

| | | |
|---|---|------------------------------------|
| ITC 3/51 Information Technology and Control Vol. 51/ No. 3 / 2022 pp. 409-428 DOI 10.5755/j01.itc.51.3.30276 | BERT-based Transfer Learning Model for COVID-19 Sentiment Analysis on Turkish Instagram Comments | |
| | Received 2021/12/07 | Accepted after revision 2022/05/04 |
| |  http://dx.doi.org/10.5755/j01.itc.51.3.30276 | |

HOW TO CITE: Karayığit, H., Akdagli, A., Acı, Ç. İ. (2022). BERT-based Transfer Learning Model for COVID-19 Sentiment Analysis on Turkish Instagram Comments. *Information Technology and Control*, 51(3), 409-428. <http://dx.doi.org/10.5755/j01.itc.51.3.30276>

BERT-based Transfer Learning Model for COVID-19 Sentiment Analysis on Turkish Instagram Comments

Habibe Karayığit

Department of Electrical and Electronics Engineering, Mersin University, 33343, Turkey
e-mail: d2014242@mersin.edu.tr

Ali Akdagli

Department of Electrical and Electronics Engineering, Mersin University, 33343, Turkey
e-mail: akdagli@mersin.edu.tr

Çiğdem İnan Acı

Department of Computer Engineering, Mersin University, 33343, Turkey
e-mail: caci@mersin.edu.tr

Corresponding author: d2014242@mersin.edu.tr

First seen in Wuhan, China, coronavirus (COVID-19) became a worldwide epidemic. Turkey's first reported case was announced on March 11, 2020—the day the World Health Organization declared COVID-19 is a pandemic. Due to the intense and widespread use of social media during the pandemic, determining social media's role and effect (i.e., positive, negative, neutral) gives us essential information about society's perspective on events. In our study, two datasets (i.e., Dataset1, Dataset2) consisting of Instagram comments on COVID-19 were com-

posed between different dates of the pandemic, and the change between users' feelings and thoughts about the epidemic was analyzed with Latent Dirichlet Allocation (LDA) and text mining algorithms. The datasets are the first publicly available Turkish datasets on the sentiment analysis of COVID-19, as far as we know. The sentiment analysis of Turkish Instagram comments was performed using machine learning models (i.e., traditional machine learning (TML), deep learning (DL), and Bidirectional Encoder Representations from Transformers (BERT)-based transfer learning). The balanced versions of these datasets (i.e., resDataset1, resDataset2) in the experiments were evaluated with the original ones. Compared with TML models (i.e., Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF)) and DL models (i.e., Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Convolutional Recurrent- Neural Networks (GCR-NN), the BERT-based transfer learning model achieved the highest classification success with 0.7864 macro-averaged F1-score values in resDataset1 and 0.7120 in resDataset2. It has been proven that using a pre-trained language model in Turkish datasets is more successful than other models in terms of classification performance.

KEYWORDS: COVID-19, Instagram, Sentiment Analysis, Deep Learning, BERT, Transfer Learning, Latent Dirichlet Allocation.

1. Introduction

COVID-19 disease, seen in about 180 countries, has had a devastating impact worldwide. COVID-19 cases have been reported in Turkey every day since March 11, 2020. The total number of cases reached 5,072,462, and the number of deaths was 43,821 in Turkey by May 12, 2021. Worldwide, 159,319,384 COVID-19 cases had been reported, and 3,311,780 people had died by that date [21]. Measures such as travel bans, quarantines, curfews, social distancing, and mask-wearing have been taken to stop the epidemic in Turkey.

In Turkey, internet use increased by 51 percent for fixed subscribers and 56 percent for mobile subscribers in the first quarter of the pandemic compared to 2019 because of curfews and social distancing [41]. The most crucial factor in this increase is social media traffic. Social media has been used frequently for outbreak-related interaction. Users have seen social networks as alternative news sources during the pandemic [30]. However, fake information and rumors can spread through social media uncontrollably. Information pollution in social media causes people to panic and fear. A study has shown that posts, comments, and content on social media during the pandemic have not been accurate [4].

Turkey's most popular social media networks are YouTube, Instagram, Whatsapp, Facebook, and Twitter [42]. Instagram is a trendy social media network that primarily shares photos and videos. Comments related to Instagram posts can be obtained if it is pub-

lic. The Instagram text limit is 2,200 characters, and the hashtag limit is 30, making it a problematic choice for natural language processing (NLP). However, with 37 million Instagram users, Turkey ranks sixth in its use [42]. Therefore, Instagram is a suitable environment for scientific research in this country despite processing data in terms of NLP.

Social media is used to disseminate advice on COVID-19, provide psychological first aid, and exchange information. Computer scientists and researchers make it an essential data source for learning about public attitudes regarding societal decisions. The sentiment analysis method is frequently used to determine the feelings and opinions of people on social media. Sentiment analysis is a controlled NLP problem and, in its simplest definition, determines whether posted comments express positive, negative, or neutral sentiments [5].

The motivation of this research can be summarized as follows: Many studies conducted to analyze content related to COVID-19 focus on English-language data on Twitter, and studies in other languages are in the minority [15]. It is the first study to conduct a COVID-19 sentiment analysis in Turkish that tries to determine the social media impact on the pandemic using the Instagram social network, as far as we know. Another motivation of the study is to make it possible to follow the sentiment changes seen in Instagram posts in the first month following the first day of the pandemic. The

obtained datasets are available online to contribute to different sentiment analysis studies [24].

In this study, COVID-19-related Instagram comments posted between March 11 and April 10, 2020, were obtained, and the datasets of these comments were divided into three sentiment moods (i.e., positive, negative, neutral). The oversampling method was applied to these clusters, and two new balanced sets (i.e., resDataset1, resDataset2) were generated due to the imbalanced distribution of the datasets. In addition to various machine learning models (i.e., Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Gated Convolutional Recurrent-Neural Networks (GCR-NN)) for COVID-19 sensitivity analysis, the Bidirectional Encoder Representations from Transformers (BERT)-based transfer learning model has been used across all datasets (i.e., Dataset1, Dataset2, resDataset1, resDataset2). Models suggested for sentiment analysis were evaluated according to classification success and duration.

This study used word n-grams (i.e., unigram, bigram, trigram) for traditional machine learning (TML) models and one-hot encoding feature vectors for deep learning (DL) models. The recommended BERT-based transfer learning model provided a macro-averaged F1-score of over 70% across the four datasets. Considering subject inferences made using Latent Dirichlet Allocation (LDA) in datasets clarifies that there is no positive opinion transfer for users, and posts are repeated without discrimination regarding their importance despite the frequent use of social media.

The contributions of this study can be summarized as follows: (1) We presented new datasets that make it possible to identify Turkish sentiment analysis on COVID-19. (2) Most social media impact and sentiment analysis studies on COVID-19 have been conducted on Twitter. Our study uses Turkish datasets obtained from Instagram, which consists of comments on COVID-19. To the best of our knowledge, no existing study has been done related to COVID-19 on Turkish comments on the Instagram social network. (3) According to the results of the experiments (i.e., LDA, word frequency, p-values of the non-linear t-test) conducted to determine the role and contribution of social media in the pandemic, it has been observed that there is no intellectual interaction

between comments on COVID-19 on social media. (4) The proposed dataset and BERT-based transfer learning achieved better classification performance in terms of the macro-averaged F1-score than other models. It has been proven that using a pre-trained language model in Turkish datasets is more successful than other models (i.e., SVM, NB, RF, CNN, LSTM, GCR-NN) in terms of classification performance.

The remainder of the article is organized as follows: Section 2 presents the datasets and sentiment analysis studies used. Section 3 sets out the materials and methods. Section 4 contains the experimental results, and Section 5 presents a discussion and a variety of information for future research based on the results achieved in this study.

2. Related Works

Many studies have been carried out about COVID-19, and a diversity of approaches have been improved. The Twitter network has been frequently used in studies conducted to explore the relationship between social media and the pandemic with data science ([2]; [11]; [39]; [6]). Other studies have collected data from Reddit [18], Yelp.com [32], YouTube [23], and Weibo [29] pages. Some of them are summarized below.

In the study of [29], data obtained from the Weibo Chinese social network beginning January 20, 2020, was evaluated via t-test using the SPSS program and sentiment analysis studies. The results of this study may be biased because Weibo users are primarily young people. A Twitter study [1] collected 167,073 unique tweets from 160,829 unique users between February 2 and March 15, 2020. Analysis of this dataset indicated that social media offered people the opportunity to directly communicate health-related information to the public and that there is a need to prevent the spread of fake news. A sentiment analysis study [2] was conducted using the NB classifier in a dataset of Arabic tweets obtained during the epidemic. The results showed that users had positive thoughts that the COVID-19 epidemic would end and that users felt social media was used positively during the epidemic. Another study analyzed the fears among Twitter users in the USA during the COVID-19 pandemic [39]. Tweets obtained in the R program via Twitter Appli-

cation Programming Interface (API) were evaluated using NB and Logistic Regression (LR) machine learning algorithms. Accuracy classification success of 91% and 74% were achieved in short tweets.

A dataset containing the terms “prevencion coronavirus” and “prevencion COVID19” was collected from the comments section of 129 Spanish language videos on YouTube [23]. Classification results were evaluated using univariate analysis and the multiple logistic regression model. It was determined that the information in Spanish on preventing COVID-19 sourced from YouTube is primarily incomplete and inaccurate. A COVID-19 dataset [13] consisting of 410,643 tweets, including #IndiaLockdown and #IndiafightsCorona hashtags, was obtained between March 22 and April 21, 2020. Results showed that there are slightly more positive sentiments than negative sentiments related to the pandemic in India. A COVID-19 dataset [36] containing 3,377,295 tweets from users in the US was obtained between November 2019 and June 2020. The researchers concluded that the proportion of negative tweets referring to Asians increased by 68.4% and that the proportion of negative tweets referring to other racial or ethnic minorities remained stable. Common themes obtained by content analysis of 3,300 random tweet subsamples were racism-accusation, anti-racism, and effects on daily life.

Datasets obtained from various social media sources (i.e., YouTube, Reddit, Wikipedia, web news) between February 7 and May 15, 2020, were analyzed using the Latent Dirichlet Allocation model in the study of [18]. Compared to other platforms, Reddit users during the pandemic period proved to be more concerned with health, coronavirus information, and the necessary interventions to stop overexposure to the media. It was found that [10] the semantic-sentimental vocabulary of words, the sentiment curve, and the portrait of the patient seeking help were heterogeneous in the dataset obtained by collecting micro-blog data on the COVID-19 outbreak from Wuhan and Henan state regions. In another sentiment analysis study [11], two types of tweets were analyzed, proving that negative opinions on social media caused fear and panic in users. The highest classification success rate of 81% was achieved with the DL classifier. Also, the fuzzy rule base based on the Gaussian function was used to predict tweets’ sentiment inferences accurately, and this model improved the success rate to 79%.

To analyze perceptions of the Indian government’s pandemic policies, data labeled #IndiaLockdown and #IndiafightsKorona were collected between March 25 and 28, 2020, in the study of [6] via the R language Twitter API. Twenty-four thousand pieces of the data were analyzed with Word Cloud. It was concluded that users found the government’s policies regarding the pandemic positive. Although the researchers determined that users had feelings of negativity, fear, disgust, and sadness about staying at home during the pandemic, they observed that positive sentiments were more prominent [15]. Positive messages were categorized as “joy” and negative messages as “anger, sadness, fear” in an epidemic dataset obtained by collecting 3,332,565 tweets in English and 3,155,277 tweets in Portuguese. A dataset showed that most tweets about the epidemic had a negative slant in the study [20]. A series of tweets containing 8-scale sentiments were collected, and the tweets were labeled as anger, anticipation, disgust, fear, joy, sadness, surprise, and trust based on COVID-19 moods. In a COVID-19 dataset [7] obtained using publicly available data posted during the 2015 Nepal earthquake and the 2016 Italy earthquake and original COVID-19 Twitter data, 83% classification accuracy was achieved. A dataset of 112,412 reviews containing a comprehensive COVID-19 outbreak restaurant rating published from January through June 2020 on Yelp.com was obtained by [32]. The Bidirectional LSTM (BiLSTM) algorithm has proven effective in generating subtopics and predictions on user sentiments. In a study [35], the effectiveness of e-learning during the COVID-19 pandemic was analyzed using the sentiments of people. A Twitter dataset containing 17,155 tweets about e-learning was utilized. Machine learning models and DL models were performed to analyze the polarity and subjectivity score of tweets’ text. COVID-19 vaccination issue was discussed in another study [37] by analyzing the global perceptions and perspectives towards vaccination using a worldwide Twitter dataset. An ensemble model named LSTM-Gated Recurrent Neural Network (LSTM-GRNN) was used with different lexicon-based methods to perform sentiment analysis.

Table 1 summarizes previous approaches using different datasets and methods for sentiment analysis of the COVID-19 pandemic on the Twitter platform. Studies have shown a large number of sentiment analysis studies on COVID-19 in English as the lan-

guage and Twitter as the social network. Studies in other social networks and languages such as Turkish are less common.

Table 1

The previous sentiment analysis studies about COVID-19 pandemic on the Twitter platform

| Related Works | Year | Platform | Best Classifier | Best Result |
|---------------|----------------|----------|--|--------------|
| [2] | April 2020 | Twitter | NB | 84.57% Acc |
| [39] | June 2020 | Twitter | NB | 91% Acc |
| [36] | September 2020 | Twitter | SVM | 91% Acc |
| [11] | December 2020 | Twitter | Gaussian membership based fuzzy rule base system | 79% Acc |
| [24] | March 2021 | Twitter | LR | 87% F1-score |
| [7] | March 2021 | Twitter | CNN | 83% Acc |
| [32] | April 2021 | Yelp.com | BILSTM | 92% F1-score |
| [35] | September 2021 | Twitter | RF, DT | 95% Acc |
| [37] | February 2022 | Twitter | LSTM-GRNN | 95% Acc |

3. Material and Methods

3.1. Datasets

In this study, the first dataset (Dataset1) was obtained from the collection of comments made on the COVID-19 posts of the Turkish magazine Instagram account “2.SayfaOfficial” on March 11, 2020. The second dataset (Dataset2) was collected from the comments on COVID-19 posts shared on the same Instagram account between March 12, 2020, and April

10, 2020. The comments collected for these datasets were acquired through the Instagram API.

Table 2 shows the frequency of comments for each category of datasets. The total number of comments gathered was 22,878. After they were labeled and cleaned, 9,708 were selected for the study; of those, there were 2,745 “positive” comments, 3,037 “negative” comments, and 3,926 “neutral” comments. In annotating the datasets, it is important to provide the necessary information to the annotators and that the annotators are experts in the annotated field. Two experts annotated COVID-19 datasets with master’s degrees from computer science-related departments. The comments collected in the datasets were manually labeled as positive, negative, and neutral in expressing the sentiment at the sentence level [38]. No keyword-based annotation was used. Samples of sentiments such as anger, pessimism, ridicule, fear, hate, swearing, and violence about the COVID-19 pandemic were labeled as “1,” or “negative,” and the comments that contained optimistic and reassuring feelings about the disease were labeled “2,” that is, “positive.” Finally, comments with no stated sentiments or expectations were labeled “neutral,” that is, “0.”

The purpose of obtaining datasets on different dates is to better observe the changes in people’s sentiments and thoughts about COVID-19. Word frequencies of datasets were obtained using a Python program. Spelling checks of the interpretations were carried out according to Turkish grammar rules before the word frequencies were calculated.

A study was carried out in the pre-processing stage to remove the noise in the data. URL links, HTML

Table 2

Comparative values of both datasets

| Comments stance | Dataset1 | Dataset2 |
|-------------------------------------|----------|----------|
| Total number of comments | 4,875 | 18,003 |
| Number of comments labeled and used | 2,887 | 6,821 |
| Number of positive comments | 671 | 2,074 |
| Number of negative comments | 1,473 | 1,564 |
| Number of neutral comments | 743 | 3,183 |

tags, hashtags, emojis, symbols, punctuation marks, and stop-word (neutral) Turkish words have been removed from datasets. All of the words have been converted to lowercase letters. According to Turkish grammar rules, spelling errors were corrected, and repeated symbols were removed. The Turkish language is agglutinative, with each added suffix changing the word's meaning [3]. Therefore, it becomes difficult to reach the root meaning of the words. The stemming process was applied to the first 30 most frequently used words in the datasets in this study.

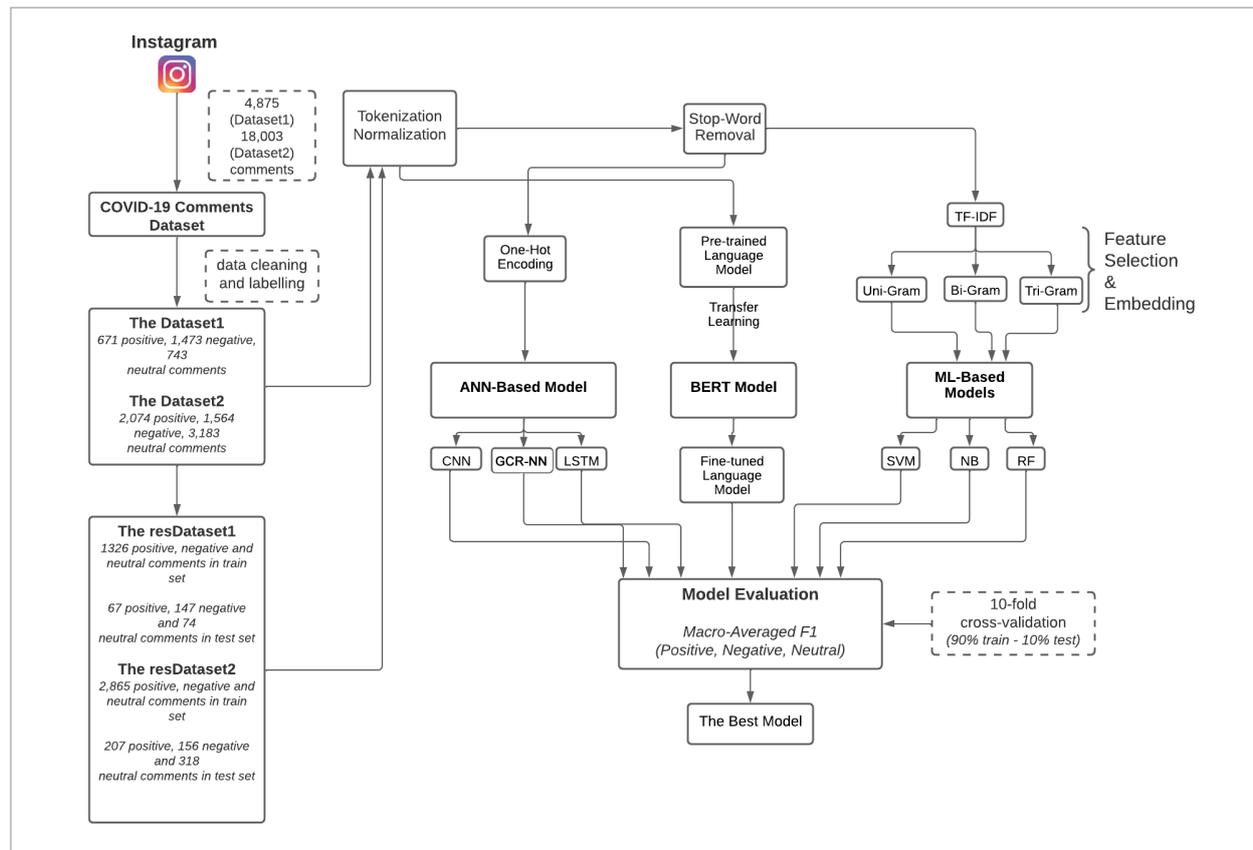
Dataset1 and Dataset2 were oversampled because the samples were unbalanced in the class distribution before the experimental studies. The oversampling method increases the number of minority class samples and equates them to the majority class [31]. Positive and neutral samples were increased randomly, equaling the number of negative samples to balance the class distribution of training data in Dataset1. Neg-

ative and neutral samples were increased randomly, equaling the number of neutral samples to balance the class distribution of training data in Dataset2.

Most of the comments labeled “neutral” in the first dataset included fake news and rumors about the person's location in the first-reported COVID-19 case. Another significant finding in the neutral comments was a repeated request that schools be closed for necessary precautions, primarily to protect children from the disease. The importance of washing hands and using alcohol-based disinfectants was also mentioned in neutral comments. Comments collected over the month after the first case was reported (March 12, 2020, to April 10, 2020), which form Dataset2, included more neutral and positive comments than Dataset1. This may indicate that people cast off fear, anxiety, and insecurity. It also shows trust in government and health officials during the pandemic [9].

Figure 1

The proposed architecture for COVID-19 sentiment analysis



3.2. Proposed Architecture

As shown in Figure 1, in this study, three models were used for COVID-19 sentiment analysis: TML models, DL models (i.e., CNN, LSTM, GCR-NN), and the BERT-based transfer learning classification model. In TML models, word n-grams (i.e., unigram, bigram, trigram) are used for feature extraction and Term Frequency-Inverse Document Frequency (TF-IDF) for feature weighting. The classifiers used in TML models are SVM, RF, and NB.

The one-hot encoding algorithm was used for feature extraction in the proposed DL models. The classifiers used in DL models were CNN, LSTM, and GCR-NN. The third proposed model, the BERT-based transfer learning classification model, used a pre-trained multi-language model of Wikipedia data in 104 languages. The models used and their contents are shown in Figure 1. Python was used as the programming language in all of the experimental studies.

3.2.1. Traditional Machine Learning Models

This section describes the feature extraction methods (i.e., unigram, bigram, trigram) used in TML models, the TF-IDF feature weighting method, and three TML algorithms. The word n-grams used for feature extraction (i.e., unigram, bigram, trigram) refer to the data's n consecutive strings of words. In n-grams, n represents the value by which word repetition is checked, and gram represents the weight of the word repetition in the sequence.

TF-IDF is a statistical evaluation measure that identifies the importance of words in a dataset by looking at the frequency of particular words. The TF-IDF value is also high if a word is used often in the target text but less often in other texts [22]. TF-IDF was used to generate the vector representation of the comments in the two datasets. Weights based on the frequency of n-grams in Dataset1 and Dataset2 were calculated with TF-IDF, and vector spaces belonging to the datasets were created.

The SVM classifier is a supervised machine learning algorithm frequently used in classification problems with successful results. The SVM algorithm uses a linear hyperplane to distinguish classes. It achieves linear separation with maximum marginal distance using support vectors in high dimensional space [22].

NB algorithm is a machine learning algorithm that performs sensitivity analysis based on conditional probability. Although it makes simple classifications,

it achieves excellent results in text classification. Unlike SVM in the NB algorithm, it does not require an input vector to perform sensitivity analysis. NB produces a classification result based on the conditional probability value of each feature. The most important advantage of the NB algorithm is that it reaches high classification accuracy values quickly.

The RF classifier is a supervised learning algorithm that creates multiple decision trees and provides a classification result by combining the values obtained from these decision trees by aggregating data. It searches for the best features in a random subset of features. The RF classifier can provide good classification results even without hyper-parameter optimization.

Table 3 shows the hyper-parameter values of TML algorithms (i.e., SVM, NB, and RF) as determined by 10-fold cross-validation with grid-search.

In this study, parameter values of SVM were defined as follows; the cost parameter (C)={0.1, 1, 2, 3, 4, 5, 10, 20, 30, 40, 50}. The SVM model gave the best results with a C=2 value in Dataset1 and a C=1 value in Dataset2. In the NB algorithm, the Multinomial NB used for multi-class categories was chosen, parameter values of NB were defined as follows; the Alpha parameter (C)={0.1, 0.2, 0.3, 0.4, 0.5, 0.6}. The NB model gave the best results with a C=0.1 value in Dataset1 and a C=0.5 value in Dataset2. The n_estimators hyper-parameter value of the RF algorithm, that is, the number of decision trees in the algorithm was 70 in all datasets.

3.2.2. Deep Learning Models

Preparing the categorical features was completed with the one-hot encoding method in the proposed

Table 3

Hyper-parameter values optimized by grid-search for TML algorithms

| Dataset | Algorithm | Hyper-parameters | Values explored |
|--------------------------|-----------|------------------|-----------------|
| Dataset1 and resDataset1 | SVM | C | 2 |
| | NB | alpha | 0.1 |
| | RF | n_estimators | 70 |
| Dataset2 and resDataset2 | SVM | C | 1 |
| | NB | alpha | 0.5 |
| | RF | n_estimators | 70 |

DL architectures. In the second step, classification results were evaluated in all datasets using DL classifiers (i.e., CNN, LSTM, GCR-NN).

One-hot encoding method enables the data to be represented as binary, making the data ready to classify. The values in the dataset are mapped to integer values for this operation. Then each integer value is marked as 1. All values except the integer index are marked as zero and represented as a binary vector [43].

In the past, the CNN classifier was used primarily to distinguish images. CNN is also frequently used in recommender systems and NLP areas [12]. CNN classifiers have embedded layers, convolution layers, pool layers, dense layers, and classification layers. The most crucial advantage of CNNs is that they assess clusters instead of examining words one by one. Thus, the meaning of a word is clarified by examining other words around it. The embedding layer prepares the input texts for classification. Input texts come in numbers, with each word represented by a unique number. The embedding layer starts with random weights and learns embedding for all words (input texts) in the training dataset. CNN uses the convolution layer to analyze the properties of the data [25]. The size of the input data is reduced at the pooling layer. The dense layer feeds the CNN neurons with all the output values from the previous layer, and each neuron provides an output value to the next layer. The classification layer is a dense layer where the classification result is produced.

The trial and error method was applied to determine the hyper-parameter values of each DL model (i.e., LSTM, CNN, GCR-NN) [16]. Table 4 shows the hyper-parameter values in all datasets (i.e., Dataset1, resDataset1, Dataset2, resDataset2) of the CNN classifier. The network structure for the CNN model was set as follows: Embedding Layer (unit=512)-CNN layer (Number of convolutional layers=1, Filter=1, Kernel=3, Activation function=ReLu)-Pooling layer-Dense layer (unit=512, Activation function=ReLu). CNN layers were tried as 1, 2, 3, respectively, and the CNN layer was chosen as one according to the classification success. The Filters value in the CNN layer was tried as 10, 50, 100, 150, respectively, and the Filters value was determined as 100 according to the classification success. Kernel_size value was tried as 3,4,5 respectively, and the value was determined as three according to classification success. Global-

Table 4

Hyper-parameters of CNN models for datasets

| Hyper-parameters | CNN models |
|---|-------------------------------|
| Number of output dimension in embedding layer | 512 |
| Number of convolutional layers | 1 |
| Filter | 100 |
| Kernel | 3 |
| Activation convolutional function | ReLU |
| Pooling | GlobalMax Pooling1D |
| Activation dense function | ReLU |
| Activation output function | softmax |
| Loss function | Categorical cross-entropy |
| Number of epoch | 15 |
| Optimizer | Adam |
| Value of fully connected units | 512 |
| Output shape | 3 |
| Number of batch size | 64 (Dataset1 and resDataset1) |
| | 32 (Dataset2 and resDataset2) |

MaxPooling1D subsamples the input representation by taking the maximum value over the time dimension [17]. GlobalMaxPooling1D was used because it is more efficient in the Pooling layer. Since the output value is three in the dense layer, softmax is used as the output function.

In this study, one embedding layer, one convolution layer (filter value 100, kernel size 3), a global max-pooling layer, and two dense output layers, the classification layer, were used in the CNN model. Output shape value was taken as three because three categorical data types (i.e., positive, negative, and neutral) in the last dense classification layer.

LSTM networks are a modified version of recurrent neural networks that enable previous data to be remembered. Gradient problems appearing in recurrent neural networks are solved with LSTM [8]. The LSTM model trains the data using backpropagation. Table 5 shows the hyper-parameter values in all datasets (i.e., Dataset1, resDataset1, Dataset2, resDataset2) of the LSTM classifier. The network structure for the LSTM model is set as follows: Embedding layer (unit=512), 1. LSTM layer (unit=256), 2. LSTM layer (unit=256), Dropout layer, Dense layer. LSTM model, different dropout values (i.e. 0.2, 0.3, 0.4, and 0.5) were tried and the optimum dropout value was found as 0.4 in the Dataset1, resDataset1. The optimum dropout value was found as 0.5 in the Dataset2, resDataset2. Likewise, the optimum recurrent dropout value was 0.4 in Dataset1, resDataset1. The optimum recurrent dropout value was used as 0.6 in Dataset1, resDataset1. The Adaptive Moment Estimation (Adam) optimizer was used in the LSTM model; the learning rate was 0.0001,

Table 5

Hyper-parameters of LSTM models for datasets

| Hyper-parameters | LSTM models |
|---|--------------------------------|
| Number of output dimension in embedding layer | 512 |
| Number of LSTM layers | 2 |
| Dropout | 0.4 (Dataset1 and resDataset1) |
| | 0.5 (Dataset2 and resDataset2) |
| Recurrent dropout | 0.4 (Dataset1 and resDataset1) |
| | 0.6 (Dataset2 and resDataset2) |
| Activation dense function | softmax |
| Loss function | categorical_crossentropy |
| Number of epoch | 10 |
| Optimizer | Adam (lr=0.0001) |
| Number of batch size | 64 |

and loss was categorical_crossentropy. During training, the batch size is 64; the number of epochs is 10.

The GCR-NN model is a combination version of the GRU, CNN, and RNN neural networks [27]. The GCR-NN model gives good classification results in sentiment analysis studies due to its stacked ensemble architecture [27].

Table 6 shows the hyper-parameter values in all datasets of the GCR-NN model. The network structure for the GCR-NN model is set as follows: Embedding layer (unit=512), 1. GRU layer (unit=64), 2. CNN layer

Table 6

Hyper-parameters of GCR-NN model for datasets

| Hyper-parameters | GCR-NN model |
|---|----------------------------|
| Number of output dimension in embedding layer | 512 |
| Number of GRU layers | 1 |
| GRU unit | 64 |
| Number of convolutional layers | 1 |
| Filter | 64 |
| Kernel | 4 |
| Activation convolutional function | ReLU |
| Pooling | MaxPooling1D (pool_size=4) |
| Dropout | 0.2 |
| Number of SimpleRNN layers | 1 |
| SimpleRNN unit | 16 |
| Value of fully connected units | 16 |
| Output shape | 3 |
| Activation output function | softmax |
| Loss function | categorical_crossentropy |
| Optimizer | Adam |
| Number of epoch | 100 (with Early_stopping) |
| Number of batch size | 16 |

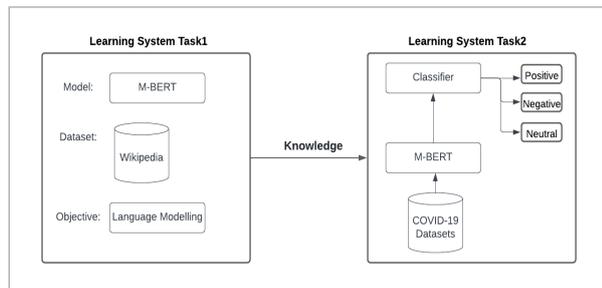
(filter=64, kernel=4), 3. MaxPooling1D layer (pool_size=4), 4. Dropout layer(0.2), 5. SimpleRNN layer (unit=16) -Dense layer(Unit=16), Dense Layer (output shape=3). Adam optimizer was used in the GCRNN model, and loss was categorical_crossentropy. During training, the batch size is 16; the number of epochs is 100 (with early stopping).

3.2.3. BERT-based Transfer Learning Model

Google has developed a BERT technique. This transformer trains general-purpose language representation models using large amounts of unlabeled text in a process known as pre-training.

Figure 2

The architecture of the BERT-based transfer learning model



Efforts to model and use a pre-trained system have gained momentum, with BERT as a replacement for developing different models in language tasks. As seen in Figure 2, the BERT pre-trained language model is trained with a considerable amount of data, allowing this model to be transferred to other smaller language processing tasks. The pre-training of the BERT model is very resource-intensive, but this task has already been done in the pre-trained language model. Researchers may fine-tune the process for different language tasks by adding the output layer. Since this situation requires significantly fewer resources for training, training time is minimized [14].

Multilingual BERT (MBERT) is based on the WordPiece dictionary, trained in 104 languages, and shares 110K of data with Wikipedia. This study contributes to the transfer of syntactic knowledge between languages by confirming that syntactic dependency relationships learned in a language are maintained in other languages with the MBERT language model [19]. A

MBERT model has data in many major languages and has sufficient data in common languages. MBERT with large data sources can be easily used to classify datasets in other languages [45]. The “bert-base-multilingual-cased” used consisted of 12 layers with 768 hidden units each and 12 attention heads. The “bert-base-multilingual-cased” model was selected from the Pytorch Huggingface transformer library, and textual data were fine-tuned.

In this study, the pre-trained model was used in Python with the *ktrain* library [33]. *ktrain* is a package that helps build, train, and deploy neural networks and other machine learning models for the DL libraries TensorFlow, Keras, and others. It has a simple unified interface that attempts to solve a wide variety of tasks with very few lines of code. *ktrain* enables easy completion of steps such as creating a model for training, examining a model, and training a model. Language and character encoding are automatically detected, and other processes continue accordingly in *ktrain*. Whether the targets are numerical or categorical is analyzed automatically. The model is configured optimally.

The pre-trained MBERT model is loaded with a randomly initiated final dense layer. There is no update during the training process, although the last dense layer was initiated randomly. Weights are updated using a newly labeled dataset to analyze contextual information extracted from the pre-trained layers of the MBERT. Since no layers are stopped, and all model layers are trainable, their weights are updated on backpropagation. The MBERT model is trained for five terms with the “fit_onecycle” method within the *ktrain* structure. The default configuration of *ktrain* was used with a $2e-5$ learning speed for five periods with categorical cross-entropy loss function and softmax prediction layer.

4. Experimental Results

In this section, the classification results and durations of the models (i.e., TML models, DL models, BERT-based transfer learning model) proposed for COVID-19 sentiment analysis were evaluated. In the final part of this section, the social media relationship between the two existing datasets (i.e., Dataset1, Dataset2) was compared with LDA and the p-values.

The Google Colab machine was used to run experimental applications in this study. Features of the Google Colab machine are Tesla K80 GPU, 12 GB RAM, and Intel Xeon CPU 2.20 GHz. The macro-averaged F1-score was used to compare the performance of the proposed models.

The macro-averaged F1-score calculates the average values obtained independently for each category (i.e., positive, negative, and neutral) [40]. Generally, the macro-averaged F1 metric is preferred for class imbalance problems. The macro-averaged F1-score performs well even in the uneven distribution of datasets. The four datasets were divided into ten parts by k-fold cross-validation method as training and test, with nine parts used for training and one for testing. The performance of the models was determined by taking the classification results of all parts and the average of the test times.

4.1. Experimental Results of TML, DL, and BERT Models

Tables 7-8 show the macro-averaged F1-score results and total training-test times evaluated using different feature extraction and classifiers.

Compared with TML models (i.e., SVM, NB, RF), CNN, LSTM, and GCR-NN models performed similarly across all versions of the first dataset (i.e., Dataset1, resDataset1). However, the one-hot encoding + LSTM model combination achieved the highest macro-averaged F1-score (0.717 in Dataset1) among TML and DL models. Also, the SVM classifier has the second-lowest classification time after the NB classifier and has the best classification macro-averaged F1-score result (0.719) in resDataset1 among the TML

Table 7

Macro-averaged F1-scores of all models for separate classes using 10-fold cross-validation in the first dataset

| Models | Feature extraction | Classifier | Positive | Negative | Neutral | F1-score | Time (sec.) |
|--|--------------------|------------|----------|----------|---------|--------------|-------------|
| Traditional machine learning models + Dataset1 | unigram+ TF-IDF | SVM | 0.672 | 0.780 | 0.650 | 0.701 | 0.155 |
| | bigram+ TF-IDF | | 0.686 | 0.783 | 0.632 | 0.700 | 0.308 |
| | trigram+ TF-IDF | | 0.686 | 0.779 | 0.610 | 0.692 | 0.470 |
| | unigram+ TF-IDF | NB | 0.658 | 0.775 | 0.615 | 0.683 | 0.104 |
| | bigram+ TF-IDF | | 0.689 | 0.786 | 0.613 | 0.696 | 0.282 |
| | trigram+ TF-IDF | | 0.680 | 0.783 | 0.609 | 0.691 | 0.366 |
| | unigram+ TF-IDF | RF | 0.658 | 0.764 | 0.656 | 0.693 | 2.949 |
| | bigram+ TF-IDF | | 0.651 | 0.752 | 0.661 | 0.688 | 6.212 |
| | trigram+ TF-IDF | | 0.645 | 0.742 | 0.658 | 0.682 | 9.774 |
| Deep learning models + Dataset1 | one-hot encoding | LSTM | 0.676 | 0.809 | 0.667 | 0.717 | 283.31 |
| | one-hot encoding | CNN | 0.645 | 0.785 | 0.639 | 0.69 | 90.121 |
| | one-hot encoding | GCR-NN | 0.606 | 0.743 | 0.624 | 0.658 | 530.21 |

| Models | Feature extraction | Classifier | Positive | Negative | Neutral | F1-score | Time (sec.) |
|---|--------------------------|------------|----------|----------|---------|--------------|-------------|
| BERT-based transfer learning model + Dataset1 | BERT text classification | | 0.665 | 0.756 | 0.806 | 0.742 | 35998.12 |
| Traditional machine learning models + resDataset1 | unigram+ TF-IDF | SVM | 0.671 | 0.771 | 0.654 | 0.699 | 0.266 |
| | bigram+ TF-IDF | | 0.685 | 0.787 | 0.674 | 0.715 | 0.414 |
| | trigram+ TF-IDF | | 0.694 | 0.788 | 0.675 | 0.719 | 0.588 |
| | unigram+ TF-IDF | NB | 0.662 | 0.773 | 0.674 | 0.703 | 0.138 |
| | bigram+ TF-IDF | | 0.682 | 0.773 | 0.684 | 0.713 | 0.259 |
| | trigram+ TF-IDF | | 0.682 | 0.771 | 0.683 | 0.712 | 0.411 |
| | unigram+ TF-IDF | RF | 0.670 | 0.737 | 0.658 | 0.688 | 3.090 |
| | bigram+ TF-IDF | | 0.671 | 0.723 | 0.650 | 0.681 | 6.296 |
| | trigram+ TF-IDF | | 0.663 | 0.726 | 0.652 | 0.680 | 9.600 |
| Deep learning models + resDataset1 | one-hot encoding | LSTM | 0.698 | 0.774 | 0.651 | 0.708 | 459.705 |
| | one-hot encoding | CNN | 0.654 | 0.762 | 0.637 | 0.684 | 145.799 |
| | one-hot encoding | GCR-NN | 0.617 | 0.768 | 0.645 | 0.677 | 700.854 |
| BERT-based transfer learning model + resDataset1 | BERT text classification | | 0.711 | 0.803 | 0.846 | 0.786 | 55100.69 |

and DL models. The SVM classifier is effective in solving text classification problems [26].

In the experiments for all versions of the first dataset, the BERT-based transfer learning model gave the best results (0.742 in Dataset1 and 0.786 macro-averaged F1-score in resDataset1). However, it is seen that the BERT model has a very long training-test period. The evaluation results of the second dataset for separate classes (i.e., positive, negative, neutral) were given in Table 8. DL models (i.e., CNN, LSTM, GCR-NN) achieved lower performance results than TML classifiers in both versions of the resDataset2.

Regarding the classification times of the second dataset versions (i.e., Dataset2, resDataset2), it is seen that the best classification time was the value of 0.244 seconds achieved by the unigram + TF-IDF + NB model. The NB classifier had the best classification time compared to other classifiers in all datasets.

The COVID-19 sentiment classification model that had the best classification result among all models was the BERT-based transfer learning model, with a macro-averaged F1-score value of 0.712 in the two versions of the second dataset.

Table 8

Macro-averaged F1-scores of all models for separate classes using 10-fold cross-validation in the second dataset

| Models | Feature extraction | Classifier | Positive | Negative | Neutral | F1-score | Time (sec.) |
|---|--------------------------|------------|----------|----------|---------|--------------|-------------|
| Traditional machine learning models + Dataset2 | unigram+ TF-IDF | SVM | 0.710 | 0.483 | 0.651 | 0.614 | 0.400 |
| | bigram+ TF-IDF | | 0.713 | 0.507 | 0.669 | 0.63 | 0.777 |
| | trigram+ TF-IDF | | 0.710 | 0.521 | 0.674 | 0.635 | 1.239 |
| | unigram+ TF-IDF | NB | 0.707 | 0.447 | 0.669 | 0.608 | 0.244 |
| | bigram+ TF-IDF | | 0.711 | 0.463 | 0.678 | 0.617 | 0.615 |
| | trigram+ TF-IDF | | 0.707 | 0.443 | 0.670 | 0.607 | 1.001 |
| | unigram+ TF-IDF | RF | 0.695 | 0.393 | 0.655 | 0.581 | 9.232 |
| | bigram+ TF-IDF | | 0.698 | 0.387 | 0.655 | 0.580 | 21.766 |
| | trigram+ TF-IDF | | 0.704 | 0.376 | 0.642 | 0.574 | 43.166 |
| Deep learning models + Dataset2 | one-hot encoding | LSTM | 0.656 | 0.496 | 0.648 | 0.600 | 530.307 |
| | one-hot encoding | CNN | 0.662 | 0.525 | 0.644 | 0.610 | 128.184 |
| | one-hot encoding | GCR-NN | 0.709 | 0.414 | 0.574 | 0.566 | 790.154 |
| BERT-based transfer learning model + Dataset2 | BERT text classification | | 0.782 | 0.578 | 0.772 | 0.711 | 86098.73 |
| Traditional machine learning models + resDataset2 | unigram+ TF-IDF | SVM | 0.641 | 0.496 | 0.705 | 0.614 | 0.466 |
| | bigram+ TF-IDF | | 0.661 | 0.515 | 0.708 | 0.628 | 0.991 |
| | trigram+ TF-IDF | | 0.673 | 0.526 | 0.709 | 0.636 | 1.460 |
| | unigram+ TF-IDF | NB | 0.654 | 0.514 | 0.693 | 0.620 | 0.330 |
| | bigram+ TF-IDF | | 0.655 | 0.512 | 0.691 | 0.619 | 0.749 |
| | trigram+ TF-IDF | | 0.666 | 0.513 | 0.689 | 0.623 | 1.140 |

| Models | Feature extraction | Classifier | Positive | Negative | Neutral | F1-score | Time (sec.) |
|---|--------------------------|------------|----------|----------|---------|--------------|-------------|
| Traditional machine learning models + resDataset2 | unigram+ TF-IDF | RF | 0.646 | 0.483 | 0.705 | 0.611 | 9.197 |
| | bigram+ TF-IDF | | 0.655 | 0.468 | 0.721 | 0.615 | 22.128 |
| | trigram+ TF-IDF | | 0.647 | 0.465 | 0.699 | 0.604 | 39.284 |
| Deep learning models + resDataset2 | one-hot encoding | LSTM | 0.675 | 0.514 | 0.604 | 0.598 | 783.155 |
| | one-hot encoding | CNN | 0.663 | 0.479 | 0.606 | 0.583 | 236.022 |
| | one-hot encoding | GCR-NN | 0.697 | 0.399 | 0.576 | 0.557 | 1033,014 |
| BERT-based transfer learning model + resDataset2 | BERT text classification | | 0.792 | 0.57 | 0.774 | 0.712 | 120039.463 |

As explained before, the classes (i.e., positive, negative, neutral) in Dataset1 and Dataset2 have an uneven distribution. The classification results reveal that the performance of each class (i.e., positive, negative, neutral) depends on the number of samples in the training set. Therefore, classes with more training samples (i.e., resDataset1, resDataset2) generally achieved better results than those with fewer samples in the training set (i.e., Dataset1, Dataset2). As seen in Tables 7-8, the BERT model gives better results than all models (i.e., TML models, DL models). Despite the success of the BERT model in classification results, the classification times are very high in all datasets.

4.2. Sentiment Analysis with Topic Extraction Using Latent Dirichlet Allocation and p-values

Topic inference in COVID-19 datasets was analyzed using the LDA method. LDA method is a statistical unsupervised learning model that can identify topics from texts given in a dataset [44]. The sentiment changes of COVID-19 datasets were investigated over time and the highlights topics were analyzed in terms of their importance [28]. In order to add meaning to the sentiment analysis in the comments, topic extraction was performed. Current topics in the datasets are classified together with the topic inference process.

The `n_components` parameter was set to three for extracting the three most discussed topics in all labels (i.e., positive, negative, and neutral). Table 9 shows LDA topics inferences and contents from Dataset1, which consists of comments labeled as positive, neutral, and negative. The translation of the topics in English is given in parenthesis.

When the topics are examined, there are more “negative” comments in Dataset1 as compared to Dataset2 considering class distribution (i.e., negative, positive, neutral). It can be said that people expressed more feelings of fear, anger, violence, disgust, sadness, and ridicule on the day Dataset1 was collected. In the comments labeled as “negative”, the virus was seen as a punishment from Allah (God), or Allah (God) was asked to punish those who transmit the virus. It is observed that in most of the “positive” comments, positive prayers and good intentions are made to Allah (God) for the end of the epidemic. In Dataset1, most of the comments labeled “neutral” included fake news and rumors about the location of the person who may have been the first person to have contracted COVID-19 in Turkey. Another significant finding in the neutral comments was the request that the necessary precautions be taken for schools to be closed and, especially, to protect children from the disease repeatedly. The importance of washing hands and us-

ing alcohol-based disinfectants was also mentioned in neutral comments.

Comments collected over the month after the first case was reported (from March 12, 2020, to April 10, 2020), which form Dataset2, included more neutral and positive comments than Dataset1. This may indicate that people cast off fear, anxiety, and insecurity during that time. It also shows trust in government and health officials during the pandemic [9]. Table 10 shows LDA topics inferences and contents obtained from Dataset2. As is seen in the results, Topic 1, Top-

ic 2, and Topic 3, which were taken from the positive comments, are about the importance of praying to God, the benefits of the bans, and opinions in favor of curfew. On the contrary, topic extraction from negative comments shows that the topics with negative comments are mostly related to people's anxiety and fear of death towards COVID-19 disease, as shown in Topic 1, Topic 2, and Topic 3 of negative comments. The word "Allah (God)" was frequently used in positive and negative comments in Dataset1 and Dataset2.

Table 9

Topic extracted from positive, negative and neutral comments in Dataset1

| Dataset1 | Topics | Topic contents |
|----------------|---------|---|
| Positive label | Topic 1 | 0.015**olsun (be)" + 0.010**geç (late)" + 0.010**kişi (person)" + 0.009**bol (plenty)" + 0.009**virus (virus)" + 0.009**öl (die)" + 0.007**rabbim (my god)" + .007**önlem (precaution)" + 0.006**yok (none)" + 0.006**insan (human)" |
| | Topic 2 | 0.009**öl (die)" + 0.009**virus (virus)" + 0.009**kork (fear)" + 0.009**allah (god)" + 0.008**ülke (country)" + 0.008**yok (none)" + 0.007**insan (human)" + 0.005**panic (panic)" + 0.005**tedbir (precaution)" + 0.005**bence (to my opinion)" |
| | Topic 3 | 0.058**allah (god)" + 0.053**okul (school)" + 0.009**virüs (virus)" + 0.007**ülke (country)" + 0.007**korusun (bless)" + 0.006**inşallah (god willing)" + 0.006**rabbim (my god)" + 0.006**temiz (clean)" + 0.005**il (city)" + 0.005**çocuk (child)" |
| Negative label | Topic 1 | 0.016**allah (god)" + 0.013**yurt (homeland)" + 0.009**yok (none)" + 0.008**okul (school)" + 0.008**ülke (country)" + 0.007**kork (fear)" + 0.007**dışına (outside)" + 0.007**insan (human)" + 0.006**diyor (says)" + 0.006**oldu (was)" |
| | Topic 2 | 0.011**virüs (virus)" + 0.010**giriş (entrance)" + 0.010**ülke (country)" + 0.009**çıkış (exit)" + 0.008**öl (die)" + 0.008**türkiye (turkey country)" + 0.008**yurt (homeland)" + 0.007**insan (human)" + 0.006**yurtdışı (abroad)" + 0.006**önlem (precaution)" |
| | Topic 3 | 0.012**virüs (virus)" + 0.009**ülke (country)" + 0.009**korona (corona)" + 0.009**il (city)" + 0.008**geç (late)" + 0.008**olsun (be)" + 0.008**çin (china)" + 0.007**kişi (person)" + 0.006**eyvah (oops)" + 0.006**allah (god)" |
| Neutral label | Topic 1 | 0.023**il (city)" + 0.010**şehir (city)" + 0.009**acaba (wonder)" + 0.008**diyor (says)" + 0.007**insan (human)" + 0.006**hastane (hospital)" + 0.006**kişi (person)" + 0.005**virüs (virus)" + 0.005**amin (amen)" + 0.005**bence (to my opinion)" |
| | Topic 2 | 0.018**okul (school)" + 0.014**şehir (city)" + 0.010**virüs (virus)" + 0.008**yok (none)" + 0.007**allah (god)" + 0.006**geç (late)" + 0.005**il (city)" + 0.005**vaka (case)" + 0.004**haber (news)" + 0.004**peki (alright)" |
| | Topic 3 | 0.018**istanbul (a turkish city)" + 0.010**cnn (news channel)" + 0.009**gün (day)" + 0.007**kişi (person)" + 0.007**kuralı (rule)" + 0.007**bakan (minister)" + 0.006**hasta (patient)" + 0.005**virüs (virus)" + 0.005**evet (yes)" |

Table 10

Topic extracted from positive, negative and neutral comments in Dataset2

| Dataset2 | Topics | Topic contents |
|----------------|---------|---|
| Positive label | Topic 1 | 0.015*"yasak (ban)" + 0.013*"inşallah (god willing)" + 0.012*"çok (lots)" + 0.010*"iyi (good)" + 0.009*"öl (die)" + 0.009*"şükür" + 0.008*"gün (day)" + 0.008*"güzel" + 0.008*"rabbim (my god)" + 0.008*"oldu (was)" |
| | Topic 2 | 0.017*"hafta (week)" + 0.014*"inşallah (god willing)" + 0.011*"gün (day)" + 0.010*"olsun (be)" + 0.009*"sokağa (to the street)" + 0.009*"yasak (ban)" + 0.008*"karar (decision)" + 0.008*"allah (god)" + 0.007*"çıkma (to go out)" + 0.007*"bakan (minister)" |
| | Topic 3 | 0.018*"yasak (ban)" + 0.015*"çıkma (to go out)" + 0.015*"allah (god)" + 0.015*"sokağa (to the street)" + 0.013*"geç (late)" + 0.011*"gelsin (come)" + 0.009*"olsun (be)" + 0.009*"bile (even)" + 0.009*"insan (human)" + 0.009*"sonunda (finally)" |
| Negative label | Topic 1 | 0.013*"insan (human)" + 0.010*"öl (die)" + 0.010*"diyor (says)" + 0.009*"allah (god)" + 0.007*"millet (nation)" + 0.006*"yasak (ban)" + 0.005*"kişi (person)" + 0.005*"sokağa (to the street)" + 0.005*"evde (at home)" + 0.005*"çalışan (worker)" |
| | Topic 2 | 0.021*"yasak (ban)" + 0.017*"sokağa (to the street)" + 0.016*"çıkma (to go out)" + 0.010*"öl (die)" + 0.007*"gün (day)" + 0.006*"evde (at home)" + 0.006*"kişi (person)" + 0.006*"diyor (says)" + 0.006*"insan (human)" + 0.005*"yok (none)" |
| | Topic 3 | 0.009*"insan (human)" + 0.009*"virüs (virus)" + 0.008*"evde (at home)" + 0.007*"geç (late)" + 0.007*"diyor (says)" + 0.005*"yok (none)" + 0.005*"öl (die)" + 0.005*"yasak (ban)" + 0.004*"gün (day)" + 0.004*"kork (fear)" |
| Neutral label | Topic 1 | 0.006*"allah (god)" + 0.006*"insan (human)" + 0.006*"yardım" + 0.005*"yasak (ban)" + 0.005*"borç (debt)" + 0.004*"olsun (be)" + 0.004*"amin (amen)" + 0.004*"gün (day)" + 0.004*"aynen (exactly)" + 0.004*"git (go)" |
| | Topic 2 | 0.009*"insan (human)" + 0.007*"öl (die)" + 0.007*"yok (none)" + 0.005*"virüs (virus)" + 0.005*"diyor (says)" + 0.005*"çalışan (worker)" + 0.004*"olan (the one-sick one)" + 0.004*"haber (news)" + 0.004*"evde (at home)" + 0.004*"kişi (person)" |
| | Topic 3 | 0.009*"inşallah (god willing)" + 0.009*"diyor (says)" + 0.008*"öl (die)" + 0.006*"yasak (ban)" + 0.006*"evde (at home)" + 0.006*"genç (young)" + 0.005*"virüs (virus)" + 0.005*"yok (none)" + 0.004*"yaşlı" + 0.004*"millet (nation)" |

The Mann-Whitney non-parametric t-test was used. The p-values belonging to each dataset were computed to determine whether there is any meaningful difference between categories (i.e., positive, negative) in datasets (i.e., Dataset1, Dataset2) [34]. Mann-Whitney non-parametric t-test p-values of Dataset1 and Dataset2 are 0.0012 and 0.0011, respectively. Since these values are $p < 0.05$, there are differences between negative comments and positive comments. The results prove no meaningful similarity between

positive and negative comments on Instagram about the COVID-19 outbreak. It has been proven by the p-values that there is no beneficial interaction in terms of supporting the epidemic process positively or preventing outbreaks in Turkey on Instagram.

The results regarding the performance comparison of the classification models and the effect of Turkish social media on the epidemic are given below:

- It is seen that the resampling method applied to Dataset1 obtained gives better results in macro-

averaged F1 metrics for SVM, NB, GCR-NN and BERT models. It can be said that resampling results positively in SVM, NB, RF and BERT models for Dataset2.

- The BERT-based transfer learning model (the recommended model) showed the best performance among all models. The highest classification results among the models were the 0.7864 macro-averaged F1-score values achieved by the BERT model in the resDataset1 dataset and the 0.7120 values in the resDataset2. As the BERT model contains millions of pre-trained data, classification results were achieved with this model without training in the four datasets. Thus, the model provided the opportunity to test datasets without training an abundance of data. The classification success of datasets with the BERT model has increased significantly compared to other models.
- Although the BERT model provided better results than other models, the classification time was very long in all datasets. The NB classifier has been the classifier with the shortest classification time in all datasets.
- CNN, LSTM, and GCR-NN models did not outperform better than TML models (i.e., SVM, NB, RF) for both datasets. Although datasets were oversampled, it is thought that more data is needed for DL architectures.
- The SVM classifier has the second-best performance results after BERT-based transfer learning. Also, the SVM classifier has the second-best classification time in all datasets. The SVM model is successful for our study when we consider it with the classification time.
- As a result of making LDA topic extractions in datasets, it is seen that the excess of negative comments (i.e., fear, panic, and anger) at the beginning of the epidemic in Dataset1 turns into positive and neutral comments in Dataset2. It is understood from the increase of positive and neutral comments in Dataset2 that people's feelings were positive or undecided in the first month of the epidemic (March 11 to April 10, 2020). According to the topics obtained with LDA,

it was observed that the number of rumors and fake news was high in both data sets, especially in neutral comments.

- Although there are many similar comments in the datasets, there are no directive and informative statements about the epidemic. In addition, it was seen that Instagram social network did not provide information or awareness-raising about the pandemic since the non-parametric t-test p-value in both datasets (i.e., Dataset1, Dataset2) was less than 0.05.

5. Conclusions

Sentiments, expressions, and opinions in datasets from social media can improve the outcomes of the COVID-19 pandemic and provide the information with the public needs. The data collected on social media is an important source for sentiment analysis of people for this reason. As far as we know, this study is the first to examine the role of social media in the COVID-19 outbreak and perform a COVID-19 sentiment analysis using the Turkish Instagram social network.

First, Instagram comments (i.e., Dataset1, Dataset2) belonging to an account were collected on specific dates at the beginning of the epidemic period. Analysis of datasets (i.e., LDA, word frequency, non-parametric t-test) was conducted to determine the role and impact of social media in the COVID-19 outbreak. Since the obtained datasets (i.e., Dataset1, Dataset2) were imbalanced, two new datasets (i.e., resDataset1, resDataset2) were acquired by oversampling both datasets for COVID-19 sentiment analysis. The classification test of sentiment classification models was applied for six different situations (i.e., SVM, NB, RF, CNN, LSTM, GCR-NN, and BERT-based transfer learning).

By increasing the data in datasets and using word-embedding methods such as Word2vec, Glove, and BERT to determine the hyper-parameter values for DL models, the classification accuracy values in such models can be increased. Consequently, the MBERT-based transfer learning model is an ideal classification model that can be used in various Turkish sentiment analysis studies.

References

1. Abd-Alrazaq, A., Alhuwail, D., Househ, M., Hai, M., Shah, Z. Top Concerns of Tweeters During the COVID-19 Pandemic: A Surveillance Study. *Journal of Medical Internet Research*, 2020, 22, e19016. <https://doi.org/10.2196/19016>
2. Alhajji, M., Al Khalifah, A., Aljubran, M., Alkhalifah, M. Sentiment Analysis of Tweets in Saudi Arabia Regarding Governmental Preventive Measures to Contain COVID-19. *Preprints 2020*, 2020040031. <https://doi.org/10.20944/preprints202004.0031.v1>
3. Alparslan, E., Karahoca, A., Bahşi, H. Security-level Classification for Confidential Documents by Using Adaptive Neuro-Fuzzy Inference Systems. *Expert Systems*, 2013, 30, 233-242. <https://doi.org/10.1111/j.1468-0394.2012.00634.x>
4. Apuke, O.D., Omar, B. Fake News and COVID-19: Modelling the Predictors of Fake News Sharing Among Social Media Users. *Telematics and Informatics*, 2021, 56, 101475. <https://doi.org/10.1016/j.tele.2020.101475>
5. Banea, C., Mihalcea, R., Wiebe, J. Sense-Level Subjectivity in a Multilingual Setting. *Computer Speech and Language*, 2014, 7-19. <https://doi.org/10.1016/j.csl.2013.03.002>
6. Barkur, G., Vibha, Kamath, G. B. Sentiment Analysis of Nationwide Lockdown due to COVID 19 Outbreak: Evidence from India. *Asian Journal of Psychiatry*, 2020, 51. <https://doi.org/10.1016/j.ajp.2020.102089>
7. Behl, S., Rao, A., Aggarwal, S., Chadha, S., Pannu, H. S. Twitter for Disaster Relief Through Sentiment Analysis for COVID-19 and Natural Hazard Crises. *International Journal on Disaster Risk Reduction*, 2021, 55, 102101. <https://doi.org/10.1016/j.ijdr.2021.102101>
8. Bengio, Y., Simard, P., Frasconi, P. Learning Long-Term Dependencies with Gradient Descent is Difficult. *IEEE Transactions on Neural Networks*, 1994, 5, 157-166. <https://doi.org/10.1109/72.279181>
9. Bhat, M., Qadri, M., ul A. Beg, N., Kundroo, M., Ahanger, N., Agarwal, B. Sentiment Analysis of Social Media Response on the Covid19 Outbreak. *Brain, Behavior, and Immunity*, 2020, 87, 136-137. <https://doi.org/10.1016/j.bbi.2020.05.006>
10. Boon-Itt, S., Skunkan, Y. Public Perception of the COVID-19 Pandemic on Twitter: Sentiment Analysis and Topic Modeling Study. *JMIR Public Health Surveillance*, 2020, 6. <https://doi.org/10.2196/21978>
11. Chakraborty, K., Bhatia, S., Bhattacharyya, S., Platos, J., Bag, R., Hassanien, A. E. Sentiment Analysis of COVID-19 Tweets by Deep Learning Classifiers-A Study to Show how Popularity is Affecting accuracy in Social Media. *Applied Soft Computing*, 2020, 97, 106754. <https://doi.org/10.1016/j.asoc.2020.106754>
12. Dang, N. C., Moreno-García, M. N., De la Prieta, F. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*, 2020, 9. <https://doi.org/10.3390/electronics9030483>
13. Das, S., Dutta, A. Characterizing Public Emotions and Sentiments in COVID-19 Environment: A Case Study of India. *Journal of Human Behavior in the Social Environment*, 2021, 31, 154-167. <https://doi.org/10.1080/10911359.2020.1781015>
14. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv preprint*, 2018, 4171-4186. <http://arxiv.org/abs/1810.04805>
15. Garcia, K., Berton, L. Topic Detection and Sentiment Analysis in Twitter Content Related to COVID-19 from Brazil and the USA. *Applied Soft Computing*, 2021, 101, 107057. <https://doi.org/10.1016/j.asoc.2020.107057>
16. García-Díaz, J.A., Cánovas-García, M., Valencia-García, R. Ontology-driven Aspect-based Sentiment Analysis Classification: An Infodemiological Case Study Regarding Infectious Diseases in Latin America. *Future Generation Computer Systems*, 2020, 112, 641-657. <https://doi.org/10.1016/j.future.2020.06.019>
17. GlobalMaxPooling1D layer, (n.d.). https://keras.io/api/layers/pooling_layers/global_max_pooling1d/ (accessed February 26, 2022).
18. Gozzi, N., Tizzani, M., Starnini, M., Ciulla, F., Paolotti, D., Panisson, A., Perra, N. Collective Response to Media Coverage of the COVID-19 Pandemic on Reddit and Wikipedia: Mixed-Methods Analysis. *Journal of Medical Internet Research*, 2020, 22, e21597. <https://doi.org/10.2196/21597>
19. Guarasci, R., Silvestri, S., De Pietro, G., Fujita, H., Esposito, M. BERT Syntactic Transfer: A Computational Experiment on Italian, French and English languages. *Computer Speech & Language*, 2022, 71, 101261. <https://doi.org/10.1016/j.csl.2021.101261>
20. Gupta, V., Jain, N., Katariya, P., Kumar, A., Mohan, S., Ahmadian, A., Ferrara, M. An Emotion Care Model using Multimodal Textual Analysis on COVID-19. *Chaos*,

- Solitons and Fractals, 2021, 144, 110708. <https://doi.org/10.1016/j.chaos.2021.110708>
21. Haberturk, Canlı Koronavirüs Haritası, Koronavirüs tablosu ve Haberleri, (2021). <https://www.haberturk.com/corona-virusu> (accessed March 23, 2021).
 22. Hernández-Castañeda, Á., García-Hernández, R. A., Ledeneva, Y., Millán-Hernández, C. E. Language-independent Extractive Automatic Text Summarization Based on Automatic Keyword Extraction. *Computer Speech & Language*, 2022, 71, 101267. <https://doi.org/10.1016/j.csl.2021.101267>
 23. Hernández-García, I., Giménez-Júlvez, T. Characteristics of Youtube Videos in Spanish on how to Prevent COVID-19. *International Journal of Environmental Research and Public Health*, 2020, 17, 1-10. <https://doi.org/10.3390/ijerph17134671>
 24. Karayigit, H., Inan Aci, Ç., Akdagli, A. COVID-19 Datasets, 2021. <https://www.kaggle.com/habibekarayigit/dataset1-11march2020> (accessed October 8, 2021).
 25. Kishore Kumar, R., Sreenivasa Rao, K. A novel Approach to Unsupervised Pattern Discovery in Speech Using Convolutional Neural Network. *Computer Speech & Language*, 2022, 71, 101259. <https://doi.org/10.1016/j.csl.2021.101259>
 26. Kwak, G.-H., Park, N.-W. Impact of Texture Information on Crop Classification with Machine Learning and UAV Images. *Applied Sciences*, 2019, 9(4), 643. <https://doi.org/10.3390/app9040643>
 27. Lee, E., Rustam, F., Washington, P.B., El Barakaz, F., Aljedaani, W., Ashraf, I. Racism Detection by Analyzing Differential Opinions Through Sentiment Analysis of Tweets Using Stacked Ensemble GCR-NN Model. *IEEE Access*, 2022, 10, 9717-9728. <https://doi.org/10.1109/ACCESS.2022.3144266>
 28. Lee, E., Rustam, F., Ashraf, I., Washington, P. B., Narra, M., Shafique, R. Inquest of Current Situation in Afghanistan Under Taliban Rule Using Sentiment Analysis and Volume Analysis, *IEEE Access*, 2022, 10, 10333-10348. <https://doi.org/10.1109/ACCESS.2022.3144659>
 29. Li, S., Wang, Y., Xue, J., Zhao, N., Zhu, T. The Impact of Covid-19 Epidemic Declaration on Psychological Consequences: A Study on Active Weibo Users. *International Journal of Environmental Research and Public Health*, 2020, 17. <https://doi.org/10.3390/ijerph17062032>
 30. Limaye, R. J., Sauer, M., Ali, J., Bernstein, J., Wahl, B., Barnhill, A., Labrique, A. Building Trust While Influencing Online COVID-19 Content in the Social Media World. *Lancet Digit. Heal.* 2, 2020, e277-e278. [https://doi.org/10.1016/S2589-7500\(20\)30084-4](https://doi.org/10.1016/S2589-7500(20)30084-4)
 31. Liu, Y., Chawla, N. V., Harper, M. P., Shriberg, E., Stolcke, A. A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech. *Computer Speech & Language*, 2006, 20, 468-494. <https://doi.org/10.1016/j.csl.2005.06.002>
 32. Luo, Y., Xu, X. Comparative Study of Deep Learning Models for Analyzing Online Restaurant Reviews in the Era of the COVID-19 Pandemic. *International Journal of Hospital Management*, 2021, 94, 102849. <https://doi.org/10.1016/j.ijhm.2020.102849>
 33. Maiya, A. S. ktrain: A Low-Code Library for Augmented Machine Learning. *ArXiv*, 2020. <http://arxiv.org/abs/2004.10703> (accessed May 2, 2021).
 34. Mancini, F., Sousa, F. S., Teixeira, F. O., Falcão, A. E. J., Hummel, A. D., da Costa, T. M., Calado, P. P., de Araújo, L. V., Pisa, I. T. Use of Medical Subject Headings (MeSH) in Portuguese for Categorizing Web-based Healthcare Content. *Journal of Biomedical Informatics*, 2011, 44, 299-309. <https://doi.org/10.1016/j.jbi.2010.12.002>
 35. Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., Ashraf, I. Sentiment Analysis and Topic Modeling on Tweets about Online Education during COVID-19. *Applied Science*, 2021, 11, 8438. <https://doi.org/10.3390/app11188438>
 36. Nguyen, T. T., Criss, S., Dwivedi, P., Huang, D., Keralis, J., Hsu, E., Phan, L., Nguyen, L. H., Yardi, I., Glymour, M. M., Allen, A. M., Chae, D. H., Gee, G. C., Nguyen, Q. C. Exploring U.S. Shifts in Anti-Asian Aentiment with the Emergence of COVID-19. *International Journal of Environmental Research and Public Health*, 2020, 17, 1-13. <https://doi.org/10.3390/ijerph17197032>
 37. Reshi, A. A., Rustam, F., Aljedaani, W., Shafi, S., Alhossan, A., Alrabiah, Z., Ahmad, A., Alsuwailam, H., Almangour, T. A., Alshammari, M. A., Lee, E., Ashraf, I. COVID-19 Vaccination-Related Sentiments Analysis: A Case Study Using Worldwide Twitter Dataset, *Healthc. (Basel, Switzerland)*, 2022, 10, 411. <https://doi.org/10.3390/healthcare10030411>
 38. Sağlam, F. Otomatik Duygu Sözlüğü Geliştirilmesi ve Haberlerin Duygu Analizi, 2019. <http://www.openaccess.hacettepe.edu.tr:8080/xmlui/handle/11655/9305> (accessed August 4, 2021).
 39. Samuel, J., Ali, G. G. M. N., Rahman, M. M., Esawi, E., Samuel, Y. COVID-19 Public Sentiment Insights and Machine Learning for Tweets Classification. *Information*, 2020, 11(6), 314. <https://doi.org/10.3390/info11060314>

40. Siddiqui, S. A., Salman, A., Malik, M. I., Shafait, F., Mian, A., Shortis, M. R., Harvey, E. S. Automatic Fish Species Classification in Underwater Videos: Exploiting Pre-trained Deep Neural Network Models to Compensate for Limited Labelled Data. *ICES Journal of Marine Science*, 2018, 75, 374-389. <https://doi.org/10.1093/icesjms/fsx109>
41. Sonsoz, Pandemi döneminde internet kullanımını arttı (Sonsöz Gazetesi), (2020). <https://sonsoz.com.tr/pandemi-doneminde-internet-kullanimi-artti/> (accessed May 2, 2021).
42. Wearesocial, Digital 2020 - We Are Social, 2020. <https://wearesocial.com/digital-2020> (accessed May 2, 2021).
43. Wei, J., Liao, J., Yang, Z., Wang, S., Zhao, Q. BiLSTM with Multi-Polarity Orthogonal Attention for Implicit Sentiment Analysis. *Neurocomputing*, 2020, 383, 165-173. <https://doi.org/10.1016/j.neucom.2019.11.054>
44. Xie, R., Chu, S. K. W., Chiu, D. K. W., Wang, Y. Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis. *Data and Information Management*, 2021, 5, 86-99. <https://doi.org/10.2478/dim-2020-0023>
45. Yoo, S. Y., Jeong, O. R. Automating the Expansion of a Knowledge Graph. *Expert Systems with Applications*, 2020, 141, 112965. <https://doi.org/10.1016/j.eswa.2019.112965>



This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (<http://creativecommons.org/licenses/by/4.0/>).