Pattern Recognition Letters 164 (2022) 224–231

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

ETCNN: Extra Tree and Convolutional Neural Network-based Ensemble Model for COVID-19 Tweets Sentiment Classification

Muhammad Umer[a], Saima Sadiq[b], Hanen karamti[d], Ala' Abdulmajid Eshmawi[c], Michele Nappi[e,*], Muhammad Usman Sana[f], Imran Ashraf[g,*]

[a] Department of Computer Science & Information Technology, The Islamia University of Bahawalpur, Bahawalpur, 63100, Pakistan
[b] Department of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, Pakistan
[c] Department of Cybersecurity, University of Jeddah, Saudi Arabia
[d] Department of computer sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O.Box 84428, Riyadh 11671, Saudi Arabia
[e] Department of Computer Science, University of Salerno, Fisciano, Italy
[f] College of Computer Science Technology, Xian University of Science and Technology, Xian, Shaanxi 710054, China
[g] Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Korea

ARTICLE INFO

Article history:
Received 30 May 2022
Revised 9 October 2022
Accepted 11 November 2022
Available online 15 November 2022

Edited by Maria De Marsico

Keywords:
Neuroinformatics
COVID-19
Ensemble model
Health informatics
Sentiment analysis

ABSTRACT

Pandemics influence people negatively and people experience fear and disappointment. With the global outspread of COVID-19, the sentiments of the general public are substantially influenced, and analyzing their sentiments could help to devise corresponding policies to alleviate negative sentiments. Often the data collected from social media platforms is unstructured leading to low classification accuracy. This study brings forward an ensemble model where the benefits of handcrafted features and automatic feature extraction are combined by machine learning and deep learning models. Unstructured data is obtained, preprocessed, and annotated using TextBlob and VADER before training machine learning models. Similarly, the efficiency of Word2Vec, TF, and TF-IDF features is also analyzed. Results reveal the better performance of the extra tree classifier when trained with TF-IDF features from TextBlob annotated data. Overall, machine learning models perform better with TF-IDF and TextBlob. The proposed model obtains superior performance using both annotation techniques with 0.97 and 0.95 scores of accuracy using TextBlob and VADER respectively with Word2Vec features. Results reveal that use of machine learning and deep learning models together with a voting criterion tends to yield better results than other machine learning models. Analysis of sentiments indicates that predominantly people possess negative sentiments regarding COVID-19.

1. Introduction

The fast spread and rapid rise in COVID-19 cases led to panic, anxiety, and fear across the globe. Each country faced elevated pressure to control the situation by using all available resources. Besides physical health, the mental health of people has a significant impact on the outbreak of pandemic disease [3]. The effect of COVID-19 on psychological health has been investigated by several studies [18]. Due to the spread of bizarre conspiracies related to COVID-19, social media platforms like Facebook, Instagram, Twitter, Reddit, and others have been monitored to control the spread of disinformation and misinformation. There is a need to devise an analytic way to investigate the public sentiments during the pandemic. Most of the researchers are dealing with healthcare problems, suggesting some preventive measures and recommend post-recovery plans but determining peoples' sentiments as presented on social media is an underexplored research area.

Human sentiments can be observed from social media posts. When data, based on social media posts, is analyzed as a whole it provides predominant thoughts of the public and temperament of the population. The information shared in the news on social media platforms includes public emotions, current trends, fashion, and health problems. Digital media has become a major information source and attracts society to be part of this platform. People now depend more on online platforms than other conventional news sources because it is easily accessible via mobiles. The huge

[*] Corresponding author.
E-mail addresses: umer.sabir@iub.edu.pk (M. Umer), cosc18151070@kfueit.edu.pk (S. Sadiq), hmkaramti@pnu.edu.sa (H. karamti), aaeshmawi@uj.edu.sa (A. Abdulmajid Eshmawi), mnappi@unisa.it (M. Nappi), m.usmansana@uog.edu.pk (M. Usman Sana), imranashraf@ynu.ac.kr (I. Ashraf).

https://doi.org/10.1016/j.patrec.2022.11.012
0167-8655/© 2022 Elsevier B.V. All rights reserved.

volume of data has brought the attention of researchers working in artificial intelligence (AI) and machine learning (ML) domains [40]. These platforms are used by companies as a tool to promote their products and services [16] as per the opinions mined from social media. Similar to users' reviews regarding products and services, people share their views on COVID-19 and such reviews can be used to find and analyze people's sentiments [26]. This also provides an opportunity to analyze the influence of the COVID-19 pandemic worldwide concerning its impact on people's temperaments.

Natural language processing (NLP) based tools and methods have been extensively used by researchers to explore comments on social media. Exploring semantics and the intrinsic meaning of the text is a difficult task, specifically adversarial text [33]. Adversarial text is the modified form of the original text, which is strategically altered to fool a trained classifier. Such modifications include deleting, replacing, or adding salient words to generate meaningful sentences [35]. Suitable feature extraction techniques coupled with the ML model are significant in handling the limitation of text classification and sentiment analysis. People express their sentiments and opinions on Facebook and Twitter about COVID-19. An automated and efficient method is required to explore the meaning of the online text, as manual analysis is long and laborious. Users mostly use wrong punctuation, nonstandard abbreviations, and jargon and emoticons in comments. The absence of voice tone and facial expression put additional complexity in dealing with such text [42].

This research focuses on online text related to COVID-19 for sentiment analysis for obtaining highly accurate sentiments. The main objective of this work is to find the sentiment of the general public to reduce the fear of the outbreak of the disease. Existing studies show poor performance for COVID-19 related Tweets' sentiments due to lack of a proper feature set [15]. This study focuses on resolving this issue and makes the following contributions

- An unlabeled dataset is taken from the IEEE data port that contains tweets regarding COVID-19. Manual annotation is laborious and time-consuming for large datasets, valence-aware dictionary for sentiment reasoning (VADER), and TextBlob is leveraged. From this perspective, the efficiency of both labeling models is investigated.
- Often preferred by the researchers, term frequency (TF) and term frequency-inverse document frequency (TF-IDF) are widely used for sentiment analysis. This study also uses Word2Vec and investigates the efficacy of various machine learning models including Random Forest (RF), Extra Tree (ET), Gradient Boosting Machine (GBM), Logistic Regression (LR), Naive Bayes (NB), Stochastic Gradient (SG) and Voting Classifier (VC) that combines LR and SG.
- A novel voting ensemble of machine learning and deep learning model is designed for sentiment analysis of COVID-19 tweets. The ETCNN combines ET and convolutional neural network (CNN) through soft voting.
- Extensive experiments have been carried out to analyze the efficacy of machine learning models and the performance of the proposed ETCNN is compared with the state-of-the-art models.

The rest of this paper follows this sequence. Related work is detailed in Section 2. It is followed by the description of the proposed model, dataset, and a brief description of ML models in Section 3. Section 4 discusses the results while in the end, the conclusion is provided in Section 5.

2. Related work

Various existing studies have applied ML-based techniques to analyze emotions from the short text commonly called sentiment analysis. Researchers have faced several challenges like unstructured text, size of data, and selection of appropriate techniques for sentiment classification [37]. Twitter is getting the attention of researchers for the tasks like information retrieval, emotion detection, determining public opinion, aggression detection, content mining, and topic modeling related to COVID-19 [11]. Regarding the analysis of COVID-19 related textual data, several different aspects have been explored such as COVID-19 detection [6], the role of the internet of things (IoT) to reduce COVID-19 spread [7], etc. Customers' feedback has been investigated using tweets by researchers [29]. The authors applied the Latent Dirichlet allocation (LDA) method to get a deep insight into the data. Authors performed topic modeling using Euclidean distance based on popularity [36]. Authors analyzed emotions using tweets from various languages [20].

Situations during the pandemic are monitored using public tweets [41]. Emotional attachments and mental condition is analyzed by tweets [8]. Some other researchers focused on the tweets of other languages to observe public sentiments [25]. Some researchers highlighted that tweets are related to a person's mental health [34] and discussed that fear of unemployment and no physical activity cause depression in people [24]. Deep learning models have been used by researchers for text classification [19,21]. Authors applied residual structure based on CNN [1,6,14] to extract local features for image-related tasks such as image segmentation and image classification.

Twitter data has been also used by researchers in tracking and analyzing crisis situations during epidemics [41]. Postnatal behavior or depression of new mothers is analyzed by exploring their emotions, language style, and social involvement from tweets [8]. Authors highlighted government policies during pandemic conditions caused by COVID-19 and also performed topic modeling using multi-lingual Twitter data in [25]. Saire et al. [34] present a positive correlation between infected persons with the number of tweets[38]. analyzed the growth of sinophobia during pandemic conditions from Twitter data.

Researchers are exploring tweets from different perspectives by analyzing the sentiments toward the COVID-19 pandemic. Authors collected tweets over twenty days of March 2020 from Europe and analyze the impact of coronavirus disease spread. The authors applied different unsupervised ML-based models to explore COVId-19-related textual data. Tweets related to COVID-19 have been scrapped by research-ers and analyzed by using diffident methods such as the NB model for analysis [2], LDA, TextBlob and NLTK libraries [22], bigram and trigram [21], etc. Authors analyzed the effect of COVID-19 symptoms on quarantine in [28]. The main focus of these studies was to explore the tweets to analyze COVID-19-related mental and psychological problems.

Other techniques like lexicon-base analysis are performed in [31]. Some researchers performed sentiment analysis of specific regions like China [15]. Some others performed topic modeling and analyzed the reasons for COVID-19- related rumors that can help management to take appropriate actions [9]. Authors claimed that Twitter has a significant impact on the behavior of the general public and information that is discussed in tweets needs to be analyzed [17].

Despite the above-cited studies and their contribution to analyzing the sentiments of COVID-19-related tweets, these studies lack in providing highly accurate results. Often the focus is placed on using different models and feature extraction techniques. This study uses the combination of machine and deep learning models to leverage the benefit of both. The ML model is fed with handcrafted features while CNN uses automatic feature extraction. This way, voting their results produces better results, as discussed in the later sections.

Table 1
Example of different sentiments extracted using VADER & TextBlob from the COVID-19 tweet corpus.

Tweet Text	Sentiment		Score	
	VADER	TextBlob	VADER	TextBlob
#NYC #CoronavirusUSA hospitalizations lowest since the lockdown began	Positive	Positive	0.57	0.64
RT @frequentbuyer1: I often wonder if I'll be among the 30,000 future #CoronavirusUSA deaths. I've already told my husband if I die I want to buried in Alaska	Negative	Negative	-0.61	-0.55

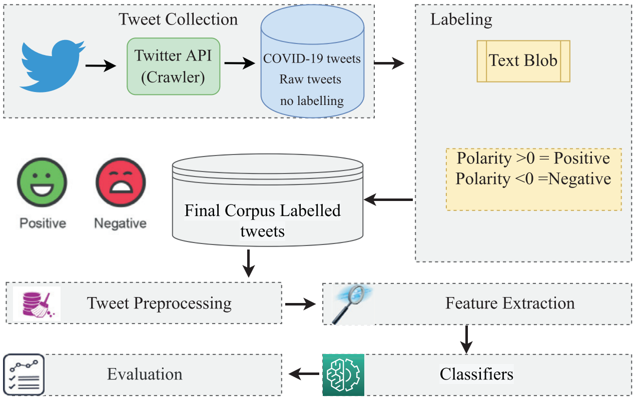

Fig. 1. Workflow of the proposed methodology for COVID-19 sentiment classification.

3. Methodology

The workflow of proposed methodology for COVID-19 sentiment analysis is presented in Fig. 1. Starting with the acquisition of the Tweets dataset, it follows the annotation of data using the TextBlob-based polarity score. The final labeled dataset is preprocessed for noise and redundant data removal. Afterward, feature extraction is performed to train the machine learning models. Models are then evaluated in terms of performance evaluation parameters like accuracy, precision, recall, and F-score.

3.1. Data collection

The COVID-19 tweet dataset [23] is obtained from the IEEE data port. The dataset includes tweet ID and score of tweets' sentiment. Tweets related to COVID-19 are extracted using relevant keywords and hashtags.

3.1.1. Dataset annotation

For data annotation, this study makes use of two well-known methods. TextBlob and VADER [40] are utilized for tweet labeling into two classes that are positive or negative. The polarity score of TextBlob ranges from -1 to 1. A polarity score of less than 0 presents a negative sentiment. A score greater than 0 shows a positive sentiment.

$$Label_{T_i} = \begin{cases} Negative, & P_i < 0 \\ Positive, & P_i > 0 \end{cases} \quad (1)$$

where T_i is the *ith* tweet and P_i represents the polarity of T_i.

Data annotation is done separately for TextBlob and VA- DER indicating that after the annotation, we have two datasets. The polarity score can vary and so do the labels for TextBlob and VADER. It will help to analyze the efficiency of these methods. Samples of dataset text along with the assigned polarities from TextBlob and VADER are given in Table 1.

The dataset contains a total of 11,858 tweets which are annotated using the selected methods. Since the polarity score varies for both these methods, the number of positive and negative tweets also varies. The number of samples for each class using TextBlob and VADER is given in Table 2.

Table 2
Count of tweets for each class using different techniques.

Sr. #	Technique	Positive	Negative	Total
1	TextBlob	7,876	3,982	11,858
2	VADER	7,103	4,755	11,858

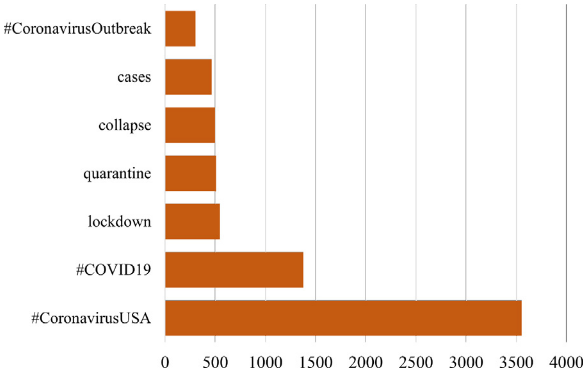

Fig. 2. Top terms in COVID-19 Sentiment Dataset.

3.2. Preprocessing

Tweets are short text and in general, are unstructured. Mostly it contains noisy data and needs to be preprocessed. The following steps are carried out for preprocessing:

1. Hashtags are generally meaningless concerning sentiments. Therefore, these hashtags are removed from the dataset.
2. All text is converted to lowercase.
3. Stopwords also have meaning in expressing sentiment. Therefore, such words are removed from the data to reduce the training time.
4. To decrease the complexity and enhance the training process of models, stemming has also been applied to transform extended words to their root forms.
5. Punctuation, user mentions, and other symbols are also removed from the tweets.

3.3. Graphical visualization and analysis of dataset

To obtain a more clear view of the COVID-19-related dataset, the data is visualized for analysis. In the first step, the top terms of the COVID-19 sentiment dataset are analyzed. People discussed the terms like ”coronavirus”, and ”COVID-19”. The most discussed terms depict the people's interest topic during pandemics. The analysis indicates that the most discussed terms are quarantine and lockdown. Fig. 2 presents the mostly used tweet terms.

Fig. 3 presents the word cloud of two classes separately. The word cloud is given for both negative and positive classes which present the frequently used terms of each category.

Table 3
Strength and weakness of feature representation technique.

Technique	Type	Strengths	Weaknesses
TF	Vectorization technique	-measure the frequency of the most used term in a document -count the occurrence of word appearing	-problem with considering the raw word frequency is that relevance does not increase the proportionality of usage
TF-IDF	Vectorization technique	-Can find similarities between documents easily -Count the frequency of each unique term in a document as well as whole corpus -Weight is directly proportional to the frequency of words in a document and inversely proportional to the frequency of words within documents. -Stop words like is, a, etc., are less significant than rarely occurring words.	-large vector size -Position of term and its co-occurring terms are not considered -Do not consider context and semantics. -Sparsity issue -Inefficient to find similarities between synonyms and differentiate in polysemy words.
Word2Vec	Prediction based technique	-Works on words' probability -Map words to target vectors -CBOW predicts the words' probability and skip-gram determines the words' context.	-Large-sized vocabulary make the model difficult to train on Word2Vec. -Consider word similarities -CBOW Take polysemy words' average, separate vectors are used to present skip-gram.

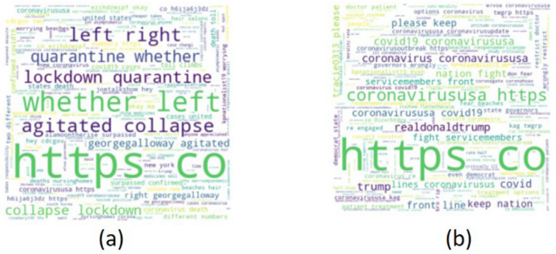

Fig. 3. Word cloud of (a) Positive tweets, and (b) Negative tweets.

3.4. Feature extraction

In this study, the techniques used for feature extraction are TF, TF-IDF, and Word2vec. These techniques have been widely used for solving classification problems. TF shows the frequency of a term in the document. While IDF is the inverse document frequency. Word2vec converts words into numerical vectors. This technique learns word associations from the dataset. The strengths and weaknesses of each of these techniques are discussed in Table 3.

3.5. Classifiers used in study

In this study, ML-based classifiers are applied for sentiment analysis of the tweets. A brief description of these models is presented for completeness.

RF is a tree-based approach, which makes accurate predictions by combining weak learners [5]. It utilizes a bootstrap dataset to train decision trees. The bootstrap is generated as a subset of the original dataset which is randomly selected.

ET is also a tree-based model that uses trees like RF [39]. It does not work on bootstrap data samples and generates trees from the actual dataset. The root node is selected based on the Gini index values.

GBM is based on boosting and the use of classification and regression tasks [10]. It minimizes the error rate by improving the model. It increases the strength of the algorithm at each next step. GBM resolves the problem of missing values efficiently.

LR is a classification model and uses statistical techniques to do that [13]. It uses a logistic function for the binary variables. It uses the sigmoid function to present the relationship between dependent and independent variables.

NB is based on the Bayes' theorem [30]. It works by calculating the prior and posterior probability of a target in the data. It considers feature Independence. It performs better on large-sized and complex multiclass data.

SG works on one versus all techniques [12]. It is based on an optimization technique and finds the most appropriate parameters. It works well on large datasets and uses a large number of samples at each step. Hyperparameter tuning is a sensitive task.

CNN is a deep neural network, widely used for image classification tasks. It efficiently learns the complex features associated with the target class during training [27]. CNN is composed of several layers like convolutional, pooling, activation, and flatten while dropout layers are also used. Features are learned from the input data at the convolutional layer while the pooling layer reduces the size of extracted features that results in low computational complexity. Max pooling is used in this study for experiments. The dropout layer aims at preventing the probability of overfitting while the flatten layer transforms the data into an array. The rectified linear unit (ReLU) is applied as an activation function in this study and the dropout rate is 0.2.

An ensemble model combines the output of more than one machine learning model and often shows better results than individual models and has been used for several kinds of tasks [42]. It determines the final result by incorporating the outputs of multiple models using a soft or hard voting criterion. This study uses two voting classifiers where the first ensemble comprises two classifiers (LR and SG) while the second voting classifier uses an ensemble of two other classifiers (ET and CNN). Hyperparameters of these models are given in Table 4.

3.6. Proposed approach

The proposed approach is called ETCNN which combines ET and CNN models by soft voting, as shown in Fig. 4. The final output will be determined based on high probability. Mathematically, it can be expressed as

$$\widehat{p} = argmax\{\sum_{i}^{n}\mathrm{ET}_i, \sum_{i}^{n}\mathrm{CNN}_i\}. \tag{2}$$

where $\sum_i^n \mathrm{ET}_i$ and $\sum_i^n \mathrm{CNN}_i$ are probabilities for each test sample.

Afterward, the probability scores for each test sample are utilized by soft voting criteria to make the final prediction. Let $Prob_{ET-Pos}$ and $Prob_{ET-Neg}$ be the probability score by the ET model and $Prob_{CNN-Pos}$, and $Prob_{CNN-Neg}$ by the CNN for positive and negative classes, respectively. The average probability for both classes can be computed as:

$$Avg_{Pos} = (Prob_{ET-Pos} + Prob_{CNN-Pos})/2$$
$$Avg_{Neg} = (Prob_{ET-Neg} + Prob_{CNN-Neg})/2$$

Table 4
Hyperparameter setting of ML models.

Classifiers	Parameters
RF	n_estimator=200, max_depth=30, random_state=52
ET	n_estimator=200, max_depth=30, random_state=52
GBM	n_estimator=200, max_depth=30, random_state=52, learning_rate=0.1
LR	penalty='l2', solver='lbfgs'
NB	alpha=1.0, binarize=0.0
SG	penalty='l2', loss='log'
VC(LR+SG)	voting='soft'
CNN	Conv (7 × 7, @64, activation='relu', padding='same'), Max pooling (2 × 2), Conv (7 × 7, @64, activation='relu', padding='same'), GlobalMax pooling (2 × 2), Dropout (0.5), Dense (32 neurons), Softmax (2), Categorical cross entropy

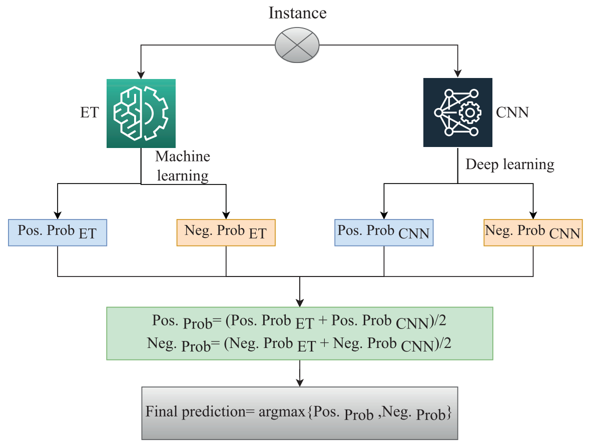
Instance
ET
CNN
Machine learning
Deep learning
Pos. Prob ET
Neg. Prob ET
Pos. Prob CNN
Neg. Prob CNN
Pos. Prob= (Pos. Prob ET + Pos. Prob CNN)/2
Neg. Prob= (Neg. Prob ET + Neg. Prob CNN)/2
Final prediction= argmax{Pos. Prob ,Neg. Prob}

Fig. 4. Framework of the Proposed Voting Classifier (ETCNN).

The final prediction will be positive class as shown below

$$ETCNN = argmax\{Avg_{Pos}, Avg_{Neg}\} \quad (3)$$

The proposed ETCNN decides the final output by combining the predicted probability of classifiers and finds the final output class based on the highest average probability. For performance evaluation of learning classifiers, this work utilizes four evaluation metrics that are accuracy, precision, recall, and F-score.

Algorithm 1 shows the step-by-step working of the proposed ensemble model.

Algorithm 1: Ensembling of ET and CNN (ETCNN).

```
Input: input data $(x, y)_{i=1}^{N}$
$M_{ET}$ = Trained_ ET
$M_{CNN}$ = Trained_ CNN
for $i = 1$ to $M$ do
    if $M_{ET} \neq 0$ & $M_{CNN} \neq 0$ & $tn\_set \neq 0$ then
        $Prob_{CNN-Pos} = M_{CNN-p}(Pos - class)$
        $Prob_{CNN-Neg} = M_{CNN-p}(Neg - class)$
        $Prob_{ET-Pos} = M_{ET-p}(Pos - class)$
        $Prob_{ET-Neg} = M_{ET-p}(Neg - class)$
        $d_f = \max(\frac{1}{N_{classifier}} \sum_{classifier}(Avg_{(Prob_{CNN-Pos}, Prob_{ET-Pos})}, Avg_{(Prob_{CNN-Neg}, Prob_{ET-Neg})}))$
    end if
    Return final label $\hat{p}$
end for
```

Table 5
Comparison of models using TF with TextBlob.

Model	Accuracy	Precision	Recall	F score
RF	0.92	0.93	0.93	0.92
ET	**0.94**	**0.94**	**0.94**	**0.94**
GBM	0.86	0.89	0.87	0.86
LR	0.92	0.93	0.92	0.92
NB	0.88	0.87	0.87	0.88
SG	0.93	0.94	0.94	0.94
VC(LR+SG)	0.92	0.93	0.92	0.92

Table 6
Performance comparison of models with TF-IDF & TextBlob.

Model	Accuracy	Precision	Recall	F score
RF	0.91	0.92	0.92	0.92
ET	**0.94**	**0.95**	**0.95**	**0.95**
GBM	0.86	0.88	0.87	0.86
LR	0.89	0.91	0.90	0.89
NB	0.88	0.89	0.89	0.88
SG	0.94	0.94	0.94	0.94
VC(LR+SG)	0.90	0.91	0.90	0.90

4. Results and discussion

Considerable experiments have been carried out to analyzing COVID-19-related tweets' sentiments. Initially, tweet data was acquired from the repository called IEEE data port. Then it is preprocessed by removing additional noise. Annotated and preprocessed data is used to train the classifiers. To prove the effectiveness of the performance of different ML models and voting ensembles, the data is split into 0.7 to 0.3 ratios for training and testing, respectively. Separate sets of experiments are performed using three feature extraction techniques that are TF, TF-IDF, and Word2vec. Results of ML models by combining different feature extraction techniques are compared for the analysis of COVID-19 tweets' sentiments.

4.1. Comparison of models using textblob with TF and TF-IDF

Results of the classifiers by combining three feature extraction techniques that are TF, TF-IDF, and Word2vec are compared. The performance of models using TF using the TextBlob annotated dataset is shown in Table 5. It can be observed that the ET model shows the best results using TF with a 0.94 accuracy score while SG also achieves good results with a 0.93 accuracy score. ET and SG showed similar results regarding the precision, recall, and F score each with a score of 0.94. NB and GBM perform poorly in analyzing sentiments of COVID-19 tweets. GBM achieves 0.86 accuracy, 0.89 precision, 0.87 recall and 0.86 F-score while NB achieve 0.8878 accuracy value, 0.87 precision, 0.87 recall and 0.88 F-score.

The results of supervised ML classifiers using TF-IDF are presented in Table 6. Performance results show that the performance of ET and SG is slightly improved by using TF-IDF. SG showed the

Table 7
Performance comparison of models with TF & VADER.

Model	Accuracy	Precision	Recall	F score
RF	0.91	0.91	0.91	0.91
ET	**0.92**	**0.93**	**0.93**	**0.93**
GBM	0.85	0.86	0.85	0.83
LR	0.91	0.92	0.92	0.91
NB	0.88	0.88	0.88	0.87
SG	0.92	0.92	0.92	0.92
VC(LR+SG)	0.91	0.92	0.92	0.92

Table 8
Performance comparison of models using TF-IDF and VADER.

Model	Accuracy	Precision	Recall	F score
RF	0.89	0.90	0.90	0.89
ET	**0.93**	**0.93**	**0.93**	**0.93**
GBM	0.85	0.86	0.85	0.83
LR	0.88	0.89	0.88	0.87
NB	0.88	0.89	0.89	0.88
SG	0.92	0.92	0.92	0.92
VC(LR+SG)	0.88	0.89	0.88	0.87

second-highest value of accuracy score. Similarly, the results of NB and GBM have been increased by TF-IDF. These two classifiers have shown poor performance as compared to other classifiers used in the experiments.

4.2. Comparison of ML-based models with VADER using TF and TF-IDF

A separate set of experiments are carried out for models using data annotated using VADER with TF and TF-IDF. VADER is based on the lexicon and is commonly used to find the texts' sentiment. VADER has performed well on short text classification problems in various types of research works [4].

The performance comparison of classifiers using TF for analyzing sentiments of COVID-19 tweets is presented in Table 7. It can be noticed from the performance comparison that TF in combination with VADER is poorer than the results achieved with TextBlob. ET achieved the best results with a 0.92 accuracy score and 0.93 precision, recall, and F-score. GBM does not show any improvement even with VADER and shows 0.85 scores for accuracy. SG and VC(LR+SG) show a similar result with a 0.92 score each for precision, recall, and F-score.

Table 8 presents the result of models utilizing TF-IDF with VADER. Clearly, the ET classifier shows a little better performance of 1% increase with a 0.93 accuracy score while other models including RF, GBM, LR, and Voting Classifier show a lowering of performance when used with TF-IDF on VADER annotated dataset.

Results indicate that the models show better performance using TF-IDFin combination with TextBlob annotated dataset. It can be noticed that ET outperforms all other classifiers and can find the tweets' sentiment with 0.94 accuracy utilizing TF-IDF in combination with TextBlob. The precision, recall, and F score value is 0.95.

4.3. Comparison of models using word2vec with textblob

Furthermore, the efficiency of models has been computed and compared utilizing Word2vec for COVID-19-related tw- eets' sentiment analysis. The results shown in Table 9 reveal that models have not shown improved results. ET achieves the highest accuracy with a 0.88 value which is lesser than the accuracy result attained by the ET model with TF and TF-IDF with TextBlob. It can be seen that Word2vec has not improved the results of models when using the TextBlob annotated dataset. The highest F-score using word2vec and TextBloe is 0.88 by the ET classifier which is lower than the F-score achieved utilizing TF-IDF with Textblob.

Table 9
Performance comparison of models using Word2vec with TextBlob.

Model	Accuracy	Precision	Recall	F-score
RF	0.87	0.88	0.87	0.87
ET	0.88	0.90	0.89	0.88
GBM	0.84	0.85	0.84	0.84
LR	0.82	0.82	0.82	0.82
NB	0.68	0.72	0.69	0.70
SG	0.83	0.83	0.83	0.83
VC(LR+SG)	0.82	0.82	0.82	0.82
ETCNN	**0.97**	**0.95**	**0.96**	**0.95**

Table 10
Performance comparison of ML models using Word2vec with VADER.

Model	Accuracy	Precision	Recall	F-score
RF	0.87	0.88	0.88	0.87
ET	0.88	0.88	0.89	0.88
GBM	0.85	0.85	0.85	0.84
LR	0.82	0.81	0.82	0.81
NB	0.66	0.74	0.66	0.68
SG	0.82	0.82	0.83	0.81
VC(LR+SG)	0.82	0.81	0.82	0.81
ETCNN	**0.95**	**0.92**	**0.92**	**0.92**

The proposed ETCNN model outperforms all other models with a 0.97 accuracy score, 0.95 precision, 0.96 recall, and 0.95 F-score with TextBlob annotated and using Word2Vec features. The proposed ensemble of machine learning and deep learning models performs better using Word2vec features and obtain the highest accuracy for the current task of sentiment classification.

4.4. Comparison of models using word2vec with VADER

For analyzing the performance of the models, the results of models using Word2vec are compared with the models using VADER. The performance comparison is shown in Table 10 which indicates that classifiers have shown poor performance using Word2vec in combination with VADER when compared with Word2vec in combination with TextBlob data. ET performed best among all with a 0.88 accuracy score, 0.88 precision, 0.89 recall, and 0.88 F-score. If we compare the effectiveness of ML models, ET is dominant among all other ML models utilized in this study. However, its performance is surpassed by the proposed ETCNN which achieves the best results even with Word2Vec with VADER data and outperforms other models with a 0.95 accuracy score. Similarly, values for other performance evaluation parameters show superior performance for sentiment classification.

The ETCNN surpassed every combination of features technique with models when it is applied using Word2vec with TextBlob. Finally, the results prove the effectiveness of the ensemble model by combining machine learning and deep learning approach with appropriate feature representation techniques. Computing the probability of the target output using machine learning and deep learning classifiers individually and then finalizing the target class with the highest probability improves the performance as compared to the separate models. The deep neural network of ETCNN makes it accurate and more efficient.

4.5. Comparison with existing studies

We also carried out a performance comparison of the proposed approach with existing studies. For this purpose, we selected

Table 11
Performance comparison with existing studies.

Ref	Year	Model	Feature	Accuracy
[32]	2022	ETC	TF-IDF + BoW	0.93
[32]	2022	RF, XGboost	TF-IDF + BoW	0.92
Current study	2022	ETCNN	Word2Vec + TextBlob	0.97

[32] as the study uses the same dataset for sentiment classification. The study performed experiments by combining TF-IDF and BoW to obtain higher accuracy with an extra tree classifier (ETC). Table 11 shows the comparison of the proposed approach with [32]. The study reports an accuracy of 0.93 with ETC while RF and XGboost obtained an accuracy of 0.92 for sentiment classification. The current study, on the other hand, obtains a far better accuracy for sentiment classification using the ensemble ETCNN.

5. Conclusion

Pandemics negatively influence people and can lead to several mental problems. Analyzing the sentiments of people and devising corresponding policies can reduce the threats to mental health. With the wide use of social media platforms like Twitter, Facebook, Instagram, etc., the use of social media data presents a potential opportunity to analyze sentiments. However, the unstructured nature of such data may lead to low classification accuracy. This study proposes a novel approach an ensemble of ET and CNN models to overcome this limitation. A large unstructured dataset is obtained, preprocessed, and annotated using TextBlob and VADER for experiments. In addition, the efficacy of TF, TF-IDF, and Word2Vec is also evaluated regarding their use with different machine learning models. Extensive experiments are performed which reveal that the ET model shows the best performance among the machine learning models when TF-IDF features are used. The proposed model obtains the best performance with a 0.97 accuracy score for Word2Vec features using TextBlob and a 0.95 accuracy score for Word2Vec features using VADER annotated dataset. The combination of machine learning and deep learning models seems to work well to obtain high accuracy for sentiment classification. Even with a medium-sized dataset, higher classification accuracy is possible if an appropriate combination of learning algorithms is used. Selecting an appropriate feature extraction approach helps to obtain better performance, however, only three feature approaches are investigated. Further experiments are needed with global vectors for word representation, continuous BoW, etc. This study uses TextBlob and VADER as base models for annotation, however, the manual and other annotation approaches need to be implemented for further experiments.

Declaration of Competing Interest

The authors declare no conflict of interest.

CRediT authorship contribution statement

Muhammad Umer: Writing – original draft, Methodology, Software. **Saima Sadiq:** Writing – original draft, Conceptualization. **Hanen karamti:** Software, Project administration. **Ala' Abdulmajid Eshmawi:** Resources, Supervision, Formal analysis. **Michele Nappi:** Writing – review & editing, Supervision. **Muhammad Usman Sana:** Project administration, Software. **Imran Ashraf:** Writing – review & editing, Methodology.

Data Availability

Data will be made available on request.

Acknowledgement

Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R192), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

[1] M. Ahmad, S. Sadiq, A.S. Alluhaidan, M. Umer, S. Ullah, M. Nappi, et al., Industry 4.0 technologies and their applications in fighting COVID-19 pandemic using deep learning techniques, Comput. Biol. Med. 145 (2022) 105418.

[2] M. Alhajji, A. Al Khalifah, M. Aljubran, M. Alkhalifah, Sentiment analysis of tweets in saudi arabia regarding governmental preventive measures to contain COVID-19(2020).

[3] R. Bhat, V.K. Singh, N. Naik, C.R. Kamath, P. Mulimani, N. Kulkarni, Covid 2019 outbreak: the disappointment in indian teachers, Asian J. Psychiatr. 50 (2020) 102047.

[4] V. Bonta, N.K.N. Janardhan, A comprehensive study on lexicon based approaches for sentiment analysis, Asian J. Comput. Sci. Technol. 8 (S2) (2019) 1–6.

[5] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[6] A. Castiglione, P. Vijayakumar, M. Nappi, S. Sadiq, M. Umer, Covid-19: automatic detection of the novel coronavirus disease from ct images using an optimized convolutional neural network, IEEE Trans. Ind. Inf. (2021).

[7] A. Castiglione, M. Umer, S. Sadiq, M.S. Obaidat, P. Vijayakumar, The role of internet of things to control the outbreak of COVID-19 pandemic, IEEE Internet Things J. (2021).

[8] M. De Choudhury, S. Counts, E. Horvitz, Predicting postpartum changes in emotion and behavior via social media, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2013, pp. 3267–3276.

[9] A. Depoux, S. Martin, E. Karafillakis, R. Preet, A. Wilder-Smith, H. Larson, The pandemic of social media panic travels faster than the COVID-19 outbreak, 2020.

[10] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. (2001) 1189–1232.

[11] K. Garcia, L. Berton, Topic detection and sentiment analysis in twitter content related to COVID-19 from brazil and the USA, Appl. Soft Comput. 101 (2021) 107057.

[12] W.A. Gardner, Learning characteristics of stochastic-gradient-descent algorithms: a general study, analysis, and critique, Signal Process. 6 (2) (1984) 113–133.

[13] A. Genkin, D.D. Lewis, D. Madigan, Large-scale bayesian logistic regression for text categorization, Technometrics 49 (3) (2007) 291–304.

[14] U. Hafeez, M. Umer, A. Hameed, H. Mustafa, A. Sohaib, M. Nappi, H.A. Madni, A CNN based coronavirus disease prediction system for chest x-rays, J. Ambient. Intell. Humaniz. Comput. (2022) 1–15.

[15] X. Han, J. Wang, M. Zhang, X. Wang, Using social media to mine and analyze public opinion related to COVID-19 in china, Int. J. Environ. Res. Public Health 17 (8) (2020) 2788.

[16] W. He, H. Wu, G. Yan, V. Akula, J. Shen, A novel social media competitive analytics framework with sentiment benchmarks, Inform. Manag. 52 (7) (2015) 801–812.

[17] B. Huang, K.M. Carley, Disinformation and misinformation on twitter during the novel coronavirus outbreak, arXiv preprint arXiv:2006.04278 (2020).

[18] A. Imran, W.S. Alnumay, A. Rashid, S. Hur, B. Ali Kashif, Z. Yousaf Bin, Prediction models for covid-19 integrating age groups, gender, and underlying conditions, Comput. Mater. Continua (2021) 3009–3044.

[19] A. Ishaq, M. Umer, M.F. Mushtaq, C. Medaglia, H.U.R. Siddiqui, A. Mehmood, G.S. Choi, Extensive hotel reviews classification using long short term memory, J. Ambient. Intell. Humaniz. Comput. 12 (10) (2021) 9375–9385.

[20] V.K. Jain, S. Kumar, S.L. Fernandes, Extraction of emotions from multilingual text using intelligent text processing and computational linguistics, J. Comput. Sci. 21 (2017) 316–326.

[21] M. Karim, M.M.S. Missen, M. Umer, S. Sadiq, A. Mohamed, I. Ashraf, Citation context analysis using combined feature embedding and deep convolutional neural network model, Appl. Sci. 12 (6) (2022) 3203.

[22] C. Kaur, A. Sharma, Twitter Sentiment Analysis on Coronavirus using Textblob, Technical Report, EasyChair, 2020.

[23] R. Lamsal, Design and analysis of a large-scale COVID-19 tweets dataset, Appl. Intell. (2020) 1–15.

[24] I. Li, Y. Li, T. Li, S. Alvarez-Napagao, D. Garcia-Gasulla, T. Suzumura, What are we depressed about when we talk about COVID-19: Mental health analysis on tweets using natural language processing, in: International Conference on Innovative Techniques and Applications of Artificial Intelligence, Springer, 2020, pp. 358–370.

[25] C.E. Lopez, M. Vasu, C. Gallemore, Understanding the perception of COVID-19 policies by mining a multilanguage twitter dataset, arXiv preprint arXiv:2003.10359 (2020).

[26] M.V. Mäntylä, D. Graziotin, M. Kuutila, The evolution of sentiment analysisa review of research topics, venues, and top cited papers, Comput. Sci. Rev. 27 (2018) 16–32.

[27] K. O'Shea, R. Nash, An introduction to convolutional neural networks, arXiv preprint arXiv:1511.08458 (2015).

[28] C.K. Pastor, Sentiment analysis of filipinos and effects of extreme community quarantine due to coronavirus (covid-19) pandemic, Available at SSRN 3574385 (2020).

[29] L. Pépin, P. Kuntz, J. Blanchard, F. Guillet, P. Suignard, Visual analytics for exploring topic long-term evolution and detecting weak signals in company targeted tweets, Comput. Ind. Eng. 112 (2017) 450–458.

[30] A. Perez, P. Larranaga, I. Inza, Supervised classification with conditional gaussian networks: increasing the structure complexity from naive bayes, Int. J. Approx. Reason. 43 (1) (2006) 1–25.

[31] D. Prabhakar Kaila, D.A.V. Prasad, et al., Informational flow on twitter–corona virus outbreak–topic modelling approach, Int. J. Adv. Res. Eng. Technol. (IJARET) 11 (3) (2020).

[32] F. Rustam, M. Khalid, W. Aslam, V. Rupapara, A. Mehmood, G.S. Choi, A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis, PLoS ONE 16 (2) (2021) e0245909.

[33] Z. Saeed, R.A. Abbasi, O. Maqbool, A. Sadaf, I. Razzak, A. Daud, N.R. Aljohani, G. Xu, Whats happening around the world? a survey and framework on event detection techniques on twitter, J. Grid Comput. 17 (2) (2019) 279–312.

[34] J.E.C. Saire, R.C. Navarro, What is the people posting about symptoms related to coronavirus in bogota, colombia? arXiv preprint arXiv:2003.11159 (2020).

[35] S. Samanta, S. Mehta, Generating adversarial text samples, in: European Conference on Information Retrieval, Springer, 2018, pp. 744–749.

[36] J. Samuel, M. Garvey, R. Kashyap, That message went viral?! exploratory analytics and sentiment analysis into the propagation of tweets, arXiv preprint arXiv:2004.09718 (2020).

[37] J. Samuel, R. Kashyap, S. Betts, Strategic directions for big data analytics in e-commerce with machine learning and tactical synopses: propositions for intelligence based strategic information modeling (SIM), J. Strategic Innovat. Sustain. (2018).

[38] L. Schild, C. Ling, J. Blackburn, G. Stringhini, Y. Zhang, S. Zannettou, " Go eat a bat, chang!": an early look on the emergence of sinophobic behavior on web communities in the face of covid-19, arXiv preprint arXiv:2004.04046 (2020).

[39] A. Sharaff, H. Gupta, Extra-tree Classifier with Metaheuristics Approach for Email Classification, in: Advances in Computer Communication and Computational Sciences, Springer, 2019, pp. 189–197.

[40] M. Umer, I. Ashraf, A. Mehmood, S. Ullah, G.S. Choi, Predicting numeric ratings for google apps using text features and ensemble learning, ETRI J. (2020).

[41] X. Ye, S. Li, X. Yang, C. Qin, Use of social media for the detection and analysis of infectious diseases in china, ISPRS Int. J. Geoinf. 5 (9) (2016) 156.

[42] A. Yousaf, M. Umer, S. Sadiq, S. Ullah, S. Mirjalili, V. Rupapara, M. Nappi, Emotion recognition by textual tweets classification using voting classifier (LR-SGD), IEEE Access (2020).