DEÜ FMD 24(72), 855-868, 2022

Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi
Dokuz Eylul University Faculty of Engineering Journal of Science and Engineering

Basılı/Printed ISSN: 1302-9304. Elektronik/Online ISSN: 2547-958X

# Deep Learning-Based Sentiment Analysis on Education During the COVID-19 Pandemic

## COVID-19 Pandemi Döneminde Eğitimde Derin Öğrenmeye Dayalı Duygu Analizi

**Kemal Karga [1], Mansur Alp Toçoğlu [2*], Aytuğ Onan [3]**

[1,2] Manisa Celal Bayar University, Technology Faculty, Software Engineering Department, Manisa, TURKEY
[3] Katip Çelebi University, Faculty of Engineering and Architecture, Computer Engineering Department, İzmir, TURKEY
*Sorumlu Yazar / Corresponding Author* *: mansur.tocoglu@cbu.edu.tr

*Geliş Tarihi / Received: 17.11.2021*
*Kabul Tarihi / Accepted: 07.02.2022*
*Araştırma Makalesi/Research Article*
*DOI:10.21205/deufmd.2022247215*
*Atıf şekli/ How to cite: KARGA, K.,TOÇOĞLU, M.A.,ONAN,A.(2022). Deep Learning-Based Sentiment Analysis on Education During the COVID-19 Pandemic. DEÜ FMD 24(72), 855-868.*

### Abstract

The global COVID-19 pandemic in 2020 has led to catastrophic economic and social disruption. The pandemic has affected almost every aspect of our lives, including health, food, business organizations, and education. An essential shift in the higher education field has been occurred with the digitalization of instruction. In attempt to combat the pandemic, several higher education institutions throughout the world have begun to offer undergraduate and graduate courses online, either asynchronously or synchronously. During this period, people make considerable use of social media to gain news, information, social connections, and support. As a result, the immense quantity of electronic text documents has been shared on the Web related to COVID-19. In this paper, we present a deep learning-based sentiment analysis approach to analyze the impact of COVID-19 pandemic on the higher education. In this regard, the predictive performance of conventional machine learning algorithms (support vector machines, naïve bayes, logistic regression, and random forest) and deep neural networks (convolutional neural network, recurrent neural network, long short-term memory, and gated recurrent unit) are compared to each other. In addition, the empirical results obtained by the bidirectional encoder representations from transformers (BERT) have been evaluated. The comprehensive empirical results with different text representation models and classification algorithms indicate that deep neural networks can yield promising results for the task of analyzing the impact of COVID-19 related text documents on the higher education.

***Keywords:*** *Deep Learning, Sentiment Analysis, Text Mining, COVID-19, Higher Education*

### Öz

2020 yılında küresel COVID-19 pandemisi, ciddi ekonomik ve toplumsal kesintilere yol açtı. Pandemi sağlık, gıda, iş organizasyonları ve eğitim dahil olmak üzere hayatımızın neredeyse her alanını etkiledi. Eğitimin dijitalleştirilmesi ile birlikte yükseköğretim alanında önemli bir değişiklik yaşanmıştır. Pandemi ile mücadele amacıyla, dünya çapında birçok yükseköğretim kurumu, eş zamanlı veya eş zamansız olarak lisans ve lisansüstü derslerini çevrimiçi olarak sunmaya başlamıştır. Bu süre zarfında insanlar haber, bilgi, destek almak için ve sosyal bağlantılar kurmak için sosyal medyadan ciddi ölçüde yararlanmaktadırlar. Bu sayede, COVID-19 ile ilgili olarak Web'de çok miktarda elektronik metin belgesi paylaşılmıştır. Bu makalede, COVID-19 salgınının yüksek öğrenim üzerindeki etkisini analiz etmek için derin öğrenime dayalı bir duygu analizi yaklaşımı sunuyoruz. Bu bağlamda, geleneksel makine öğrenimi algoritmalarının (vektör destek makineleri, naive bayes, lojistik regresyon ve rastgele orman) ve derin sinir ağlarının (evrişimli sinir ağı, tekrarlı sinir ağı, uzun süreli bellek ve gated tekrarlı birim) performansları karşılaştırılmıştır. Buna ek olarak, transformerlardan gelen çift yönlü enkoder gösterimleri (BERT) tarafından elde edilen ampirik sonuçlar da değerlendirilmiştir. Farklı metin gösterim modelleri ve sınıflandırma algoritmalarına sahip kapsamlı ampirik sonuçlar, derin sinir ağlarının COVID-19 ile ilgili metin belgelerinin yüksek eğitim üzerindeki etkisini analiz etme görevi için umut verici sonuçlar verebileceğini göstermektedir.

***Keywords:*** *Derin Öğrenme, Duygu Analizi, Metin Madenciliği, COVID-19, Yüksek Öğrenim*

## 1. Introduction

Throughout the world, the COVID-19 epidemic that began in 2020 has caused devastating economic and social turmoil. The epidemic has had an impact on practically every element of our lives, including our health, food, corporate organizations, and educational institutions [1]. The COVID-19 pandemic process has brought together government, academic, and industrial organizations to collaborate toward a common aim of averting an outbreak. In the domains of health resource management, social policy formation, epidemic prevention and treatment, and vaccine research, this has resulted in a variety of outcomes [2]. Simultaneously, numerous social media posts and news articles about social policies, epidemic prevention and treatment practices, and vaccine development processes implemented in various countries worldwide during the COVID-19 outbreak have been shared on the Internet's media and communication platforms. It has been discovered that unofficial Internet sharing platforms account for a significant number of posts regarding infectious diseases and epidemics, and that these platforms supply the world with the first and most up-to-date information.

The World Health Organization (WHO) found that all major outbreaks were first shared via unauthorized Internet sites [3]. News stories uploaded on social media and other online communication channels can help track and monitor infectious disease outbreaks. To investigate the COVID-19 outbreak time-spatially, several countries throughout the world have established a substantial number of real-time, interactive mobile or online geographic information systems, websites, and applications. For acquiring accurate and timely information regarding the COVID-19 outbreak, advances in information and communication technologies, as well as data obtained from a range of sources, are crucial.

To combat the pandemic, several higher education institutions around the world have begun to offer undergraduate and graduate courses online, either asynchronously or synchronously [4]. With the digitalization of instruction, a significant shift has occurred in the field of higher education. The COVID-19 pandemic poses a significant challenge to educational systems [5]. As a marker of social inequality, Internet access in this context has served to further divide students into those who have access to a stable connection and those who do not, in a relatively short period of time [6]. In addition, practice sessions are vital parts of the learning process for students at some colleges, including natural science, engineering, and medicine, which cannot be properly handled by online learning [5, 6].

The COVID-19 pandemic has established a valuable experimental setting for measuring the effectiveness of online courses at universities, as well as identifying potential difficulties and positive experiences in the process. People make extensive use of social media during this time to obtain news, information, social connections, and support. As a result, an enormous amount of electronic text documents linked to COVID-19 have been disseminated on the Web. The digitalization that took place in education with the pandemic also caused many online contents to be created by students and educators expressing their views on the new process. In this regard, the main objective of this paper is to present a sentiment classification framework of COVID-19 pandemic on higher education. To the best of our knowledge, this is the first comprehensive analysis based on text mining, machine learning and deep learning techniques to identify sentiment orientation of undergraduate students towards COVID-19 period.

In this paper, we present a deep learning-based sentiment analysis approach to analyze the impact of COVID-19 pandemic on the higher education. In this regard, the predictive performance of conventional machine learning algorithms (i.e., support vector machines, Naïve Bayes, logistic regression, and random forest) and deep neural networks (i.e., convolutional neural network, recurrent neural network, long short-term memory, and gated recurrent unit). In addition, the empirical results obtained by the bidirectional encoder representations from transformers (BERT) have been evaluated. The comprehensive empirical results with different text representation models and classification algorithms indicate that deep neural networks can yield promising results for the task of analyzing the impact of COVID-19 related text documents on the higher education.

The rest of this paper is structured as follows: In Section 2, related work on sentiment analysis has been presented. In Section 3, the methodology of the study (i.e., the dataset, text representation models, conventional machine learning algorithms, neural language models, and deep neural networks) has been briefly discussed. Section 4 presents the experimental procedure and the empirical results. In Section 5, the concluding remarks of the study have been given.

## 2. Material and Method

This section introduces the text corpus, the machine learning-based sentiment analysis approach, and the deep-learning-based sentiment analysis approach.

### 2.1. Corpus

In this study, we generated a new dataset named "TURCOVIDEDU" to be used in the classification phase. To do so, we collected the raw data by conducting a survey among university students. Each participant is asked to write the advantages and disadvantages of university education during COVID-19 pandemic. As a result, we achieved a raw dataset which is composed of 12,018 documents. 5,882 of these documents are labeled as positive and the rest 6,136 documents are labeled as negative. To validate the dataset, we conducted an annotation process where two annotators labeled each document as positive, negative, or non-related. If there is a conflict between annotators, we eliminated the related document. After the validation process, we eliminated 606 documents in total. Table 1 shows the distribution of the raw and annotated documents in the dataset. In addition, we calculated the Cohen's kappa (K) metric as 0.82 which indicates an almost perfect agreement among the annotators.

**Table 1**. The distribution of the raw and annotated documents in TURCOVIDEDU dataset

| | Positive | Negative | Total |
|---|---|---|---|
| Raw | 5,882 | 6,136 | 12,018 |
| Annotated | 5,588 | 5,824 | 11,412 |

After the dataset construction phase, we preprocessed the TURCOVIDEDU dataset to be used for the empirical research. First, we converted all letters to lowercase. Then, we removed numeric characters, extra spaces, and punctuation marks, respectively. Next, we normalized the dataset by stemming each term using Snowball-stemmer (SS) [34].

### 2.2. Machine Learning Based Sentiment Analysis

There are two main stages in machine learning-based sentiment analysis: extracting features from the data and representing them as feature vectors and training supervised learning algorithms on the feature vectors to obtain the learning model. The class labels for unseen instances have been determined using the obtained learning model [35]. We used three term weighting schemes (i.e., term-presence, term-frequency, and TF-IDF) and three N-gram models to perform machine learning-based sentiment analysis (namely, bigram, unigram, and trigram model). Four conventional machine learning algorithms (i.e., support vector machines, Naïve Bayes, logistic regression, and random forest) have been utilized. The remainder of this section briefly discusses feature extraction schemes and supervised learning models.

#### 2.2.1. Feature Construction

Converting text documents into feature vectors is a critical task when processing text documents with supervised learning algorithms. In text mining and information retrieval tasks, a frequently used and successful scheme is the bag-of-words (BOW) framework. A text document is viewed as a bag of words in this framework and is represented by a vector containing all the words encountered in the document, without regard for syntax, word orderings, or grammar [36]. Each text document has been represented in this framework using the frequency of each word. To obtain the learning model, the set of features was used to train the supervised learning algorithm. There are three types of weighting schemes that can be used with the bag-of-words framework: term presence (TP), term frequency (TF), and TF-IDF.

#### 2.2.2. Classification Algorithms

We considered four supervised learning methods to obtain learning models based on the feature sets. The algorithms have been summarized as follows:

Support vector machines (SVMs) are supervised learning algorithms for classification and regression tasks. SVM computes a hyperplane in a higher-dimensional space to denote the boundary between instances of distinct classes [37].

The Naïve Bayes algorithm (NB) is a statistical supervised learning algorithm based on Bayes' theorem and the assumption of conditional independence [38].

Logistic regression (LR) is a linear classification algorithm that provides a framework for solving classification problems using linear regression. A linear classification scheme was created using a linear regression model and transformed target variables in this scheme [39].

The random forest (RF) algorithm is a combination of the bagging and random subspace algorithms. Decision trees were used as the base learner in this algorithm. Each tree was constructed using bootstrap samples taken from the training data. A random feature selection process was used to generate diversity among the base learners. Thus, even in the presence of noisy or irrelevant data, the model produces satisfactory results [40].

### 2.3. Deep Learning Based Sentiment Analysis

The text corpus was represented by three word embedding schemes for the deep learning-based sentiment analysis (namely, word2vec, fastText and GloVe). The four conventional deep learning architectures were used to process text (i.e., convolutional neural network, recurrent neural network, gated recurrent unit, and long short-term memory). In addition, the empirical results obtained by the bidirectional encoder representations from transformers (BERT) have been evaluated. The remainder of this section briefly discusses neural language models and deep learning architectures.

#### 2.3.1. Neural Language Models

A representation based on word embedding enables the learning of distributed expressions for words in low-dimensional space [41]. We considered three different word embedding schemes (word2vec, fastText, and GloVe) in conjunction with deep learning architectures for this study. We have briefly described the representation schemes. The word2vec model is an unsupervised, computationally efficient model for learning word embeddings from text documents. The word2vec model is composed of two components: a continuous bag of words (CBOW) model and a continuous skip-gram model [42]. The CBOW model predicts the target word based on the context words that surround it over a k-word window. In comparison, the skip-gram model predicts the target word's context words.

FastText is a computationally efficient representation scheme for learning word embeddings from text documents. Each word was treated as a collection of character *n*-grams in this scheme [43]. When compared to word2vec, the fastText scheme can perform better on morphologically rich languages and rare words [44].

The global vectors (GloVe) model is an unsupervised prediction algorithm for generating vector representations of words. The global matrix factorization scheme has been used to incorporate the local context-based learning of the word2vec model. Training was conducted using global statistics on word-word co-occurrence extracted from the text corpus. Linear structures in the word vector space have been extracted using the training procedure [45, 46].

### 2.3.2. Deep Learning Architectures

Deep learning architectures enable the acquisition of multi-level representations of features. The architectures are designed to identify learning models through the hierarchical processing of multiple layers/or stages of nonlinear information [47]. The remainder of this section discusses the deep learning architectures used in the empirical analysis in detail.

Convolutional neural networks (CNNs) are architectures based on deep neural networks that process data using a grid-based topology. CNN is defined by a particular type of mathematical operation known as convolution. Convolution has been performed using one or more convolutional layers. An input layer, an output layer, and hidden layers comprise a typical convolutional neural network architecture. The architecture's hidden layers are divided into several categories, including convolutional layers, pooling layers, fully connected layers, and normalization layers [48].

Another type of deep learning architecture is the recurrent neural network (RNN), which is used to process sequential data. The connections between neurons in an RNN form a directed graph. Internal state has been used to process the sequence of inputs in this architecture. As a result, the architecture is suitable for sequential tasks such as speech recognition [49].

Another deep learning architecture based on recurrent neural networks is long short-term memory networks (LSTM). The exploding or vanishing gradient problem plagues conventional RNN architectures. RNNs are incapable of properly handling arbitrarily long input sequences. As a result, LSTM employs forget gates to circumvent the issues. Backpropagation of error is permitted in the LSTM architecture up to a finite number of time steps. A typical LSTM unit consists of a cell and three different types of gates: an input gate, an output gate, and a forget gate. The gates' open and close operations have been used to specify which information should be preserved and when it should be accessed [50].

Another deep learning architecture based on recurrent neural networks is the gated recurrent unit (GRU). There are two gates in a typical GRU architecture (namely, the reset gate and the update gate) [51].

Google discovered Bidirectional Encoder Representations from Transformers (BERT) in 2018 [52]. BERT is a new pre-training bidirectional language presentation model. BERT employs Masked Language Modeling in the pre-training phase to enable pre-trained deep bidirectional representations and the Next Sentence Prediction method to discover semantic relationships between sentences. Due to the heavily engineered tasks used during the pre-training process, BERT, as the first fine-tuning-based representation model, enables the pre-trained model to be fine-tuned for a wide variety of classification tasks.

## 3. Results

The experimental method and findings are discussed in this section.

### 3.1 Experimental Procedure

In the empirical analysis, we focused on the identification of efficacy of the COVID-19 pandemic shutdowns in undergraduate education by comparing classification performances of four well-known conventional machine learning algorithms (SVM, NB, LR, RF), deep learning architectures (LSTM, GRU, RNN, CNN) and BERT classifier. We also used pre-trained models for the experiments on deep learning architectures and BERT classifier. In all cases, we utilized 10-fold cross-validation and used hyper-parameter optimization. We set input sequence length as 100 of each embedding layer and used binary_crossentropy for the loss function and adamax as the optimizer in all deep learning architectures. We constructed the layers of the dense neural network (DNN) of each architecture as input, hidden and output, respectively. Besides, the dropout rate of all network is set to 0.3, hidden layer neuron number as 50, and the activation function as relu. Regarding the structure of the CNN architecture, we set the convolutional layer parameters, activation function as softmax, kernel size as 16 and filter value as 100. We used pooling as the next layer with a pooling size of 2. In the next step, we flattened the 2D pooled feature maps to a one-dimension vector. In the architectures of LSTM, GRU and RNN, we set the unit size as 100. In the BERT classifier, we set the parameters of the model as batch size: 32, learning rate: 5e-5, number of training epoch: 4 and optimizer: Adam. In addition, the technical specifications of the computer we used in the experiment section are as follows: CPU name: AMD Ryzen 5 4500U, CPU processing speed: 2.38 GHz, CPU cache size: 8 MB, CPU core: 6, RAM type: DDR4, RAM size: 8 GB, disk name: NVMe SSD and disk size: 512 GB.

Firstly, we focused on the effects of the normalization process of the dataset. To do so, we compared the predictive performances of four different conventional machine learning algorithms (namely, SVM, NB, LR and RF) and four different deep learning architectures (namely, CNN, RNN, LSTM, GRU) using the dataset in two stemming forms which are raw and SS. According to the overall accuracy values displayed in Tables 2 and 3, we achieved higher performances for all cases by using the stemmed dataset rather than using the raw form. As a result, unless otherwise stated, we stemmed each term using SS stemmer in all following experiments.

**Table 2**. Comparison of accuracy results of four conventional machine learning algorithms in terms of two different stemmer methods

| **Stemmer** | **SVM** | **NB** | **LR** | **RF** |
|---|---|---|---|---|
| Raw | 0.8962 | 0.8925 | 0.8922 | 0.8570 |
| SS | 0.9026 | 0.8968 | 0.8941 | 0.8671 |

**Table 3**. Comparison of accuracy results of four deep learning architectures in terms of two different stemmer methods

| **Stemmer** | **CNN** | **RNN** | **LSTM** | **GRU** |
|---|---|---|---|---|
| Raw | 0.9081 | 0.8946 | 0.9047 | 0.9029 |
| SS | 0.9084 | 0.8977 | 0.9083 | 0.9102 |

Table 4 displays the comparison of the classification performances of four conventional machine learning algorithms (namely, SVM, NB, LR and RF) using three different feature extraction methods (namely, TF, TP and TF-IDF) and three N-gram models (unigram, bigram, and trigram). Regarding the performances of the algorithms, SVM, NB and LR performed slightly similar results and RF achieved the lowest accuracy result among others. On the other hand, representation of the data using unigram model and TF-IDF weighting scheme performed the highest average accuracy result as 0.8901.

**Table 4.** Comparison of classification performances of four conventional machine learning algorithms using different representation methods.

| Representation | SVM | NB | RF | LR | Average |
|---|---|---|---|---|---|
| Unigram, TF | 0.8840 | 0.8963 | 0.8684 | 0.8959 | 0.8862 |
| Unigram, TP | 0.8893 | 0.8974 | 0.8663 | 0.8955 | 0.8871 |
| Unigram, TF-IDF | 0.9026 | 0.8968 | 0.8671 | 0.8941 | **0.8901** |
| Bigram, TF | 0.8850 | 0.8919 | 0.8672 | 0.8919 | 0.8840 |
| Bigram, TP | 0.8886 | 0.8937 | 0.8636 | 0.8932 | 0.8848 |
| Bigram, TF-IDF | 0.8991 | 0.8915 | 0.8634 | 0.8912 | 0.8863 |
| Trigram, TF | 0.8847 | 0.8905 | 0.8642 | 0.8920 | 0.8828 |
| Trigram, TP | 0.8877 | 0.8929 | 0.8645 | 0.8917 | 0.8842 |
| Trigram, TF-IDF | 0.8983 | 0.8910 | 0.8631 | 0.8910 | 0.8858 |

The aim in the next experiment is to compare the predictive performances of two different embedding layer construction forms (namely, self-trained and pre-trained) using three different embedding methods (namely, Glove, Word2Vec SG and Word2Vec CBOW) in conjunction with four deep learning architectures (namely, CNN, RNN, LSTM and GRU). In the self-trained form, we trained the embedding layer using the dataset TURCOVIDEDU. However, in the pre-trained form, we utilized a new raw dataset, which is composed of 116K documents related to COVID-19 and education news, in the training process of the embedding layers. Tables 5 and 6 show the predictive performances in terms of accuracy values of the four deep learning architectures in conjunction with three different embedding methods constructing the embedding layers as self-trained and pre-trained, respectively. Regarding the predictive performances of the four deep learning architectures, GRU outperformed other classifiers in almost all cases. LSTM performed the second highest performances; CNN achieved the third highest performances and the lowest predictive performances are obtained by RNN classifier. Regarding the predictive performances of the embedding methods, Word2Vec type methods performed higher accuracy values in almost all cases compared to Glove. Considering the performances of using self-trained and pre-trained embedding layer construction forms, we achieved higher accuracy values for both forms in different cases. Using self-trained embedding layers in all CNN architectures and Glove related results provided higher performances. On the other hand, RNN, LSTM and GRU architectures obtained higher accuracy values using pre-trained embedding layers constructed by Word2Vec type embedding methods. In addition, the combination of GRU, Word2Vec SG embedding method and pre-trained construction form achieved the highest accuracy value as 0.9130 among others. Figure 1 shows the comparison of the average accuracy results of the four deep learning architectures in terms of the two embedding layer construction forms and Figure 2 displays the comparison of the average accuracy results of the three embedding methods in terms of two embedding layer construction forms.

**Table 5**. Comparison of accuracy values of four deep learning architectures using self-trained embedding layers constructed by using three different embedding methods

| Embedding Method | CNN | RNN | LSTM | GRU |
|---|---|---|---|---|
| Glove | 0.9041 | 0.8896 | 0.9064 | 0.9090 |
| Word2Vec SG | 0.9053 | 0.8924 | 0.9062 | 0.9069 |
| Word2Vec CBOW | 0.9047 | 0.8924 | 0.9010 | 0.9019 |

**Table 6**. Comparison of accuracy values of four deep learning architectures using pre-trained embedding layers constructed by using three different embedding methods

| Embedding Method | CNN | RNN | LSTM | GRU |
|---|---|---|---|---|
| Glove | 0.8946 | 0.8885 | 0.9040 | 0.9055 |
| Word2Vec SG | 0.9016 | 0.8973 | 0.9117 | **0.9130** |
| Word2Vec CBOW | 0.8993 | 0.8934 | 0.9041 | 0.9074 |


**Figure 1.** Comparison of the average accuracy results of four deep learning architectures in terms of two embedding layer construction forms.


**Figure 2.** Comparison of the average accuracy results of three embedding methods in terms of two embedding layer construction forms.

In Table 7, we compared the classification performances of the BERT classifier using four different forms of BERTurk pre-trained language model in terms of accuracy, f-score, recall and precision values. According to the results, we obtained the highest accuracy value as 0.9666 in this study using the BERTurk_cased pre-trained language model. In addition, we compared the performances of cased and uncased models, in which cased model performed higher results in all cases. On the other hand, decreasing the vocabulary size of a pre-trained model to 128K did not perform higher performance in all cases.

**Table 7**. Comparison of accuracy values of four deep learning architectures using pre-trained embedding layers constructed by using three different embedding methods

| **Pre-trained language models** | **accuracy** | **f-score** | **recall** | **precision** |
|---|---|---|---|---|
| BERTurk_uncased | 0.956117 | 0.956112 | 0.956259 | 0.956102 |
| BERTurk_cased | **0.966607** | 0.966595 | 0.966640 | 0.966560 |
| BERTurk_uncased_128K | 0.957909 | 0.957896 | 0.957975 | 0.957846 |
| BERTurk_cased_128K | 0.958517 | 0.958510 | 0.958644 | 0.958472 |

Figure 3 displays the comparison of the average accuracy results of all machine learning approaches (namely, conventional machine learning algorithms, deep learning algorithms and BERT classifier) used in the empirical analysis. BERT classifier performed the highest average accuracy value as 0.9598. Despite similar results of LSTM and GRU architectures, GRU achieved the second highest average accuracy result as 0.9086. On the other hand, RNN architecture performed the lowest performance as 0.8931 among deep learning architectures. Regarding the average accuracy results of the conventional machine learning algorithms, NB and LR algorithms performed the highest values among conventional algorithms which are followed by SVM classifier. Lastly, RF achieved the lowest result as 0.8653 among all classifiers of the empirical analysis.

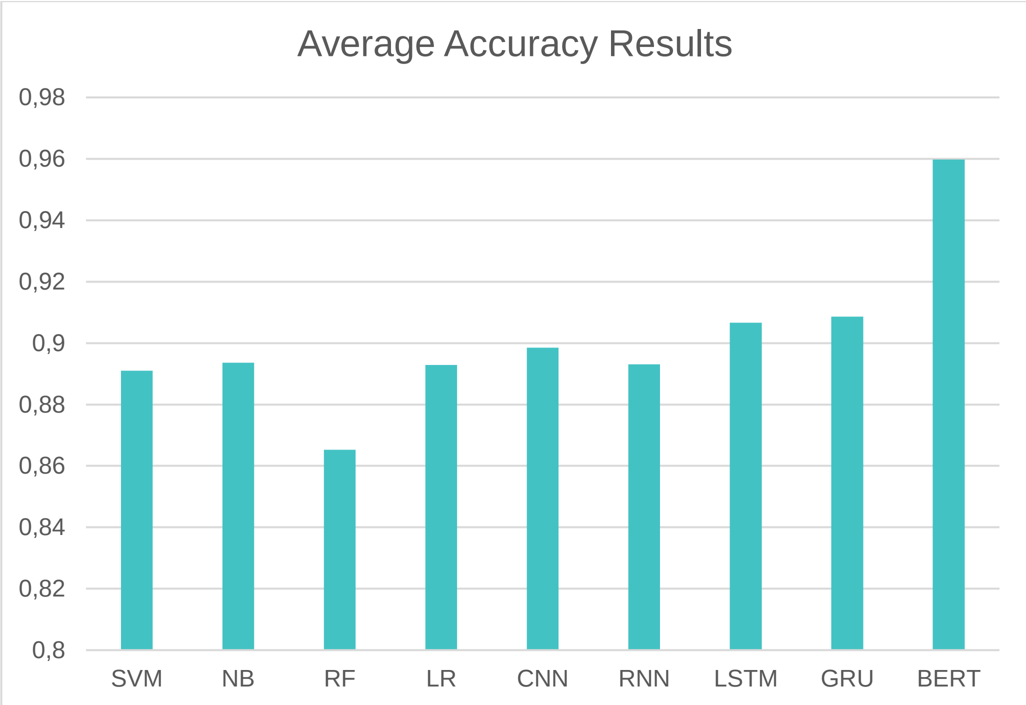


**Figure 3.** Comparison of the average accuracy performances of all three machine learning approaches.

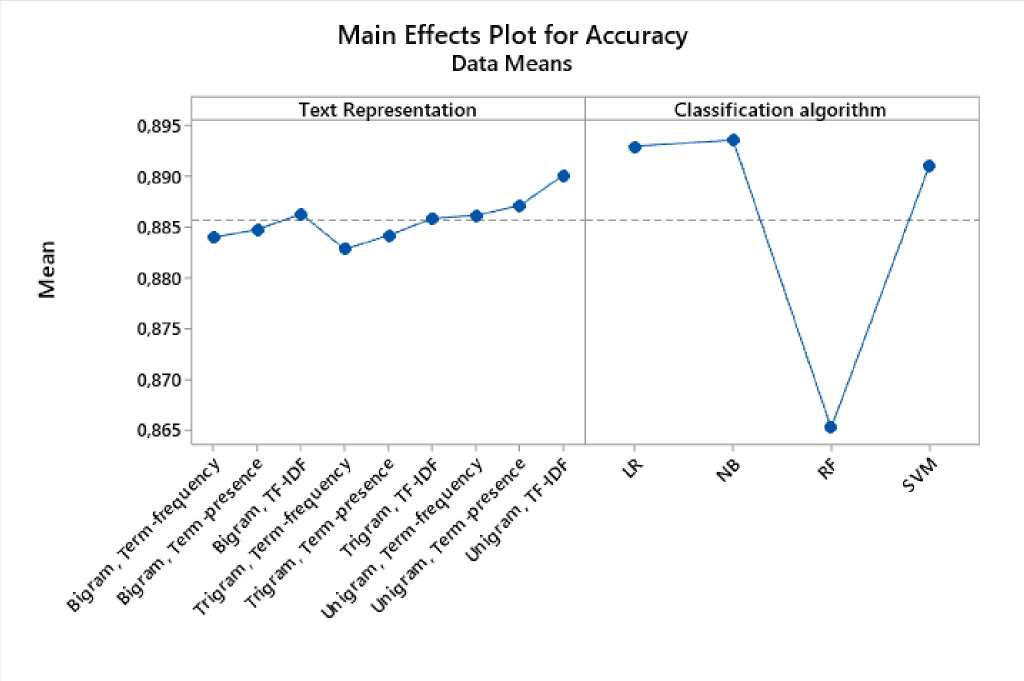


**Figure 4.** The main effects plot for accuracy values obtained by machine learning architectures.

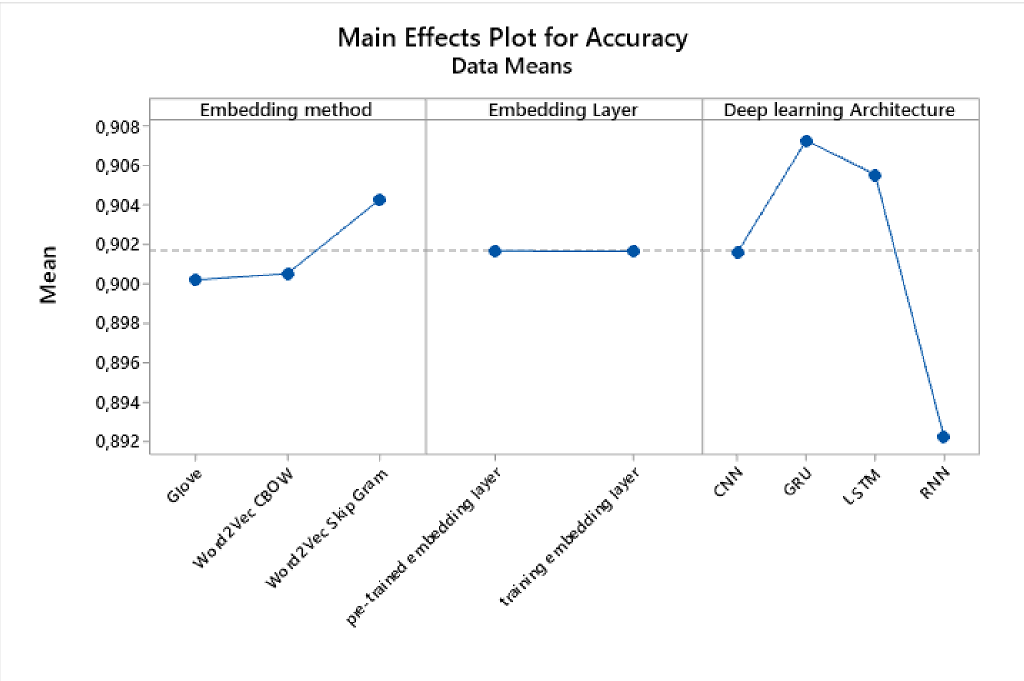


**Figure 5.** The main effects plot for accuracy values obtained by deep learning architectures.

To summarize the main findings of the empirical results, the main effects plot for accuracy values for conventional machine learning algorithms and deep learning architectures have been depicted in Figure 4 and Figure 5, respectively. As it can be observed from Figure 4, regarding conventional machine learning-based sentiment analysis unigram text representation scheme in conjunction with TF-IDF term weighting outperforms the other text representation schemes. Regarding the predictive performance of conventional supervised learning methods, Naïve Bayes algorithm outperforms the other classifiers. In Figure 5, the results obtained by deep learning-based sentiment analysis have been summarized. As can be observed from Figure 5, word2vec (Skip Gram model) outperforms the other neural language models. Training embedding layer yields higher predictive performance compared to pre-trained embedding layer. In addition, gated recurrent unit outperforms the other conventional deep neural networks for the task.

## 4. Discussion and Conclusion

The automatic identification of positive and negative effects of COVID-19 pandemic in undergraduate education can be regarded as a challenging task in natural language processing since the related cases can be encountered on social media platforms frequently. There are three main contributions in this study. First, we collected and validated a new dataset, composed of 11,412 validated documents, by conducting a survey among undergraduate students. Second, we performed an extensive empirical analysis of three different machine learning approaches. Third, we created a new pre-trained embedding layer by using a new raw dataset, composed of 116,085 documents, to be used in deep learning architectures.

We achieved encouraging predictive results in the empirical analysis of this study by performing three different machine learning approaches which are four conventional machine learning algorithms (SVM, NB, LR, RF), four deep learning architectures (CNN, RNN, LSTM, GRU) and the BERT classifier. According to the overall results, BERT classifier outperformed both deep learning approaches and conventional machine learning algorithms in all cases by achieving an average accuracy result as 0.9598. The second highest average accuracy values are performed by deep learning architectures GRU, LSTM and CNN, respectively.

We compared the performances of three different embedding methods (namely, Glove, Word2Vec SG and Word2Vec CBOW) in the construction phase of the embedding layers in deep learning architectures. Regarding the overall results, Word2Vec SG performed the highest average accuracy value as 0.9059. In addition, we constructed a new pre-trained embedding layer by using a new raw dataset to be used in deep learning architectures. According to the results, we achieved higher performances than using self-trained embedding construction form in general.

This study can be extended in several dimensions. First, the number of the documents in two datasets can be increased. Second, it can be beneficial to propose a novel deep learning architecture.


## Acknowledgment

We are thankful to the participants who attended to the survey.


## References


[1] Chawla, S., Mittal, M., Chawla, M., and Chawla, and Goyal, L.M. 2020. Corona virus - SARS-CoV-2: An insight to another way of natural disaster, EAI Endorsed Trans. Pervasive Health Technol, Vol. 6, pp. 25-33. DOI: 10.4108/eai.28-5-2020 164823

[2] Wang, L.L., and Lo, K. 2021. Text mining approaches for dealing with the rapidly expanding literature on COVID-19, Brief. Bioinform, Vol. 22, pp. 781–799. DOI: 10.1093/bib/bbaa296

[3] Gajewski, N.K, Peterson, A.E., Chitale, R.A., and Pavlin, J.A. 2014. A review of evaluations of electronic event-based biosurveillance systems, PLoS One, Vol. 9, DOI: 10.1371/journal.pone.0111222

[4] Bismala, L., and Manurung, Y.M. 2021. Student satisfaction in e-learning along the COVID-19 pandemic with importance performance analysis, Int. J. Eval. Res. Educ. (IJERE), Vol. 10, DOI: 10.11591/ijere.v10i3.21467

[5] Daniel, S.J. 2020. Education and the COVID-19 pandemic, Prospects (Paris), Vol. 49, pp. 1–6. DOI: 10.1007/S11125-020-09464-3

[6] Bilecen, B. 2020. Commentary: COVID-19 pandemic and higher education: International mobility and students' social protection, International Migration, Vol. 58, pp. 263–266. DOI: 10.1111/imig.12749

[7] Onan, A., Korukoğlu, S., and Bulut, H. 2016 Ensemble of keyword extraction methods and

classifiers in text classification, Expert Syst. Appl., Vol. 57, pp. 232–247. DOI: 10.1016/J.ESWA.2016.03.045

[8] Onan, A. 2016. Classifier and feature set ensembles for web page classification,J. Inf. Sci., Vol. 42, pp. 150–165. DOI: 10.1177/0165551515591724

[9] Onan, A., Korukoğlu, S., and Bulut, H. 2016. A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification, Expert Systems With Applications, Vol. 62, pp. 1–16. DOI: 10.1016/j.eswa.2016.06.005

[10] Onan, A., and Korukoğlu, S. 2017. A feature selection model based on genetic rank aggregation for text sentiment classification,Journal of Information Science, Vol. 43, pp. 25–38. DOI: 10.1177/0165551515613226

[11] Onan, A. 2017. Hybrid supervised clustering based ensemble scheme for text classification, Kybernetes, Vol. 46, pp. 330–348. DOI: 10.1108/K-10-2016-0300

[12] Onan, A. and Tocoglu, M.A. 2021. A term weighted neural language model and stacked bidirectional LSTM based framework for sarcasm identification, IEEE Access, Vol. 9, pp. 7701–7722. DOI: 10.1109/ACCESS.2021.3049734

[13] Onan, A., Korukoğlu, S., and Bulut, H. 2017 A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification, Information Processing and Management Vol. 53, pp. 814–833. DOI: 10.1016/j.ipm.2017.02.008

[14] Toçoğlu, M.A., and Onan, A. 2019. Satire detection in Turkish news articles: A machine learning approach, in Communications in Computer and Information Science, Cham: Springer International Publishing, pp. 107–117. DOI: 10.1007/978-3-030-27355-2_8

[15] Onan, A. 2018. Review spam detection based on psychological and linguistic features, 26th Signal Processing and Communications Applications Conference (SIU), 2-5 May, Izmir, Turkey

[16] Onan, A. 2018. An ensemble scheme based on language function analysis and feature engineering for text genre classification, Journal of Information Science, Vol. 44, pp. 28–47. DOI: 10.1177/0165551516677911

[17] Jahanbin, K., and Rahmanian, V. 2020. Using twitter and web news mining to predict COVID-19 outbreak, Asian Pacific Journal of Tropical Medicine, Vol. 13, pp. 378-380. DOI: 10.4103/1995-7645.279651

[18] Ordun, C., Purushotham S., and Raff, E. 2020. Exploratory analysis of covid-19 tweets using topic modeling, UMAP, and DiGraphs, https://arxiv.org/abs/2005.03082 (Date of Access: 06.04.2020)

[19] Peng, Z., Wang, R., Liu, L., and Wu, H. 2020. Exploring urban spatial features of COVID-19 transmission in Wuhan based on social media data, ISPRS International Journal of Geo-Information, Vol. 9, DOI: 10.3390/ijgi9060402

[20] Li, D., Chaudhary, H., and Zhang, Z. 2020. Modeling spatiotemporal pattern of depressive symptoms caused by COVID-19 using social media data mining, International Journal of Environmental Research and Puplic Health, Vol. 17, DOI: 10.3390/İJERPH17144988

[21] Chen, N., Zhong, Z., and Pang, J. 2021. An exploratory study of COVID-19 information on Twitter in the Greater Region, Big Data and Cognitive Computing, Vol. 5, DOI: 10.3390/bdcc5010005

[22] Boon-Itt, S., and Skunkan, Y. 2020. Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study, JMIR Public Health and Surveill., Vol. 6, DOI: 10.2196/21978

[23] Onan, A., 2021. COVID-19 ile İlgili Sosyal Medya Gönderilerinin Metin Madenciliği Yöntemlerine Dayalı Olarak Zaman-Mekansal Analizi, European Journal of Science and Technology, Vol. 26, pp. 138-143. DOI: 10.31590/ejosat.957020

[24] Onan, A., and Toçoğlu, M.A. 2020. Weighted word embeddings and clustering-based identification of question topics in MOOC discussion forum posts, Computer Applications in Engineering Education, Vol. 29, pp. 675–689. DOI: 10.1002/cae.22252

[25] Onan, A. 2021. Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach, Computer Applications in Engineering Education, Vol. 29, pp. 572–589. DOI: 10.1002/cae.22253

[26] Bustillos, R.O., Cabada, R.Z., Estrada, M.L.B, and Perez, Y.H. 2019. Opinion mining and emotion recognition in an intelligent learning environment ,Computer Applications in Engineering Education, Vol. 27, pp. 90–101. DOI: 10.1002/cae.22059

[27] Cabada, R.Z., Estrada, M. L. B., and Bustillos, R. O. 2018. Mining of Educational Opinions with Deep Learning, Journal of Universal Computer Science, Vol. 24, pp. 1604–1626.

[28] Nguyen, H. T., and Nguyen, M.L. 2018. Multilingual opinion mining on YouTube – A convolutional N-gram BiLSTM word embedding, Information Processing and Management Vol. 54, pp. 451–462. DOI: 10.1016/j.ipm.2018.02.001

[29] Lin, Q., Zhu, Y. , Zhang, S., Shi , P., Guo, Q., and Niu, Z. 2019. Lexical based automated teaching evaluation via students' short reviews, Computer Applications in Engineering Education, Vol. 27, pp. 194–205. DOI: 10.1155/2021/5596518

[30] López, M. B., Alor-Hernández G., Sánchez-Cervantes, J. L., Pilar Salas-Zárate M., and Paredes-Valverde, M.A. 2018. EduRP: an Educational Resources Platform based on Opinion Mining and Semantic Web, Journal of Universal Computer Science, Vol. 24, pp. 1515–1535. DOI: 10.3217/JUCS-024-11-1515

[31] Chen, T., Peng, L., Jing, B., Wu, C., Yang, J., and Cong, G. 2020. The impact of the COVID-19 pandemic on user experience with online education platforms in China, *Sustainability*, Vol. 12, DOI: 10.3390/su12187329

[32] Komasawa, N., Terasaki, F., Nakano, T., Saura, R. , and Kawata, R. 2020. A text mining analysis of perceptions of the COVID-19 pandemic among final-year medical students, Acute Medicine Surgery, Vol. 7, pp. DOI: 10.1002/ams2.576

[33] Kim, E.-J., Kim J. J., and Han, S.-H. 2021. Understanding student acceptance of online learning systems in higher education: Application of social psychology theories with consideration of user innovativeness, Sustainability, Vol. 13, DOI: 10.3390/su13020896

[34] Porter, M. F. 2001. A language for stemming algorithms. http://snowball.tartarus/texts/introduction (Date of Access: 05.10.2001)

[35] Lane, P. C. R., Clarke D., and Hender, P. 2012. On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data, Decision Support System, Vol. 53,pp.712–718. DOI: 10.1016/J.DSS.2012.05.028

[36] Hackeling, G. 2017. Mastering machine learning with scikit-learn -, 2nd ed, Birmingham, England: Packt Publishing, 254p.

[37] Vapnik, V. 2014. The nature of statistical learning theory. New York, NY: Springer, 314p.

[38] Li, X., Li, S., Li, J., Yao, J., and Xiao, X. 2021. Detection of fake-video uploaders on social media using Naive Bayesian model with social cues, Scientific Reports, Vol. 11, DOI: 1038/s41598-021-95514-5

[39] Hastie, T., Tibsharani, R., and Friedman, J. 2009. Springer Series in Statistics The Elements of, Math. Intell, Vol. 27, pp. 83–85.

[40] Breiman, L., Last, M., and Rice, J. 2006. Random forests: Finding quasars, in Statistical Challenges in Astronomy, New York: Springer-Verlag, pp. 243–254.

[41] Bengio, Y., and Senecal, J.S. 2008. Adaptive importance sampling to accelerate training of a neural probabilistic language model, IEEE Transactions on Neural Networks, Vol. 19, DOI: 10.1109/TNN.2007.912312

[42] Rezaeinia, S. M., Rahmani, R., Ghodsi, A., and Veisi, H. 2019. Sentiment analysis based on improved pre-trained word embeddings, Expert System with Application, Vol. 117, pp. 139–147. DOI: 10.1016/j.eswa.2018.08.044

[43] Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. Efficientestimation of word representations in vector space, https://arxiv.org/abs/1301.3781.(Date of Access: 07.09.2013)

[44] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. 2016. Fasttext. zip: Compressing text classification models. https://arxiv.org/abs/1612.03651 (Date of Access: 12.12.2016)

[45] Di, W., Bhardwaj, A., and Wei, J. 2018. Deep Learning Essentials: Your hands-on guide to the fundamentals of deep learning and neural network modeling. Packt Publishing, 284p.

[46] Pennington, J., Socher, R., and Manning, C. 2014. Glove: Global vectors for word representation, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), October 2014, Doha, Qatar, 1532-1543

[47] LeCun, Y., 1989. Generalization and network design strategies, Vol. 19, Amsterdam: Elsevier

[48] Elman, J.L. 2020. Finding structure in time, in Connectionist psychology: A text with readings, Psychology Press, 352p.

[49] Zhang, L., Wang, S., and Liu, B. 2018. Deep learning for sentiment analysis: A survey, Wiley Interdiscip. Rev. Data Min. Knowl. Discov., Vol. 8, DOI: 10.1002/widm.1253

[50] Rojas-Barahona, L.M. 2016. Deep learning for sentiment analysis: Language and Linguistics Compass, Language and Linguist. Compass, Vol. 10, pp. 701–719. DOI: 10.111/Inc3.12228

[51] Cho, K., et al. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. https://arxiv.org/abs/1406.1078 (Date of Access: 03.06.2014)

[52] Devlin, J., Chang , M.-W., Lee, K. , and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional Transformers for language understanding. https://arxiv.org/abs/1810.04805 (Date of Access: 11.10.2018)