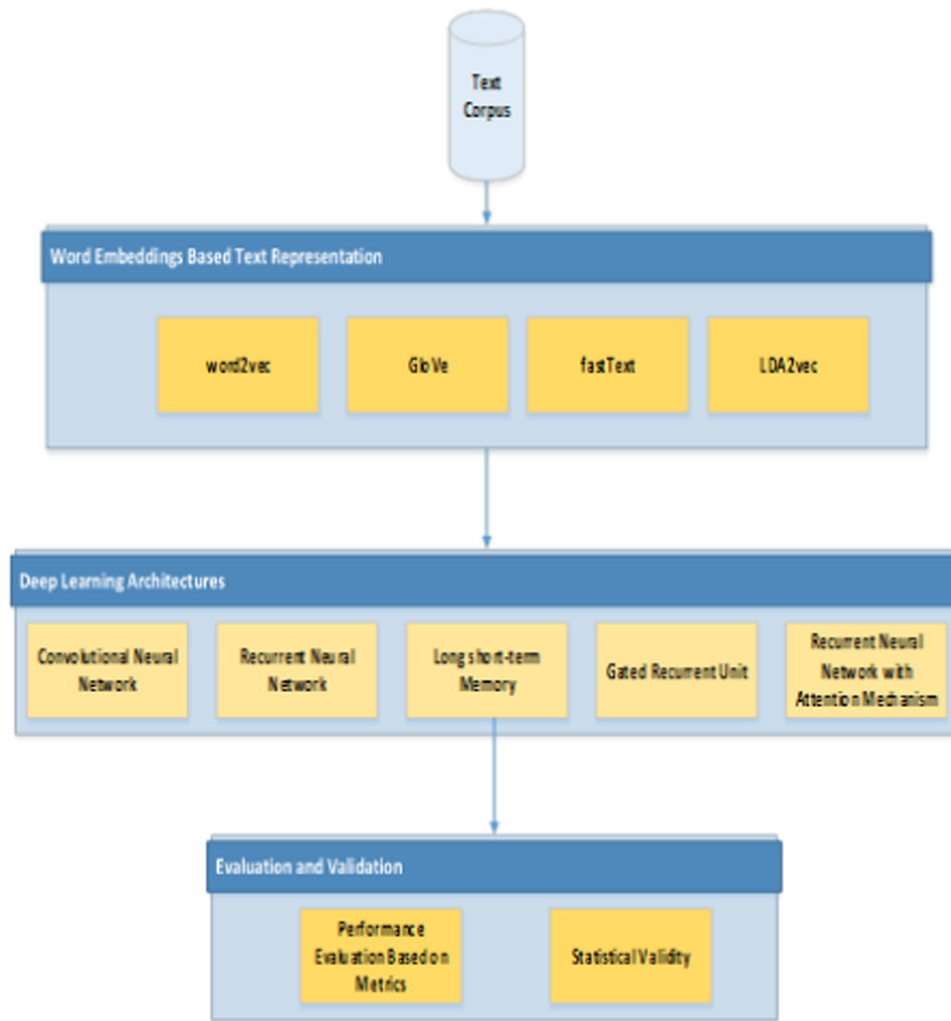


FIGURE 3 Deep learning-based sentiment analysis framework

For the empirical analysis, the lowest predictive performance in terms of classification accuracy has been achieved by the CNN architecture. Among the compared configurations, the highest predictive performance has been achieved by RNN-AM in conjunction with GloVe word embedding scheme-based representation, with a classification accuracy of 98.29%.

Regarding the predictive performances listed in Table 5, the same patterns mentioned for the schemes in terms of classification accuracies are still valid. The highest *F*-measure values have been obtained by RNN-AM, which is followed by gated recurrent units. Regarding the predictive performance of word embedding schemes in terms of *F*-measure values, GloVe word embedding scheme outperforms the other word embedding schemes. In Figure 4, main effect plot for classification accuracy values has been presented.

In the empirical results listed in Section 4.1, the results for machine learning-based sentiment has been presented. As it can be observed from the results listed in

Section 4.2, deep learning-based architectures outperform conventional supervised learning algorithms and ensemble learning methods for sentiment classification.

For comparison, Table 6 presents classification accuracies obtained by earlier related works for sentiment analysis on SET. For the empirical analysis presented in Section 4.1, unigram features with term-frequency-based representation outperform other N-gram models, i.e., bigram and trigram models. For results presented in Sultana et al [73] bigram features yield higher predictive performance compared to unigram features. As it can be observed from the results, deep learning-based architectures yield higher predictive performance compared to conventional machine learning classifiers for sentiment analysis. For results presented in Cabada et al [19] CNN architecture outperforms long short-term memory architecture. For the empirical analysis presented in Section 4.2, recurrent neural network-based architectures outperform CNN.

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	84.32	85.73	87.15	88.84	91.33
word2vec (Skip-gram)	200	200	84.34	85.83	87.19	88.99	91.46
word2vec (Skip-gram)	300	100	84.43	85.91	87.26	89.04	91.58
word2vec (Skip-gram)	300	300	84.50	85.93	87.41	89.23	91.65
word2vec (CBOW)	200	100	84.52	85.98	87.46	89.31	91.74
word2vec (CBOW)	200	200	84.56	86.01	87.50	89.35	91.87
word2vec (CBOW)	300	100	84.65	86.06	87.54	89.50	92.08
word2vec (CBOW)	300	300	84.73	86.08	87.60	89.60	92.25
fastText (Skip-gram)	200	100	84.80	86.16	87.65	89.69	92.30
fastText (Skip-gram)	200	200	84.87	86.19	87.68	89.77	92.37
fastText (Skip-gram)	300	100	84.92	86.22	87.74	89.80	92.53
fastText (Skip-gram)	300	300	84.99	86.32	87.78	89.90	92.68
fastText (CBOW)	200	100	85.05	86.38	87.95	89.97	92.81
fastText (CBOW)	200	200	85.17	86.41	88.01	90.09	93.13
fastText (CBOW)	300	100	85.23	86.52	88.08	90.24	93.21
fastText (CBOW)	300	300	85.26	86.59	88.10	90.40	93.62
GloVe	200	100	85.46	86.92	88.50	91.02	94.87
GloVe	200	200	85.49	86.98	88.54	91.15	94.95
GloVe	300	100	85.54	87.01	88.59	91.22	95.59
GloVe	300	300	85.58	87.03	88.76	91.27	98.29
LDA2Vec	200	100	85.29	86.67	88.14	90.57	93.93
LDA2Vec	200	200	85.33	86.70	88.23	90.64	94.11
LDA2Vec	300	100	85.40	86.77	88.30	90.75	94.41
LDA2Vec	300	300	85.43	86.88	88.36	90.95	94.64

Abbreviations: CBOW, continuous bag of words; CNN, convolutional neural network; GRU, gated recurrent unit; LSTM, long short-term memory; RNN-AM, recurrent neural network with attention mechanism.

4.3 | Statistical validity of empirical results

To evaluate the statistical significance of the results presented in Section 4.1 and Section 4.2, we have performed one-way analysis of variance (ANOVA) test in Minitab statistical program. The results for the one-way ANOVA test of overall results obtained by the conventional supervised learning algorithms, ensemble learning methods, and deep learning algorithms have been presented in Table 7, where DF, SS, MS, F , and p denote degrees of freedom, adjusted sum of squares, adjusted mean square, F value, and probability value, respectively. DF is the quantity of information in the data. The adjusted SS term reflects the quantity of variation that is explained by each model term. The adjusted SS error represents the variation in the data that the predictors do not reveal. The adjusted SS total denotes the total variation in the data. F -statistics (F) is the test statistic to identify whether a term is associated with the

response. In addition, the probability value (p) is utilized to make a decision about the statistical significance of the terms and model [59].

For the one-way ANOVA test results presented in Table 7, it can be observed that the F -statistics value of 594.24 for classification accuracy values on machine learning approaches and F -statistics value of 636.42 for F -measure values on machine learning approaches indicates that there is a statistically significant difference ($p < .001$) for the means of at least two machine learning approaches. According to the one-way ANOVA test results presented in Table 7, there are statistically meaningful differences between the predictive performances obtained by conventional supervised learning algorithm, ensemble learning methods and deep learning algorithms.

The 95% confidence interval for the compared algorithms based on the pooled standard deviation has been presented in Figure 5. The confidence intervals for the mean values of classification accuracies obtained by

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	0.89	0.92
word2vec (Skip-gram)	200	200	0.85	0.87	0.88	0.89	0.92
word2vec (Skip-gram)	300	100	0.85	0.87	0.88	0.89	0.92
word2vec (Skip-gram)	300	300	0.85	0.87	0.88	0.89	0.92
word2vec (CBOW)	200	100	0.85	0.87	0.88	0.90	0.92
word2vec (CBOW)	200	200	0.85	0.87	0.88	0.90	0.92
word2vec (CBOW)	300	100	0.85	0.87	0.88	0.90	0.92
word2vec (CBOW)	300	300	0.85	0.87	0.88	0.90	0.92
fastText (Skip-gram)	200	100	0.85	0.87	0.88	0.90	0.93
fastText (Skip-gram)	200	200	0.85	0.87	0.88	0.90	0.93
fastText (Skip-gram)	300	100	0.85	0.87	0.88	0.90	0.93
fastText (Skip-gram)	300	300	0.85	0.87	0.88	0.90	0.93
fastText (CBOW)	200	100	0.86	0.87	0.88	0.90	0.93
fastText (CBOW)	200	200	0.86	0.87	0.88	0.90	0.93
fastText (CBOW)	300	100	0.86	0.87	0.88	0.91	0.94
fastText (CBOW)	300	300	0.86	0.87	0.89	0.91	0.94
GloVe	200	100	0.86	0.87	0.89	0.91	0.95
GloVe	200	200	0.86	0.87	0.89	0.91	0.96
GloVe	300	100	0.86	0.88	0.89	0.91	0.96
GloVe	300	300	0.86	0.88	0.89	0.92	0.98
LDA2Vec	200	100	0.86	0.87	0.89	0.91	0.94
LDA2Vec	200	200	0.86	0.87	0.89	0.91	0.94
LDA2Vec	300	100	0.86	0.87	0.89	0.91	0.95
LDA2Vec	300	300	0.86	0.87	0.89	0.91	0.95

Abbreviations: CBOW, continuous bag of words; CNN, convolutional neural network; GRU, gated recurrent units; LSTM, long short-term memory; RNN-AM, recurrent neural network with attention mechanism.

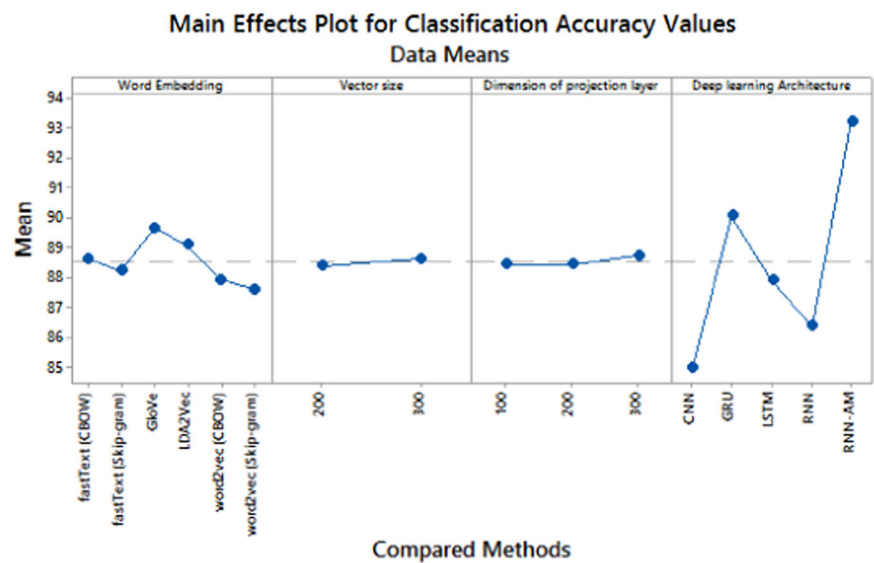
**FIGURE 4** Main effects plot for classification accuracy values (deep learning)

TABLE 6 Comparison of earlier research contributions

Reference	Methods/architectures	Classification accuracy
Sultana et al [73]	Multilayer perceptron	78.33
Sultana et al [73]	Support vector machines	78.75
Nguyen et al [54]	Unigram features + Naive Bayes	85.30
Nguyen et al [54]	Bigram features + Naive Bayes	87.50
Nguyen et al [54]	word2vec + LSTM	87.60
Nguyen et al [54]	word2vec + Bi-LSTM	92.00
Kandhro et al [45]	word2vec + LSTM	89.00
Bustillos et al [18]	Bernoulli Naive Bayes	76.77
Bustillos et al [18]	CNN + LSTM	88.26
Cabada et al [19]	Multilayer perceptron	90.42
Cabada et al [19]	CNN	92.46
Cabada et al [19]	LSTM	90.92
Cabada et al [19]	CNN + LSTM	92.15
Our proposal	Glove + RNN-AM	98.29

Abbreviations: CNN, convolutional neural network; LSTM, long short-term memory; RNN-AM, recurrent neural network with attention mechanism.

the compared algorithms support the statistical analysis results presented in Table 7. As it can be observed from Figure 5, it has been divided into three different regions based on the statistical significances between the results of algorithms. Hence, the predictive performance differences obtained by supervised learning methods, deep learning methods and ensemble learning methods are statistically significant.

4.4 | Discussions

Based on the empirical analysis on student evaluations of teaching, several insights follow:

TABLE 7 One-way ANOVA test results

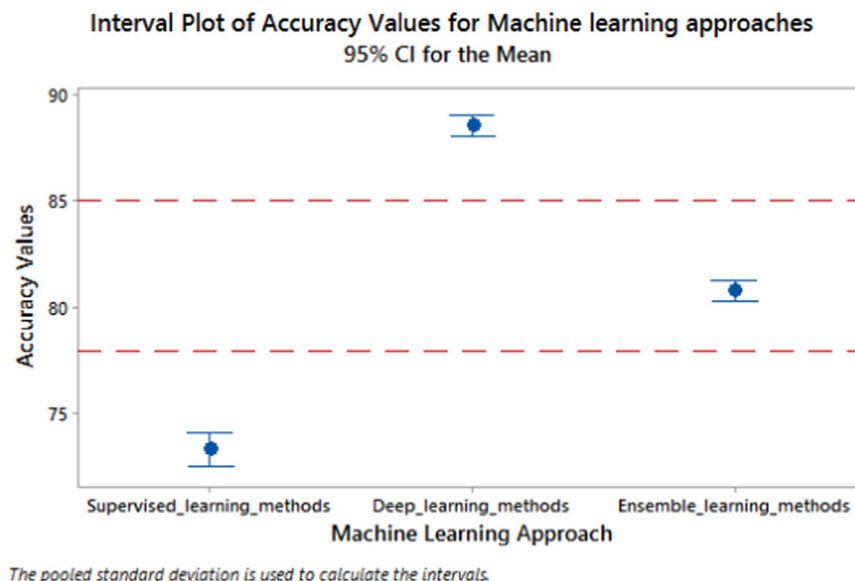
Classification accuracy values					
Source	DF	SS	MS	F value	p Value
Machine learning approach	2	8505	4252.67	594.24	0.000
Error	297	2125	7.16		
Total	299	10631			

F-measure values					
Source	DF	SS	MS	F value	p Value
Machine learning approach	2	0,6562	0,328119	636,42	0.000
Error	297	0,1531	0,000516		
Total	299	0,8094			

Abbreviations: ANOVA, analysis of variance; DF, degrees of freedom; MS, mean square; SS, sum of square.

- In the empirical analysis, conventional machine learning classifiers, ensemble learning methods, and deep learning architectures have been taken into consideration. Ensemble learning methods generally yield higher predictive performance compared with the conventional classification algorithms and deep learning architectures outperform ensemble learning-based methods. For the RNN-AM in conjunction with GloVe word embedding scheme-based representation a classification accuracy of 98.29% has been obtained.
- The experimental results indicate that deep learning-based frameworks can yield promising results on model machine learning/data mining-based tasks in the field of education.
- Text mining and machine learning techniques can be employed on SET to help administrators of higher educational institutions. In addition, it provides feedback to instructors regarding the problematic parts of teaching and learning process. Many higher educational institutions employ SET to analyze the performance of instructors [27]. This feedback can be utilized for evaluating the learning experience and other attributes of the learning process.
- The presented text mining-based sentiment analysis framework has been employed on SET crawled from a popular instructor review website. Sentiment analysis has many application fields in the context of education. Evaluating the progress of group discussion [77], recognizing and regulating e-learners' emotions [32] and identifying learning-related emotions of students on text feedbacks [11] are to name a few. The presented scheme may also be employed for course evaluation surveys, discussion forums, and blog posts. Deep

FIGURE 5 Interval plot of accuracy values for the machine learning approaches. CI, confidence interval



learning-based sentiment analysis framework can be integrated to online learning portals for real-time sentiment analysis of student feedback.

Despite its promising predictive performance, the proposed scheme has several limitations:

- The corpus utilized in the empirical analysis has been crawled from Ratemyprofessors.com. The dominant participants on the website are localized to instructors and schools in USA. In the field of machine learning and data science, the results may be sensitive to the corpus. This can be regarded as one limitation of the study. The presented scheme, however, follows a machine learning-based approach to text sentiment classification. Hence, the framework employed can be easily employed for sentiment analysis on other languages with appropriate preprocessing.
- The objective of this study is to present sentiment analysis-based evaluation model for SET. Many higher educational institutes utilize student information systems to organize key tasks of learning process, such as registration, grade entries, and course assignments for instructors. These systems tend to have modules dedicated to course evaluation surveys. Student information systems have the immense quantity of student opinions regarding courses, schools and instructors. For institutions with survey modules, machine learning and deep learning frameworks can be easily integrated. For resource crunched universities in developing nations, student information systems may not be utilized to handle key activities of learning process. This can be regarded as one limitation of the study. For such institutions, open source platforms may be utilized to handle with course registration and course evaluation, as well.

5 | CONCLUSIONS

In this paper, we have presented a text mining approach for opinion mining on instructor evaluation reviews. We have collected a text corpus containing 154,000 reviews. We have presented comprehensive empirical analysis on the text corpus, with the use of conventional supervised learning algorithms (NB, SVMs, logistic regression, K-NN, and RF), ensemble learning methods (AdaBoost, Bagging, and random subspace) and deep learning architectures (CNN, RNN, bidirectional RNN-AM, GRU, and long short-term memory). Three conventional text representation schemes (TP, TF, and TF-IDF weighting) have been employed in conjunction with conventional machine learning algorithms. The four word embedding schemes (word2vec, GloVe, fastText, and LDA2vec) have been utilized in conjunction with the deep learning architectures. The empirical analysis indicates that deep learning-based schemes can yield more promising results compared with the ensemble learning methods and supervised learning methods for sentiment classification. Among the compared configurations, the highest predictive performance has been achieved by RNN with attention mechanism in conjunction with GloVe word embedding scheme-based representation, with a classification accuracy of 98.29%.

ORCID

Aytuğ Onan  <http://orcid.org/0000-0002-9434-5880>

REFERENCES

1. A. I. Adekitan, and E. Noma-Osaghae, *Data mining approach to predicting the performance of first year student in a university*

- using the admission requirements, *Educ. Inf. Technol.* **24** (2019), no. 2, 1527–1543.
2. A. I. Adekitan and O. Salau, *The impact of engineering students' performance in the first three years on their graduation result using educational data mining*, *Heliyon* **5** (2019), no. 2.
3. P. Adinolfi et al., *Sentiment analysis to evaluate teaching performance*, *Int. J. Knowl. Soc. Res.* **7** (2016), no. 4, 86–107.
4. D. W. Aha, D. Kibler, and M. K. Albert, *Instance-based learning algorithms*, *Mach. Learn.* **6** (1991), 37–66.
5. N. Alami, M. Meknassi, and N. Ennahnani, *Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning*, *Expert Syst. Applic.* **123** (2019), 195–211.
6. A. Almasri, E. Celebi, and R. S. Alkhawaldeh, *EMT: Ensemble meta-based tree model for predicting student performance*, *Scientific Programming*, 2019.
7. N. Altrabsheh, M. Cocea, and S. Fallahkhair, *Predicting learning-related emotions from students' textual classroom feedback via Twitter*, *Proceedings of the 8th International Conference on Educational Data Mining*, 2015.
8. E. Andersson, C. Dryden, and C. Variawa, *Methods of applying machine learning to student feedback through clustering and sentiment analysis*, *Proceedings of the Canadian Engineering Education Association (Vancouver, BC, 2018)*, 2018.
9. W. E. Becker, and M. Watts, *How departments of economics should evaluate teaching*, *Am. Econ. Rev.* **89** (1999), 344–349.
10. J. Beemer et al., *Ensemble learning for estimating individualized treatment effects in student success studies*, *Int. J. Artif. Intell. Educ.* **28** (2018), no. 3, 315–335.
11. A. Bermingham and A. F. Smeaton, *Classifying sentiment in microblogs: is brevity an advantage*, *Proceedings of the ACM International Conference on Information and Knowledge Management*, 2010.
12. S. S. Boswell, *Ratemyprofessors is hogwash (but I care): effects of ratemyprofessors and university-administered teaching evaluations on professors*, *Comput. Hum. Behav.* **56** (2016), 155–162.
13. M. Braga, M. Paccagnella, and M. Pellizzari, *Evaluation students' evaluations of professors*, *Econ. Educ. Rev.* **41** (2014), 71–88.
14. L. Breiman, *Bagging predictors*, *Mach. Learn.* **4** (1996), no. 2, 123–140.
15. L. Breiman, *Random forests*, *Mach. Learn.* **45** (2001), no. 1, 5–32.
16. B. Brockx, R. K. Van, and D. Mortelmans, *The student as a commentator: Students' comments in student evaluations of teaching*, *Procedia Soc. Behav. Sci.* **69** (2012), 1122–1133.
17. J. Brownlee, *Machine learning mastery*. <http://machinelearningmastery.com/discover-feature-engineering-howtoengineer-features-and-how-to-get-good-at-it>
18. O. R. Bustillos et al., *Opinion mining and emotion recognition in an intelligent learning environment*, *Comput. Applic. Eng. Educ.* **27** (2019), no. 1, 90–101.
19. 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.
20. P. Chaovalit and L. Zhou, *Movie review mining: A comparison between supervised and unsupervised classification approaches*, *Proceedings of the 38th Annual Hawaii Conference on System Sciences (Big Island, HI, 2005)*, IEEE, 2005.
21. K. Cho et al, *Learning phrase representations using rnn encoder-decoder for statistical machine translation*, *arXiv preprint* (2014).
22. D. Cireşan, U. Meier, and J. Schmidhuber, *Multi-column deep neural networks for image classification*, *arXiv preprint*. (2012).
23. H. Cui, V. Mittal, and M. Datar, *Comparative experiments on sentiment classification for online product reviews*, *AAAI Conference (Boston, MA, 2006)*, 2006, pp. 1265–1270.
24. L. Deng, and D. Yu, *Deep learning: Methods and applications*, *Found. Trends Signal Process.* **7** (2014), no. 3 to 4, 197–387.
25. S. Y. Deng, and X. Que, *Research on the teaching assessment of students of science and engineering teachers in a university*, *Comput. Applic. Eng.* **27** (2019), no. 1, 5–12.
26. W. Di, A. Bhardwaj, and J. Wei, *Deep learning essentials: Your hands-on guide to the fundamentals of deep learning and neural network modelling*, *Packt Publishing*, New York, NY, 2018.
27. L. P. Dringus, and T. Ellis, *Using data mining as a strategy for assessing asynchronous discussion forums*, *Comput. Educ.* **45** (2005), no. 1, 141–60.
28. J. L. Elman, *Finding structure in time*, *Cognit. Sci.* **14** (1990), no. 2, 179–211.
29. E. Fersini, E. Messina, and F. A. Pozzi, *Sentiment analysis: Bayesian ensemble learning*, *Decis. Support Syst.* **68** (2014), 26–38.
30. Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, *Proceedings of the Thirteenth International Conference on Machine Learning (Bari, Italy, 1996)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1996, pp. 148–156.
31. X. Glorot, A. Bordes, and Y. Bengio, *Domain adaptation for large-scale sentiment classification: A deep learning approach*, *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, 2011.
32. N. Godbole, M. Srinivasaiyah, and S. Skiena, *Large-scale sentiment analysis for news and blogs*, *Proceedings of the International Conference on Weblogs and Social Media*, 2007.
33. P. Gonçalves et al., *Comparing and combining sentiment analysis methods*, *Proceedings of the first ACM Conference on Online Social Networks*, 2013.
34. K. M. Gregory, *How undergraduates perceive their professors: A corpus analysis of rate my professor*, *J. Educ. Technol. Syst.* **40** (2011), no. 2, 169–193.
35. G. Gutierrez et al., *Mining: Students comments about teacher performance assessment using machine learning algorithms*, *Int. J. Comb. Optim. Probl. Inf.* **9** (2018), no. 3, 26–40.
36. G. Hackeling, *Mastering machine learning with scikit-learn*, *Packt Publishing*, New York, 2017, p. 238.
37. T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, *Springer*, New York, NY, 2009.
38. T. K. Ho, *The random subspace method for constructing decision forests*, *IEEE Transactions on Pattern Analysis and Machine Learning* **20** (1998), no. 8, 832–44.
39. S. Hochreiter, and J. Schmidhuber, *Long short-term memory*, *Neural Comput.* **9** (1997), no. 8, 1735–1780.
40. Z. Hu et al., *Review sentiment analysis based on deep learning*, *Proceedings of the 12th International Conference on e-Business Engineering (Beijing, China, 2015)*, IEEE, 2015.
41. W. Y. Hwang, Y. H. Li, and R. Shadiev, *Exploring effects of discussion on visual attention, learning performance and*

- perceptions of students learning with support, *Comput. Educ.* **116** (2018), 225–236.
42. R. K. Jena, *Sentiment mining in a collaborative learning environment: Capitalising on big data*, *Behav. Inf. Technol.* **38** (2019), no. 9, 986–1001.
 43. T. Joachims, *Text categorization with support vector machines*, *Proceedings of the European Conference on Machine Learning*, 1998.
 44. A. Joulin et al., *Fasttext.zip: Compressing text classification models*. arXiv preprint, 2016.
 45. I. A. Kandhro et al., *Sentiment analysis of students' comment using long-short term model*, *Indian J. Sci. Technol.* **12** (2019), no. 8, 1–16.
 46. Y. LeCun, *Generalization and network design strategies*, Elsevier, Amsterdam, 1989.
 47. D. Lewis, *Naïve Bayes at forty: The independence assumption in information retrieval*, *Proceedings of the 10th European Conference on Machine Learning*, 1998.
 48. X. Li and X. Wu, *Constructing long short-term memory based deep recurrent neural network for large vocabulary speech recognition*, arXiv preprint (2014).
 49. Q. Lin et al., *Lexical based automated teaching evaluation via students' short reviews*, *Comput. Applic. Eng. Educ.* **27** (2019), no. 1, 194–205.
 50. A. McCallum and K. Nigam, *A comparison of event models for Naïve Bayes text classification*, *Proceedings of AAAI-98 Workshop on Learning for text categorization*, 1998, pp. 41–48.
 51. W. Medhat, A. Hassan, and H. Korashy, *Sentiment analysis algorithms and applications: A survey*, *Ain Shams Eng. J.* **5** (2014), no. 4, 1093–1113.
 52. T. Mikolov et al., *Efficient estimation of word representations in vector space*. arXiv preprint, 2013.
 53. C. E. Moody, R. Johnson, and T. Zhang, *Mixing Dirichlet topic models and word embeddings to make lda2vec 2014*, May 11, 2019, available at <https://www.datacamp.com/community/tutorials/lda2vec-topic-model>
 54. P. X. Nguyen et al., *Deep learning versus traditional classifiers on Vietnamese students' feedback corpus*, *Proceedings of the 5th NAFOSTED Conference on Information and Computer Science*, 2018.
 55. A. Onan, *An ensemble scheme based on language function analysis and feature engineering for text genre classification*, *J. Inf. Sci.* **44** (2018), no. 1, 28–47.
 56. A. Onan, *Topic-enriched word embeddings for sarcasm identification*, *Proceedings of CSOC 2019*, 2019, pp. 293–304.
 57. A. Onan, H. Bulut, and S. Korukoglu, *An improved ant algorithm with LDA-based representation for text document clustering*, *J. Inf. Sci.* **43** (2017), no. 2, 275–292.
 58. A. Onan, S. Korukoğlu, and H. Bulut, *A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification*, *Expert Syst. Applic.* **62** (2016), 1–16.
 59. A. Onan, S. Korukoğlu, and H. Bulut, *A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification*, *Inf. Process. Manage.* **53** (2017), no. 4, 814–833.
 60. B. Pang, L. Lee, and S. Vaithyanathan, *Thumbs up sentiment classification using machine learning techniques*, *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, Assoc. Comput. Linguist., 2002, pp. 79–86.
 61. P. Pang, and L. Lee, *Opinion mining and sentiment analysis*, *Found. Trends Inf. Retr.* **2** (2008), 1–135.
 62. J. Pennington, R. Socher, and C. Manning, *Glove: Global vectors for word representation*, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.
 63. R. Prabowo, and M. Thelwall, *Sentiment analysis: A combined approach*, *J. Informetrics* **3** (2009), 143–157.
 64. S. Rani and P. Kumar, *A sentiment analysis system to improve teaching and learning*, *Computer* **50** (2017), no. 5, 36–43.
 65. Ratemyprofessors.com, *About ratemyprofessors.com 2019*, May 11, 2019, available at <http://www.ratemyprofessors.com/About.jsp>
 66. S. M. Rezaeinia et al., *Sentiment analysis based on improved pre-trained word embeddings*, *Expert Syst. Applic.* **117** (2019), 139–147.
 67. L. M. Rojas-Barahona, *Deep learning for sentiment analysis*, *Lang. Linguist. Compass* **10** (2016), no. 12, 701–719.
 68. 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. Manage.* **32** (2018), no. 3, 431–447.
 69. 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. Manage.* **32** (2018), no. 3, 431–447.
 70. C. Dos Santos and M. Gatti, *Deep convolutional neural networks for sentiment analysis of short texts*, *Proceedings of the 25th International Conference on Computational Linguistics (Dublin, Ireland, 2014)*, Dublin City Univ. Assoc. Comput. Linguist., 2014, pp. 69–78.
 71. A. Severyn and A. Moschitti, *Twitter sentiment analysis with convolutional neural networks*, *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2015.
 72. E. Stamatatos, *A survey of modern authorship attribution methods*, *J. Am. Soc. Inf. Sci. Technol.* **60** (2009), no. 3, 548–556.
 73. J. Sultana et al., *Prediction of sentiment analysis on educational data based on deep learning approach*, *Proceedings of the 21st Saudi Computer Society National Computer Conference (Riyadh, Saudi Arabia, 2018)*, 2018.
 74. D. Tang et al., *Coooll: A deep learning system for twitter sentiment classification*, *Proceedings of the 8th International Workshop on Semantic Evaluation*. 2014, pp. 208–212.
 75. T. T. Thet, J. C. Na, and C. S. Khoo, *Aspect-based sentiment analysis of movie reviews on discussion boards*, *J. Inf. Sci.* **36** (2010), no. 6, 823–848.
 76. 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.
 77. F. Tian et al., *Recognizing and regulating elearners' emotions based on interactive Chinese texts in e-learning systems*, *Knowl. Based Syst.* **55** (2014), 148–164.
 78. T. D. Ullmann, *Automated analysis of reflection in writing: Validating machine learning approaches*, *Int. J. Artif. Intell. Educ.* (2019), 1–41.
 79. V. Vapnik, *Statistical learning theory*, John Wiley & Sons, New York, NY, 1998.

80. B. C. Wallace et al., *A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews*, J. Am. Med. Informat. Assoc. **21** (2014), no. 6, 1098–1103.
81. 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.
82. X. Zhang and Q. Yu, *Hotel reviews sentiment analysis based on word vector clustering*. 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCI). 2017.

AUTHOR BIOGRAPHY



Aytuğ Onan was born in İzmir, Turkey, in 1987. He received the B.S. degree in computer engineering from the İzmir University of Economics, Turkey, in 2010, and the M.S. degree in computer engineering and the Ph.D. 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, İzmir Katip Celebi University, Turkey, since April 2019. He has published several journal articles on machine learning and computational linguistics. Dr. Onan has been reviewing for several international journals, including Expert Systems with Applications, Plos One, the International Journal of Machine Learning and Cybernetics, the Journal of Information Science.

How to cite this article: Onan A. Mining opinions from instructor evaluation reviews: A deep learning approach. *Comput Appl Eng Educ.* 2019;1–22. <https://doi.org/10.1002/cae.22179>