Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Social media-based COVID-19 sentiment classification model using Bi-LSTM

Mohamed Arbane [a], Rachid Benlamri [b], Youcef Brik [a], Ayman Diyab Alahmar [c,*]

[a] LASS Laboratory, Mohamed Boudiaf University, M'sila, 28000, Algeria
[b] University of Doha for Science and Technology, Doha, PO Box 24449, Qatar
[c] Department of Software Engineering, Lakehead University, Thunder Bay, P7B 5E1, Ontario, Canada

ARTICLE INFO

Keywords:
Bi-LSTM
COVID-19
Deep learning
Natural language processing
Sentiment classification
Social media

ABSTRACT

Internet public social media and forums provide a convenient channel for people concerned about public health issues, such as COVID-19, to share and discuss information/misinformation with each other. In this paper, we propose a natural language processing (NLP) method based on Bidirectional Long Short-Term Memory (Bi-LSTM) technique to perform sentiment classification and uncover various issues related to COVID-19 public opinions. Bi-LSTM is an improved version of conventional LSTMs for generating the output from both left and right contexts at each time step. We experimented with real datasets extracted from Twitter and Reddit social media platforms, and our experimental results showed improved metrics compared with the conventional LSTM model as well as recent studies available in the literature. The proposed model can be used by official institutions to mitigate the effects of negative messages and to understand peoples' concerns during the pandemic. Furthermore, our findings shed light on the importance of using NLP techniques to analyze public opinion and to combat the spreading of misinformation and to guide health decision-making.

1. Introduction

Social networking and microblogging services with millions of users, such as Twitter and Reddit, have the power of shaping public opinion. Recent statistics on Twitter popularity, as one of the largest microblogging social services, indicate that Twitter has 340 million users communicating 500 million tweets per day (Omnicore, 2021). Fig. 1 shows the increase in monetizable Twitter daily active users over recent years. Twitter is seeing a record number of users flock to its service amid the coronavirus pandemic (Washington-Post, 2021). An interesting finding on the effect of Twitter on shaping public opinion is that 71% of Twitter users say they use Twitter network to get their news (Omnicore, 2021).

To assess the impact of social media messages related to COVID-19 pandemic, studies related to public opinion on health news and practices have gained high importance. In particular, text analysis of Twitter data has been the focus of many studies, enabling researchers to analyze large samples of user-generated content to discover insights that can inform decision making and early response strategies. Twitter platform is experiencing a massive infusion of information related to COVID-19 issues. Tweets are good source to measure public awareness of COVID-19 pandemic trends and to understand how the pandemic impacted people's lives, their well-being and socio-economic habits, and their reactions to government measures such as lockdown plans and other restrictions.

On social media sites, people express their positive and negative opinions. They can also share their concerns and thoughts on public health decisions related to COVID-19. Such concerns include the developed vaccines, government's stay-at-home orders, social distancing, proper ventilation, face masks, and personal hygiene. This makes famous social networking sites, such as Twitter and Reddit, an important platform for people awareness of best health practices to overcome the pandemic. However, this powerful platform might also work on the opposite direction by acting as a tool to spread misinformation and negative tweets, and to oppose government's restrictions, which can worsen the pandemic.

The objective of this research is to investigate the use of deep learning models and natural language processing (NLP) techniques such as Sentiment Analysis (SA) in helping policy makers and communities to stop the spread of misinformation, fake news, and incitement of insurrection. Sentiment analysis, or public opinion mining, is defined as the process of using NLP and machine learning (ML) to classify opinions and emotions in subjective information. SA is considered as one of the most popular research areas in the field of NLP as it provides a means to review and analyze opinions that are held by any number of individuals (Liu, Shi, Ji, & Jia, 2019; Tyagi & Tripathi, 2019).

* Corresponding author.
E-mail addresses: mohamed.arbane@univ-msila.dz (M. Arbane), rachid.benlamri@udst.edu.qa (R. Benlamri), youcef.brik@univ-msila.dz (Y. Brik), aalahmar@lakeheadu.ca (A.D. Alahmar).

https://doi.org/10.1016/j.eswa.2022.118710
Received 7 November 2021; Received in revised form 26 June 2022; Accepted 25 August 2022
Available online 30 August 2022

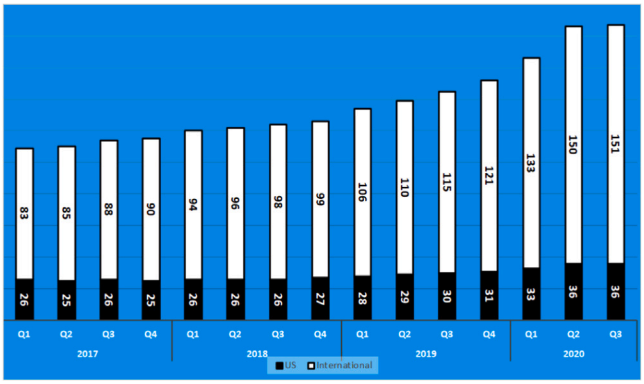


Fig. 1. Monetizable Twitter daily active users over years (2017–2020).
Source: Data source: Twitter.

In this paper, we investigate a deep learning model for sentiment analysis to uncover various issues related to COVID-19 from public opinion. Our investigation was guided by the following research questions (RQ):

(RQ1) What is the performance of various deep learning and NLP algorithms for sentiment classification related to COVID-19 tweets and which one performs better?
(RQ2) How can novel techniques in NLP and deep learning such as Bi-Directional LSTM efficiently analyze the user behavior in online discussion forums?
(RQ3) How can we help government and health care decision makers combat the spreading of false news and negative tweets /comments that may impact people's health?

To address the above research questions, we focused on analyzing COVID-19 related tweets to detect polarity sentiments relating to COVID-19 based on the public opinion on Twitter and Reddit. Specifically, we developed a deep learning framework that is effective for detecting the actual sentiment behind social media messages related to COVID-19 based on Bidirectional Long Short-Term Memory (Bi-LSTM) and embedding matrix.

The major contributions of this paper are as follows:

- An improved recurrent neural networks (RNNs) framework based on Bidirectional Long Short-Term Memory (Bi-LSTM) for sentiment classification of COVID-19–related tweets, which produces better results than other familiar deep learning methods.
- An interesting polarity analysis of worldwide COVID-19 tweets and insights about collective reactions on coronavirus outbreak on social media.
- A supervised deep learning sentiment detection model for Twitter and Reddit to foresee people's reactions during the COVID-19 pandemic.

The remaining sections of the paper are organized as follows. Section 2 discusses related work and highlights the research challenges in this field. Section 3 describes the main components of the proposed COVID-19 Sentiment Classification framework. In Section 4, we present and discuss the experimental results. Section 5 presents a comparative analysis with similar works. Finally, Section 6 concludes the paper and provides future research directions.

2. Related work

Many researchers have addressed the application of NLP and deep learning methods in COVID-19 sentiment classification. Previous research has shown that Bi-LSTM method is one of the most successful methods in sentiment analysis. For example, Xu, Meng, Qiu, Yu, and Wu (2019) have compared Bi-LSTM use in sentiment analysis with the state-of-art methods such as Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), traditional LSTM, and Naïve Bayes (NB). The obtained results showed that Bi-LSTM outperforms the traditional techniques, provides a better capture of context information, and has higher precision, recall and F1 than the other methods. Wu et al. (2021) demonstrated in their research that Bi-LSTM units can effectively solve the gradient problems and can well capture the contextual semantic information.

Jelodar, Wang, Orji, and Huang (2020) reported the use of automated extraction of COVID-19 discussions from online Reddit discussion forum using topic modeling to uncover issues related to COVID-19 from public opinions. They have also investigated how to use LSTM recurrent neural network for COVID-19 sentiment classification. Their model achieved an accuracy of 81.15%.

Nemes and Kiss (2020) analyzed the sentiments and manifestations (hashtags, comments, tweets, and posts) from Twitter based mainly on ‘covid’ and ‘coronavirus’ keywords. Using RNN to classify emotions, they developed a model to analyze the emotional nature of various tweets. They have classified the tweets into various classes of emotional strength (weakly positive/negative, strongly positive/negative). Their comparisons were made mainly against TextBlob (a third-party sentiment analyzer) and they have shown that the recurrent neural network model provides good performance and prediction in text classification of this kind.

Samuel, Ali, Rahman, Esawi, Samuel, et al. (2020) identified coronavirus specific Tweets and used the R statistical software, along with its sentiment analysis packages, for developing two machine learning classification models and comparing their effectiveness in classifying coronavirus Tweets of varying lengths. They observed a classification accuracy of 91% for short Tweets, with the Naïve Bayes method. They also observed that the logistic regression classification method provided a reasonable accuracy of 74% with shorter Tweets, and both methods showed relatively weaker performance for longer Tweets. Their research provided insights into coronavirus fear sentiment progression.

Mansoor, Gurumurthy, Prasad, et al. (2020) used various machine learning models such as LSTM and ANN to perform a sentiment classification and analysis of Coronavirus tweets. Their study focused on how the sentiment of people in different countries has changed over time. They used tweets related to online learning and work from home (WFH) and they observed the change in sentiment over time. Their experimental results showed an accuracy of 84.5% using LSTM and 76% using ANN on the coronavirus tweets dataset. They also observed that the fear emotion has been significantly greater than the trust emotion throughout the pandemic. Regarding countries, they found that the countries with a greater proportion of positive sentiment include Bangladesh, Pakistan, Mali, and South Africa. Whereas, Australia, India, USA, Turkey, UK, and Brazil are among the countries of higher proportion of negative sentiments. The calculated emotions score showed that Thailand, Vietnam, and Poland have the highest fear score, whereas the highest trust scores were exhibited by Oman, Syria, and Kazakhstan.

Imran, Daudpota, Kastrati, and Batra (2020) conducted a study to analyze the reaction of people from different cultures to Coronavirus and citizen's sentiment about Coronavirus actions taken by different countries. They used Deep LSTM models for estimating the emotions and sentiment polarity from tweets extracted during the period from early February 2020 until the end of April 2020. They justified the use of only the initial few weeks of the pandemic by the argument that “people usually get accustomed to the situation over time and an initial phase is enough to grasp the general/overall behavior of the masses towards a crisis and the policies adopted by respective governments” (Imran et al., 2020). They had four main conclusions: (1) NLP-based deep learning models can provide cultural and emotional insight across cross-cultural trends; (2) the observations of users' concern and their response to respective Governments' decision on Coronavirus resonate with sentiments analyzed from the Twitter posts; (3) there is a high correlation between the sentiments expressed between the neighboring countries within a region (e.g., Pakistan and India, similar to the USA and Canada, have similar polarity trends, unlike Norway and Sweden); and (4) both positive and negative emotions were equally observed concerning Coronavirus lockdowns; however, in Canada, Pakistan, and Norway the average number of positive tweets was more than the negative ones (Imran et al., 2020).

Rustam et al. (2021) performed COVID-19 sentiment analysis on a Twitter dataset from which Tweets were extracted by an in-house built crawler that uses the Tweepy library. The sentiments were extracted using the TextBlob library. They evaluated the performance of various ML classifiers using their proposed feature set that was formed by concatenating the bag-of-words and the term frequency-inverse document frequency. Tweets were classified as positive, neutral, or negative. Their results show that Extra Trees Classifiers outperform all other models by achieving a 0.93 accuracy score.

Naseem, Razzak, Khushi, Eklund, and Kim (2021) analyzed a dataset of 90,000 COVID-19-related tweets for sentiment classification using different sets of features and classifiers. The tweets were collected in the early stages of the pandemic (from February to March 2020) and have been labeled into positive, negative, and neutral sentiment classes. Their findings indicate that most people favored the lockdown and stay home order in February; however, their opinion shifted by mid-March due to misinformation being spread on social media. They concluded that there is a need to develop agile and proactive public health presence to combat the spread of fake news.

Also during the pandemic, several research works have been developed to better understand the impact of the emotions expressed by people as reported in Choudrie, Patil, Kotecha, Matta and Pappas (2021). The latter research applied deep learning techniques based on transfer learning and with a Robustly Optimized BERT Pretraining called 'Roberta' to classify and analyze people emotions. Their research findings suggest that worry was more prominent in the earlier months of lockdown, while hate, anger, and depression became more intense later on. These results were confirmed with a sentiment classifier accuracy of 80.33%. Additionally, Wagle, Kaur, Kamat, Patil, and Kotecha (2021) proposed a system for explaining why one particular piece of information is classified as misinformation by building an explainable AI-assisted multimodal credibility assessment system. Their method is characterized by rating the credibility of misleading forums or blogs using multiple crucial modalities which would lead to an insight information consumption by users.

Gaikwad, Ahirrao, Phansalkar, and Kotecha (2021) showed how social media is used to spread online radicalization, and suggested that better detection and classification of social media extremism is needed. They also investigated data standardization and bias, and its validation using statistical techniques.

Baydogan and Alatas (2021) proposed a novel approach for sentiment classification within a labeled twitter dataset based on the Social Spider Optimization Algorithm (SSA). In this work, the authors showed that the social networks could also be analyzed in terms of optimization search. A comparison between their method and some traditional machine learning techniques was evaluated where their method achieved the highest performance metrics. Similarly, Akyol and Alatas (2020) examined the sentiment analysis challenge as an optimization and multiobjective problem using Whale Optimization Algorithm (WOA) and Social Impact Theory-based Optimization (SITO) methods. Furthermore, Yildirim, Yildirim, and Alatas (2021) proposed a new approach for sentiment analysis task by combining the sunflower optimization and chaos theory. The experiments of their proposed system were conducted on the Trip Advisor dataset. More recently, Sitaula and Shahi (2022) represented each tweet by combining both syntactic and semantic information. Then, they designed a multi-channel convolutional neural network (MCNN) for sentiment classification using domain-specific (ds) methods with bag of words (BOW). Their experiments were conducted on tweets collection with three sentiment classes. The results showed that feature combination and MCNN model achieved 71.3% in term of accuracy.

3. Framework methodology

In this section we describe the main components of the proposed COVID-19 Sentiment Classification framework. A deep learning model based on bidirectional LSTM is used for analyzing COVID19-related tweets polarity. The deep learning model is trained using thousands of tweets posted during the COVID-19 health crisis. The overall architecture of the proposed system is illustrated in Fig. 2.

3.1. Datasets description

In this study we use three datasets to develop and validate the proposed system. Two datasets collected from Twitter and one dataset collected from Reddit platform. The first Twitter dataset (Miglani, 2020) consists of 44,957 COVID-relevant tweets collected from users who have applied the following hashtags: #coronavirus, #coronavirusoutbreak, #covid19. These tweets were collected from March to April 2020, and were originally manually labeled for five different polarities (Extremely Negative, Negative, Neutral, Positive and Extremely Positive). For the purpose of this study, and in order to have an overall dataset with homogeneous polarity set, we re-grouped the five polarities into three polarities, thus merging extremely negative with negative tweets, and extremely positive with positive tweets. The outcome of such restructuring resulted into a dataset consisting of 17,031 negative, 8334 neutral and 19,592 positive tweets. Fig. 3 shows sample tweets from this dataset.

The second Twitter dataset (Smith, 2020) is much larger and consists of 460,286 tweets collected in a similar way to the first dataset, but labeled using 13 polarities. These are: sadness, enthusiasm, neutral, worry, love, fun, surprise, hate, empty, happiness, boredom, relief and anger. We re-grouped these polarities into three polarities as follows: Negative polarity class includes sadness, hate, boredom, and anger; neutral polarity class includes neutral, empty, and surprise; and positive polarity class includes love, fun, happiness, relief, worry and enthusiasm.

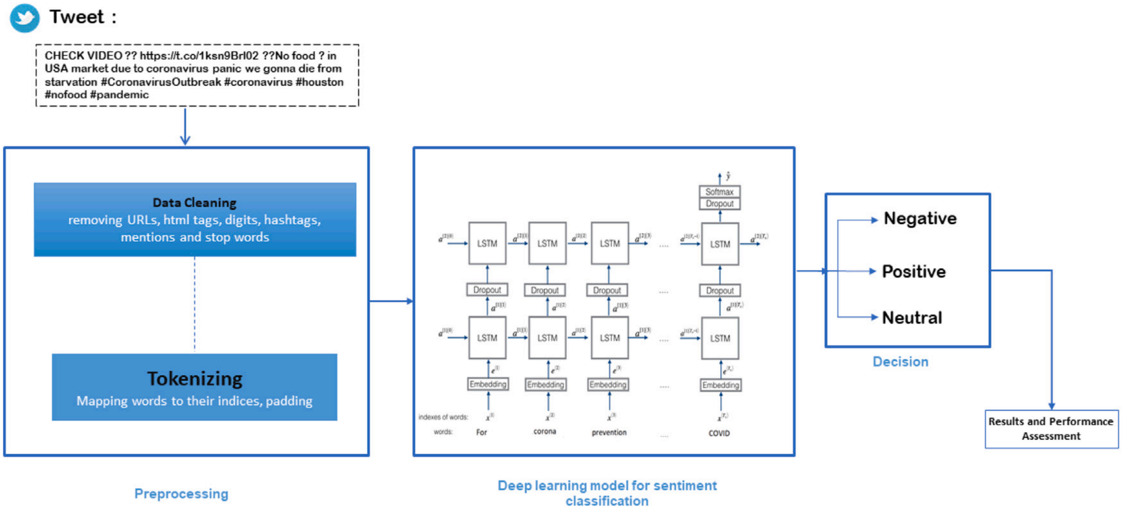
The third dataset used in this work consists of 563,079 COVID-19–related comments from Reddit platform. These were collected between January 20, 2020 and March 19, 2020. The collected comments were manually grouped by Jelodar et al. (2020) into five polarity classes (Very Positive, Positive, Neutral, Negative, and Very Negative). Similarly to the first dataset, we created a new version of this dataset by re-grouping the Reddit comments into three polarity classes. It should be noted here that both of the original and the new datasets were used for validate our model.

We selected these datasets on the basis of open source data, and our goal was to compare our model with other models that have used the same datasets. Therefore, as long as the comments/tweets are related to the Coronavirus pandemic, it does not matter when the data was collected.

3.2. Preprocessing steps

3.2.1. Data cleaning

Tweets are full of noise because of the presence of irregular expressions, http addresses, hashtags, incomplete words, and out of dictionary words. Therefore, a number of text preprocessing and data cleaning steps are needed before COVID-19 tweets can be successfully used for any machine learning task. The first preprocessing task is to remove URLs, html tags, digits, and hashtags, which get embedded in the original data. The next step is to remove stop words that have a very little meaning in natural language, such as "a", "an", "the", and "so". Stop words are often removed from text before training deep learning models, as these words occur in abundance, where little or no unique information can be used for classification or clustering. To perform the above-mentioned data cleaning operations, we used the NLTK Python library.

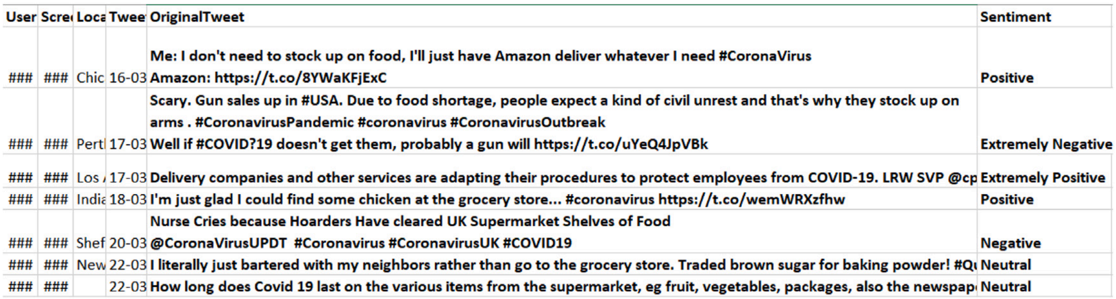


Tweet :

CHECK VIDEO ?? https://t.co/1ksn9Brl02 ??No food ? in USA market due to coronavirus panic we gonna die from starvation #CoronavirusOutbreak #coronavirus #houston #nofood #pandemic

Data Cleaning
removing URLs, html tags, digits, hashtags, mentions and stop words
Tokenizing
Mapping words to their indices, padding
Preprocessing
LSTM
Dropout
Embedding
Softmax
$\hat{y}$
$a^{[1]\langle 0\rangle}$
$a^{[2]\langle 0\rangle}$
$e^{\langle 1\rangle}$
$x^{\langle 1\rangle}$
indexes of words:
words:
For
corona
prevention
COVID
Deep learning model for sentiment classification
Negative
Positive
Neutral
Decision
Results and Performance Assessment

Fig. 2. The overall framework of the proposed system.



User	Scre	Loca	Twee	OriginalTweet	Sentiment
###	###	Chic	16-03	Me: I don't need to stock up on food, I'll just have Amazon deliver whatever I need #CoronaVirus Amazon: https://t.co/8YWaKFjExC	Positive
###	###	Pert	17-03	Scary. Gun sales up in #USA. Due to food shortage, people expect a kind of civil unrest and that's why they stock up on arms . #CoronavirusPandemic #coronavirus #CoronavirusOutbreak Well if #COVID?19 doesn't get them, probably a gun will https://t.co/uYeQ4JpVBk	Extremely Negative
###	###	Los /	17-03	Delivery companies and other services are adapting their procedures to protect employees from COVID-19. LRW SVP @cp	Extremely Positive
###	###	India	18-03	I'm just glad I could find some chicken at the grocery store... #coronavirus https://t.co/wemWRXzfhw	Positive
###	###	Shef	20-03	Nurse Cries because Hoarders Have cleared UK Supermarket Shelves of Food @CoronaVirusUPDT #Coronavirus #CoronavirusUK #COVID19	Negative
###	###	New	22-03	I literally just bartered with my neighbors rather than go to the grocery store. Traded brown sugar for baking powder! #Qu	Neutral
###	###		22-03	How long does Covid 19 last on the various items from the supermarket, eg fruit, vegetables, packages, also the newspap	Neutral

Fig. 3. Sample tweets from the first dataset used in this work.

3.2.2. *Tokenizing*

After the data cleaning step, we split all data sentences into words using the word tokenizer process. The latter creates an internal dictionary of unique words and assigns an integer to every word. Thus, the output of word tokenization is sequences of indexes representing the key values in the tokenizer created dictionary. Since tweets have different number of words, the text-to-sequence transformation process produces sequences of different dimensions. Such sequences would require different computation steps on the Long-Short Term Memory (LSTM) based deep learning model. However, since any deep learning model needs to know the exact number of feature dimensions, and that number must be the same for both training and prediction on each observation, we need to convert the sequences into a well-structured matrix (embedding matrix). Consequently, padding is used to address the issue of sequences of varying length. This is done in this study by padding all sequences to a 54 sequence-length, which represents the tweet with the largest number of words found in our datasets. The example provided below illustrates the processes of text-to-sequence and padding.

Let us consider the following preprocessed tweet: “Coronavirus Australia Woolworths give elderly disabled dedicated shopping hours amid COVID- outbreak”.

After text-to-sequence mapping, the text above will be represented by a sequence of 12 indexes:

[26, 674, 4360, 247, 259, 1412, 1762, 12, 194, 91, 1, 66].

After padding the sequence above, we obtain the following results:

[26 674 4360 247 259 1412 1762 12 194 91 1 66 0 … 0].

3.3. *Deep learning model for sentiment classification*

In this study, we model the problem of sentiment polarity classification as a supervised sequence labeling problem. We consider three sentence polarities labeled as follows, 0 denotes Negative, 1 denotes Neutral, and 2 denotes Positive. Formally, given the input training pair (x; Y), the process of training the model aims to fine-tune the optimal parameter weights, where $Y_i \subset \{0, 1, 2\}$ and $x_i \subset R^d$ is the word embedding in a d-dimensional vector space (Pennington, Socher, & Manning, 2014).

Human vocabulary comes in free text. In order to make the deep learning model understand and process the natural language, we need to transform the free-text words into numeric values. One of the simplest transformation approaches is to do a one-hot encoding, in which each distinct word stands for one dimension of the resulting vector and a binary value indicates whether the word is present (one) or not (zero).

One problem of one-hot encoding is associating distant numbers to similar words. To address this shortcoming, in this study we used word embedding (Pennington et al., 2014) which is a dense representation of words in the form of numeric vectors. The advantage of using word embedding is that similar words are located together in the vector space, and arithmetic operations on word vectors can pose semantic or syntactic relationships. For example, the semantic distance between the vectors associated to “king” and “queen” will be similar to that between the vectors associated to “man” and “woman” as shown in Fig. 4.

Since our inputs and outputs are sequential with the same length, we use Recurrent Neural Networks (RNNs) (Sherstinsky, 2020). These networks are a class of artificial neural sequence model as shown in Fig. 5, where connections between units form a directed cycle. It takes arbitrary embedding sequences $x = (x_{t-1} \ldots x_t \ldots, x_{t+1})$ as input, and uses its internal memory network to exhibit dynamic temporal behavior. It consists of a hidden unit h and an optional output y. The parameter t is the time step and is also the length of input sentence in this text sequence learning task. At each time step t, the hidden state h_t of the

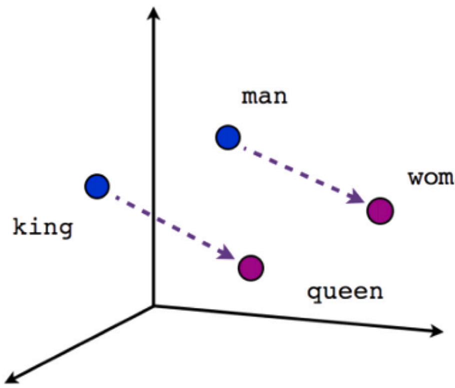


Fig. 4. Example of the semantic distance between the words in word embedding space.

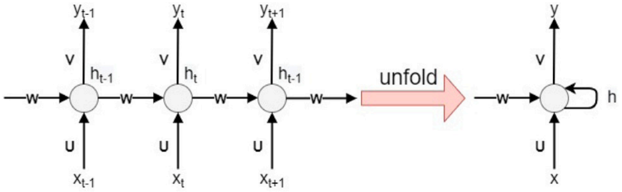


Fig. 5. A representation of RNN Cell unit.

RNN is computed based on the previous hidden state h_{t-1}, and the input at the current step x_t is computed as shown below:

$$h_t = g(U_{x_t} + W_{h_{t-1}}) \tag{1}$$

where U and W are weight matrices of the network; and g is a non-linear activation function, such as a ReLu function.

In particular, we use an architecture template where firstly the input passes through an Embedding layer followed by a stack of RNNs, which comes out towards a Dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). The Dropout effectively grays out certain neurons during training. This operation strengthens the overall model by reducing the likelihood of overfitting. Finally, the Dropout outputs pass through a dense layer with SoftMax activation function.

To overcome the problem of vanishing gradients (Hochreiter, 1998) that RNN models suffer from, we replace them by LSTM Networks (Hochreiter & Schmidhuber, 1997). In fact, LSTM models respond to this problem by using gates that can learn to remember states and move them through the network. The equations for activation, update, forget, and output gates, along with the current and previous states are given below, where symbols have their usual meanings.

$$\Gamma_f^{(t)} = \sigma(W_f\left[a^{(t-1)}, x^{(t)}\right] + b_f) \tag{2}$$

$$\Gamma_u^{(t)} = \sigma(W_u\left[a^{(t-1)}, x^{(t)}\right] + b_u) \tag{3}$$

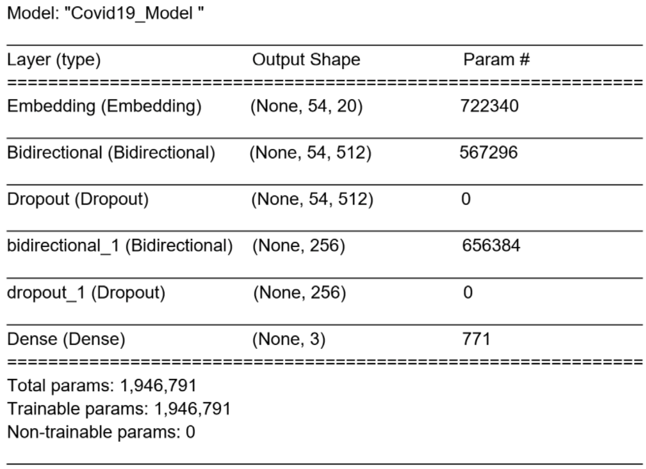
$$\tilde{c} = tanh(W_c\left[a^{(t-1)}, x^{(t)}\right] + b_c) \tag{4}$$

$$c^t = \Gamma_f^{(t)} * c^{(t-1)} + \Gamma_u^{(t)} * \tilde{c}^{(t)} \tag{5}$$

$$\Gamma_o^{(t)} = \sigma(W_o\left[a^{(t-1)}, x^{(t)}\right] + b_o) \tag{6}$$

$$a^{(t)} = \Gamma_o^{(t)} * tanh(c^{(t)}) \tag{7}$$

Furthermore, in this study we use the Bi-directional LSTMs (Bi-LSTM) (Schuster & Paliwal, 1997) which are improved version of conventional LSTMs for generating the output from both left and right contexts at each time step. For Bi-LSTMs architecture, we use the same conventional LSTM nodes except that the input can move from both sides (i.e., from left to right as well as from right to left). The loss function used in the last layer of our model is weighted SoftMax cross entropy loss function, defined as follows.



```
Model: "Covid19_Model "
_____________________________________________________________________
Layer (type)                   Output Shape              Param #
=====================================================================
Embedding (Embedding)          (None, 54, 20)            722340
_____________________________________________________________________
Bidirectional (Bidirectional)  (None, 54, 512)           567296
_____________________________________________________________________
Dropout (Dropout)              (None, 54, 512)           0
_____________________________________________________________________
bidirectional_1 (Bidirectional) (None, 256)              656384
_____________________________________________________________________
dropout_1 (Dropout)            (None, 256)               0
_____________________________________________________________________
Dense (Dense)                  (None, 3)                 771
=====================================================================
Total params: 1,946,791
Trainable params: 1,946,791
Non-trainable params: 0
_____________________________________________________________________
```

Fig. 6. Proposed Bi-LSTM model parameters.

$$\hat{Y} = \frac{e^{xi}}{\sum_{i=1}^{m} e^{xi} + \varepsilon} \tag{8}$$

However, we use Mean Squared Error (MSE) loss in other layers. As hyperparameters, we can set Adam optimization, tuned learning rate and input sequence length. For the evaluation of validation and accuracy, we use the mean absolute percentage error.

The main advantage of our system is to replace regular LSTM with bidirectional LSTM so that our system can learn from both directions in the text, from left-to-right and right-to-left, unlike systems that use normal LSTM that allows their system to learn from left-to-right only. We also use the method of word embedding that helps the system to understand words close in meaning as well as appropriate choices of preprocessing techniques, parameters and hyper-parameters like tokenizing, word padding, loss function, optimization algorithm (Adam), learning rate (0.001), and number of units in Bi-LSTM layers. All these helped in obtaining a significant improvement in learning results.

4. Experimental results and discussion

Many experiments were conducted in this study to answer the three research questions (RQ1, RQ2 and RQ3) addressed in this research. In particular, four scenarios were considered using different datasets. In order to answer research question RQ1, we compared the performance of the proposed model in the four different scenarios with the models reported in related works section(state-of-the-art models). Below we describe the experimental setup, results and discussion for each of these scenarios also we chose each performance metric according to the system it was compared to, in each scenario . In all the scenarios our algorithms were implemented on a PC-INTEL i7-7700k CPU 4.20 GHz, with 16 GB and NVIDIA GTX1070 GPU.

4.1. First scenario

Our first experiment was conducted on the first Twitter dataset (Miglani, 2020) where loss and accuracy are used in training and validation phases to evaluate the proposed Bi-LSTM model. We recall here that this dataset has three different polarities (Positive, Neutral, and Negative). Figs. 6 and 7 show the Bi-LSTM model parameters and its performance using the first dataset, respectively. The obtained training accuracy was 95% and 86% in the validation phase.

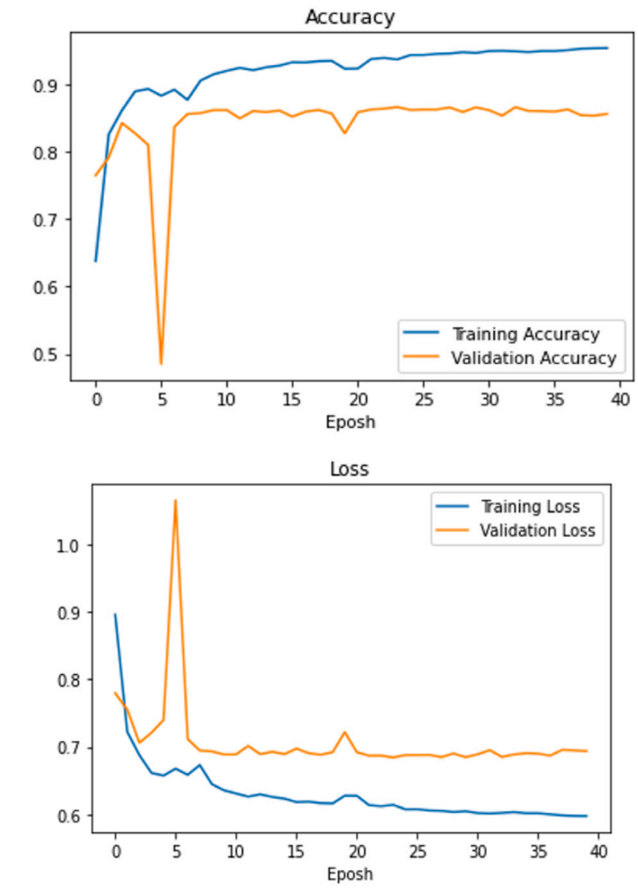


Fig. 7. Performance of the first scenario.

Table 1
Performance metrics of the second scenario in validation phase.

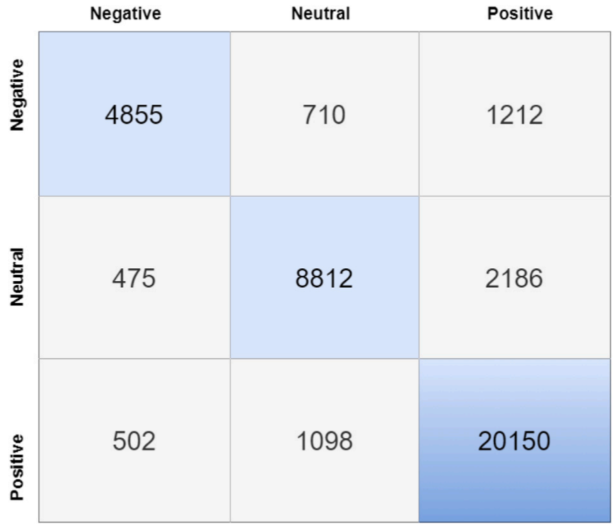
	Precision	Recall	F1-score
Negative	83%	72%	77%
Neutral	83%	77%	80%
Positive	86%	93%	89%

4.2. Second scenario

In order to verify the robustness of our sentiment classification model, we performed an experiment similar to that described in the first scenario, but applied on a tweet dataset that is ten times larger. We recall here that our model was used without re-training its parameters to know how it behaves with the new dataset while keeping the same labels with three polarity classes. Fig. 8 shows the obtained confusion matrix from this experiment. The obtained accuracy of our model was 84.54%. Furthermore, we can extract from this matrix some significant performance metrics such as Precision, Recall, and F1-score as depicted in Table 1.

The precision values for Negative, Neutral, and Positive classes were 83%, 83%, and 86%, respectively. While F1-score values were 77%, 80%, and 89%, respectively. In addition, the Recall assessment on positive tweets of our model achieved the best scores in the validation phase with 93%. However, there is a fairly large portion of negative and neutral tweets that are classified as positive ones.

In order to make a meaningful comparison between our second scenario and models that have been developed for the same purpose with the same data, we report our performances in Table 2 with those of Imran et al. (2020). It is worth noticing that the two models evaluated in Imran et al. (2020) have only two polarities (Negative and Positive). Hence, the comparison revolves around the ground-truth of each tweet since the data comes from the same source (i.e. Twitter data). We clearly see from Table 2 that our model outperforms the two models of Imran et al. (2020), which justifies the robustness of our Bi-LSTM model even without re-training its parameters (i.e. transfer learning).



	Negative	Neutral	Positive
Negative	4855	710	1212
Neutral	475	8812	2186
Positive	502	1098	20150

Fig. 8. Confusion matrix of the second scenario.

4.3. Third scenario

In this scenario we tested our sentiment classification model with data produced from another social media platform that has more than three polarities. This scenario is related to the original Reddit dataset that is labeled into five different polarities (Very Positive, Positive, Neutral, Negative and Very Negative). For our Bi-LSTM model, we kept the same model that is used in the first scenario without re-training (i.e. model parameters freezing), except for the last layer where five units and Softmax function have been considered. Fig. 9 shows the obtained accuracy and loss curves, where the values of accuracy and loss were 94.55% and 0.2345%, respectively.

Table 3 provides more details about our model's performance in terms of precision, recall, and F1-score. It is clear that the Very Positive class provides the lowest performance compared to the other classes. The recall rate obtained for Very Positive class was too small with only 14% compared to other polarities. This is justified by the confusion between Positive and Very Positive classes where some Very Positive comments have been classified as Positive ones as shown by the confusion matrix in Fig. 10.

By comparing our results using Reddit dataset that has 5 classes with the model of Jelodar et al. (2020) that has reached 81.15% in terms of accuracy on the same dataset, we can clearly see the superiority of our model by more than 13%. It is worth noticing here that Jelodar et al. (2020) used only 112,888 comments to validate their model. In our case, we took the entire dataset after its cleaning (451,554 comments) to prove that our model can give good results without retraining on large amount of data coming from different platforms.

4.4. Fourth scenario

In this experiment, we transformed the Reddit dataset polarity from 5-classes to 3-classes, similarly to scenarios 1 and 2. The objective of this experiment was to test our model with a similar class-polarity using the Reddit dataset which consists of 563,079 COVID-19 related

Table 2
Performance comparison of the second scenario with similar works.

NLP Model	Validation accuracy	Validation loss
LSTM + FastText (Imran et al., 2020)	82%	X
LSTM without pre-trained embedding (Imran et al., 2020)	83%	X
Our second scenario	**84.54%**	**0.69**

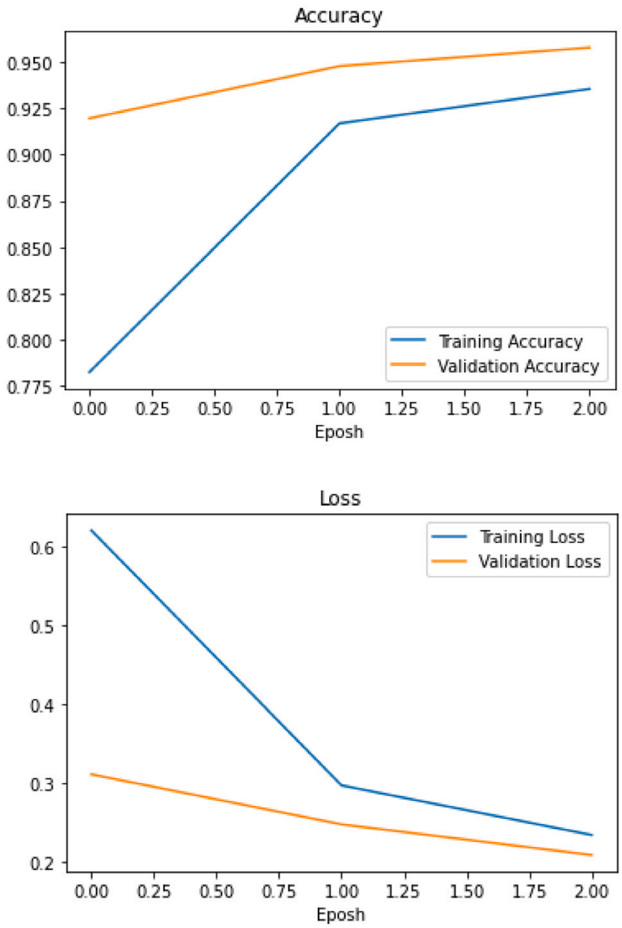
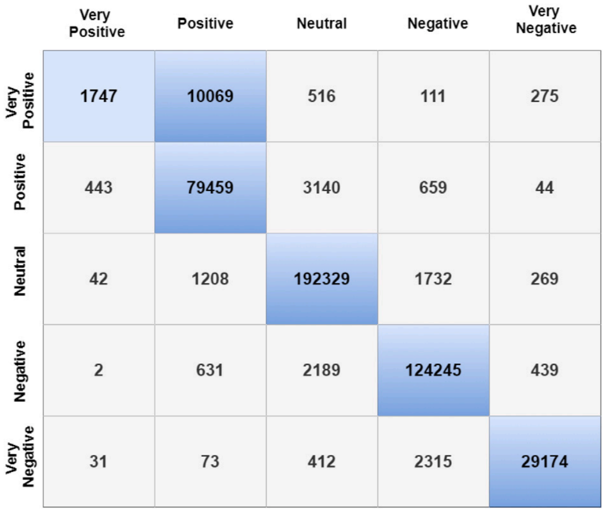


Fig. 9. Performance of the third scenario with five classes (Very Positive, Very Positive, Neutral, Negative and Very Negative).



	Very Positive	Positive	Neutral	Negative	Very Negative
Very Positive	1747	10069	516	111	275
Positive	443	79459	3140	659	44
Neutral	42	1208	192329	1732	269
Negative	2	631	2189	124245	439
Very Negative	31	73	412	2315	29174

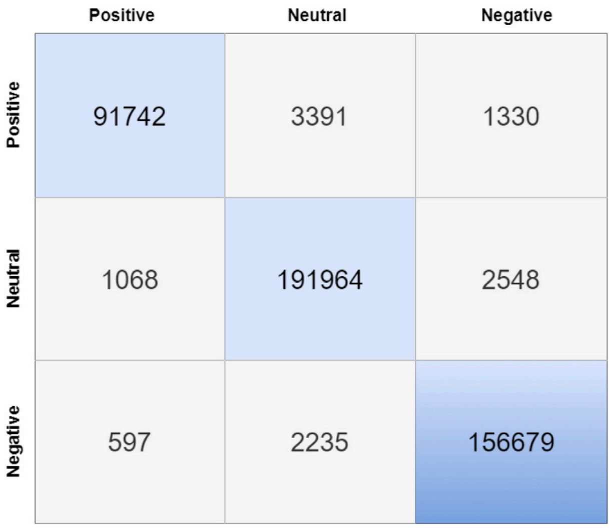
Fig. 10. Confusion matrix of third scenario with five classes.

Table 3
Performance metrics of third scenario with five classes.

	Precision	Recall	F1-score
Very negative	97%	91%	94%
Negative	96%	97%	97%
Very positive	77%	14%	23%
Positive	87%	95%	91%
Neutral	97%	98%	98%

Table 4
Performance metrics of fourth scenario with three classes.

	Precision	Recall	F1-score
Negative	98%	98%	98%
Positive	98%	95%	97%
Neutral	97%	98%	98%



	Positive	Neutral	Negative
Positive	91742	3391	1330
Neutral	1068	191964	2548
Negative	597	2235	156679

Fig. 11. Confusion matrix of fourth scenario with three classes.

comments. In fact, the majority of government and health care institutions, who consider Twitter and other social media for analyzing people perception of COVID-19 pandemic, care mostly about the three main classes (i.e., Positive, Negative, and Neutral). The new Reddit dataset polarity was constituted by merging Very Positive and Positive comments into one single class (Positive class). Similarly, Negative and Very Negative comments were merged into a new Negative class. For validation, we changed only the last layer by three units with SoftMax function. Fig. 12 shows the obtained accuracy and loss curves. The confusion matrix and statistical performances (i.e., precision, recall, and F1-score of each class) are presented in Fig. 11 as well as in Table 4.

It is clearly observed that the performance of our model using three polarities has considerably increased with 3% compared to when using five polarities. The obtained results can be justified by reducing the confusion between the Positive and Very positive classes. The obtained F1-score scores show that our Bi-LSTM model presents very satisfactory performances with 98%, 97%, and 98% for Negative, Positive, and Neutral classes, respectively.

Table 5
Comparative analysis of the proposed method with similar approaches.

Scenario	Dataset	Original polarities	Mapped polarities	Method	Validation subset size	Validation accuracy (%)
1	First Twitter dataset	5	3	Our first scenario	3798 tweets	86
2	Second Twitter dataset	13	3	LSTM without pretrained embedding (Imran et al., 2020)	40,000 tweets	83
				LSTM + GloVe (Imran et al., 2020)		82
				LSTM + FastText (Imran et al., 2020)		82
				Roberta (Choudrie, Patil et al., 2021)		80.33
				FastText + ds + BOW + MCNN (Sitaula & Shahi, 2022)		71.3
				Our second scenario		**84.54**
3	Third dataset from Reddit	5	5	Jelodar et al. (2020)	112,888 comments	81.15
				Our third scenario	451,554 comments	**94.55**
4			3	Our fourth scenario		**97.52**

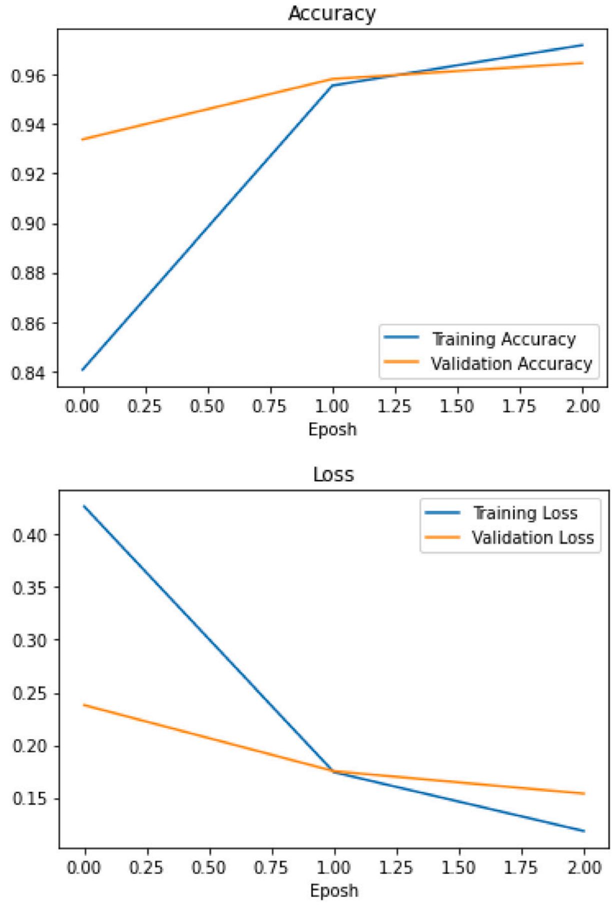


Fig. 12. Performance of third scenario with three classes (Positive, Neutral and Negative).

5. Comparative analysis

In order to give an idea on where our COVID-19 sentiment classification system ranks performance-wise, we compare our different scenarios with works that used the same experimental protocol, the same performance measures, and the same datasets. Table 5 summarizes the different parameters of each scenario as well as the obtained results compared to similar works. In the four scenarios evaluated above, our system showed consistent performance proving its superiority over models reported in related work section (state-of-the-art models). Also, the experiments carried out in these scenarios address and fulfill the objectives set in research questions RQ2 and RQ3. For research question RQ2, the experimental results shown in scenarios 2, 3 and 4 demonstrate that our Bi-Directional LSTM (RNN) model can successfully provide polarity insight across comments and tweets with very good accuracy. In particular, our system can be very useful to non-native English-speaking users who may find it sometimes difficult to discriminate between polarity in some ambiguous tweets/comments. Thus, our system can help them interpret the right polarity class.

In order to address research question RQ3, the results obtained in the four scenarios show that the proposed system can be used reliably to classify large batches of tweets and other social media comments. Most decision makers from governments and health organizations currently rely on social media to help them understand peoples' perception of their COVID-19 policies, plans and imposed restrictions. Performing such a task manually on thousands of tweets every day is not feasible as it takes a lot of time in reporting, sending and analyzing text by experts. To address this shortcoming, the proposed system can automatically classify the polarity of thousands of COVID-19 social media posts and Tweets, and help decision makers identify the most propagating misinformation as quickly as possible and address the issue before affecting a larger population in a timely manner.

It is worth mentioning that older adult's processing of online information and misinformation differs from youngsters. Choudrie et al. (2021) showed that older adults are often left confused about the accuracy of online COVID-19 information. Therefore, they prefer and trust traditional (non-online) media over online/digital media for COVID-19 information. The authors concluded that older adults are quite immune to online misinformation. The datasets that we used in this study do not have an age attribute for users, therefore, it was not possible for us to conduct a similar analysis. However, a possible future research direction is to focus on datasets with user age attributes and investigate the different patterns of online information processing between youngsters and older adults.

6. Conclusion

With the growth of social media platforms in recent years, more attention is paid to automated sentiment analysis systems using people's social media posts/tweets because of their influence on public opinion. This paper presented an improved NLP model for sentiment classification of COVID-19–related issues from social media. The proposed model is based on a Bi-LSTM system, which is tested using four different scenarios conducted on real datasets extracted from Twitter and Reddit social media platforms. The experimental results showed the superiority of our model over the conventional LSTM model as well as other recent state-of-the-art models. The proposed system can be used by healthcare authorities to help understand the fears and needs of those affected by the COVID-19 pandemic. Moreover, the proposed model can be efficiently utilized by official institutions to mitigate the effects of negative comments.

It is clear that the Very Positive class provides the lowest performance compared to the other classes. The recall rate obtained for Very Positive class was too small with only 14% compared to other polarities. This is justified by the confusion between Positive and

Very Positive classes where some Very Positive comments have been classified as Positive ones as shown by the confusion matrix.

Even though the Bi-LSTM method demonstrated promising results and significantly improved the sentiment classification on the social media datasets; there is still room for improvement, such as overcoming the limitations of Bi-LSTM being a time-consuming method and the considerable confusion between Very Positive and Positive classes. Therefore, we recommend that these limitations of Bi-LSTM should be the focus of future research in this area. Moreover, in this research, we used only English social media comments/tweets from Twitter and Reddit. To generalize our work and to be able to analyze global COVID-19 sentiments, we need multi-language social media datasets. Thus, we plan to explore models that process social media posts for COVID-19 in languages other than English.

CRediT authorship contribution statement

Mohamed Arbane: Conceptualization, Software, Validation, Writing – original draft, Writing – review & editing. **Rachid Benlamri:** Conceptualization, Supervision, Methodology, Writing – review & editing. **Youcef Brik:** Conceptualization, Software, Writing – review & editing. **Ayman Diyab Alahmar:** Conceptualization, Supervision, Methodology, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

We extend a sincere thank you to the authors of Jelodar et al. (2020) for sharing their dataset with us.

References

Akyol, S., & Alatas, B. (2020). Sentiment classification within online social media using whale optimization algorithm and social impact theory based optimization. *Physica A: Statistical Mechanics and its Applications*, *540*, Article 123094.

Baydogan, C., & Alatas, B. (2021). Sentiment analysis in social networks using social spider optimization algorithm. *Tehnički Vjesnik*, *28*(6), 1943–1951.

Choudrie, J., Banerjee, S., Kotecha, K., Walambe, R., Karende, H., & Ameta, J. (2021). Machine learning techniques and older adults processing of online information and misinformation: a covid 19 study. *Computers in Human Behavior*, *119*, Article 106716.

Choudrie, J., Patil, S., Kotecha, K., Matta, N., & Pappas, I. (2021). Applying and understanding an advanced, novel deep learning approach: A Covid 19, text based, emotions analysis study. *Information Systems Frontiers*, *23*(6), 1431–1465.

Gaikwad, M., Ahirrao, S., Phansalkar, S., & Kotecha, K. (2021). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*, *9*, 48364–48404.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107–116. http://dx.doi.org/10.1142/S0218488598000094.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Imran, A. S., Daudpota, S. M., Kastrati, Z., & Batra, R. (2020). Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets. *IEEE Access*, *8*, 181074–181090. http://dx.doi.org/10.1109/ACCESS.2020.3027350.

Jelodar, H., Wang, Y., Orji, R., & Huang, S. (2020). Deep sentiment classification and topic discovery on novel coronavirus or covid-19 online discussions: Nlp using lstm recurrent neural network approach. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2733–2742. http://dx.doi.org/10.1109/JBHI.2020.3001216.

Liu, R., Shi, Y., Ji, C., & Jia, M. (2019). A survey of sentiment analysis based on transfer learning. *IEEE Access*, *7*, 85401–85412. http://dx.doi.org/10.1109/ACCESS.2019.2925059.

Mansoor, M., Gurumurthy, K., Prasad, V., et al. (2020). Global sentiment analysis of COVID-19 tweets over time. arXiv preprint arXiv:2010.14234.

Miglani, A. (2020). Coronavirus tweets NLP - Text classification, version 1. https://www.kaggle.com/datatattle/covid-19-nlp-text-classification. [Date accessed: 01.11.2020].

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). Covidsenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *IEEE Transactions on Computational Social Systems*, http://dx.doi.org/10.1109/TCSS.2021.3051189.

Nemes, L., & Kiss, A. (2020). Social media sentiment analysis based on COVID-19. *Journal of Information and Telecommunication*, 1–15. http://dx.doi.org/10.1080/24751839.2020.1790793.

Omnicore (2021). Twitter by the numbers- stats, demographics and fun facts. http://www.omnicoreagency.com/twitter-statistics/. [Date accessed: 21.01.2021].

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* EMNLP, (pp. 1532–1543). http://dx.doi.org/10.3115/v1/D14-1162.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos One*, *16*(2), Article e0245909.

Samuel, J., Ali, G., Rahman, M., Esawi, E., Samuel, Y., et al. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), 314. http://dx.doi.org/10.3390/info11060314.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. http://dx.doi.org/10.1109/78.650093.

Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, *404*, Article 132306. http://dx.doi.org/10.1016/j.physd.2019.132306.

Sitaula, C., & Shahi, T. B. (2022). Multi-channel CNN to classify nepali covid-19 related tweets using hybrid features. arXiv preprint arXiv:2203.10286.

Smith, S. (2020). Coronavirus (covid19) tweets, version 19. https://www.kaggle.com/smid80/coronavirus-covid19-tweets. [Date accessed: 01.02.2021].

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Tyagi, P., & Tripathi, R. (2019). A review towards the sentiment analysis techniques for the analysis of twitter data. In *Proceedings of 2nd international conference on advanced computing and software engineering* ICACSE, (pp. 91–95). http://dx.doi.org/10.2139/ssrn.3349569.

Wagle, V., Kaur, K., Kamat, P., Patil, S., & Kotecha, K. (2021). Explainable AI for multimodal credibility analysis: Case study of online beauty health (mis)-information. *IEEE Access*, *9*, 127985–128022.

Washington-Post (2021). Twitter sees record number of users during pandemic, but advertising sales slow. http://www.washingtonpost.com/. [Date accessed: 21.01.2021].

Wu, H., He, Z., Zhang, W., Hu, Y., Wu, Y., & Yue, Y. (2021). Multi-class text classification model based on weighted word vector and bilstm-attention optimization. In *International conference on intelligent computing* (pp. 393–400). Springer.

Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment analysis of comment texts based on BiLSTM. *IEEE Access*, *7*, 51522–51532.

Yildirim, S., Yildirim, G., & Alatas, B. (2021). A new plant intelligence-based method for sentiment analysis: Chaotic sunflower optimization. *Journal of Computer Science*, (Special), 35–40.