Knowledge-Based Systems 226 (2021) 107126

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

TClustVID: A novel machine learning classification model to investigate topics and sentiment in COVID-19 tweets

Md. Shahriare Satu [a], Md. Imran Khan [b], Mufti Mahmud [c], Shahadat Uddin [d], Matthew A. Summers [e], Julian M.W. Quinn [e], Mohammad Ali Moni [e,f,*]

[a] Department of Management Information Systems, Noakhali Science & Technology University, Noakhali, 3814, Bangladesh
[b] Department of Computer Scienc & Engineering, Gono Bishwabidyalay, Savar, Dhaka, 1344, Bangladesh
[c] Department of Computer Science, and Medical Technology Innovation Facility, Nottingham Trent University, Clifton Campus, Clifton, Nottingham – NG11 8NS, UK
[d] Complex Systems Research Group, Faculty of Engineering, The University of Sydney, Darlington, NSW 2008, Australia
[e] The Garvan Institute of Medical Research, Healthy Ageing Theme, Darlinghurst, NSW 2010, Australia
[f] WHO Collaborating Centre on eHealth, UNSW Digital Health, School of Public Health and Community Medicine, Faculty of Medicine, University of New South Wales, Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 2 August 2020
Received in revised form 1 May 2021
Accepted 3 May 2021
Available online 6 May 2021

Keywords:
COVID-19
Twitter data
Machine learning
TClustVID
Classification
Topics modeling

ABSTRACT

COVID-19, caused by SARS-CoV2 infection, varies greatly in its severity but presents with serious respiratory symptoms with vascular and other complications, particularly in older adults. The disease can be spread by both symptomatic and asymptomatic infected individuals. Uncertainty remains over key aspects of the virus infectiousness (particularly the newly emerging variants) and the disease has had severe economic impacts globally. For these reasons, COVID-19 is the subject of intense and widespread discussion on social media platforms including Facebook and Twitter. These public forums substantially influence public opinions and in some cases can exacerbate the widespread panic and misinformation spread during the crisis. Thus, this work aimed to design an intelligent clustering-based classification and topic extracting model named TClustVID that analyzes COVID-19-related public tweets to extract significant sentiments with high accuracy. We gathered COVID-19 Twitter datasets from the IEEE Dataport repository and employed a range of data preprocessing methods to clean the raw data, then applied tokenization and produced a word-to-index dictionary. Thereafter, different classifications were employed on these datasets which enabled the exploration of the performance of traditional classification and TClustVID. Our analysis found that TClustVID showed higher performance compared to traditional methodologies that are determined by clustering criteria. Finally, we extracted significant topics from the clusters, split them into positive, neutral and negative sentiments, and identified the most frequent topics using the proposed model. This approach is able to rapidly identify commonly prevailing aspects of public opinions and attitudes related to COVID-19 and infection prevention strategies spreading among different populations.

© 2021 Published by Elsevier B.V.

1. Introduction

COVID-19 has become a global concern as a major and dangerous public health threat. The World Health Organization (WHO) declared COVID-19 a Public Health Emergency of International Concern (PHEIC) on February 28, 2020. During the 1960s various coronaviruses were identified as infectious to humans in the upper respiratory tract, notably human coronavirus 229E and OC43 [1]. Numerous coronaviruses may circulate in wild mammalian populations, with some causing minor human health problems. However, this picture changed with the emergence of severe acute respiratory syndrome (SARS-CoV) at 2002 and the Middle East Respiratory Syndrome coronavirus (MERS-CoV) at 2012 that infect respiratory tract epithelial tissues to cause serious and often deadly respiratory disease [2]. Pandemic coronavirus SARS-CoV2 causes the pandemic disease COVID-19 that shows flu and pneumonia-like symptoms with cardiovascular complications with severity ranging from undetectable to rapidly lethal. The spread of this disease has been causing huge economic disruption, personal health fears, and uncertainties that have dominated both the news and social media.

The massive use of web and mobile technologies gives opportunities to the population to share their opinions on social media platforms such as Facebook and Twitter. Emotion plays a

* Corresponding author at: The Garvan Institute of Medical Research, Healthy Ageing Theme, Darlinghurst, NSW 2010, Australia.
E-mail address: m.moni@unsw.edu.au (M.A. Moni).

https://doi.org/10.1016/j.knosys.2021.107126
0950-7051/© 2021 Published by Elsevier B.V.

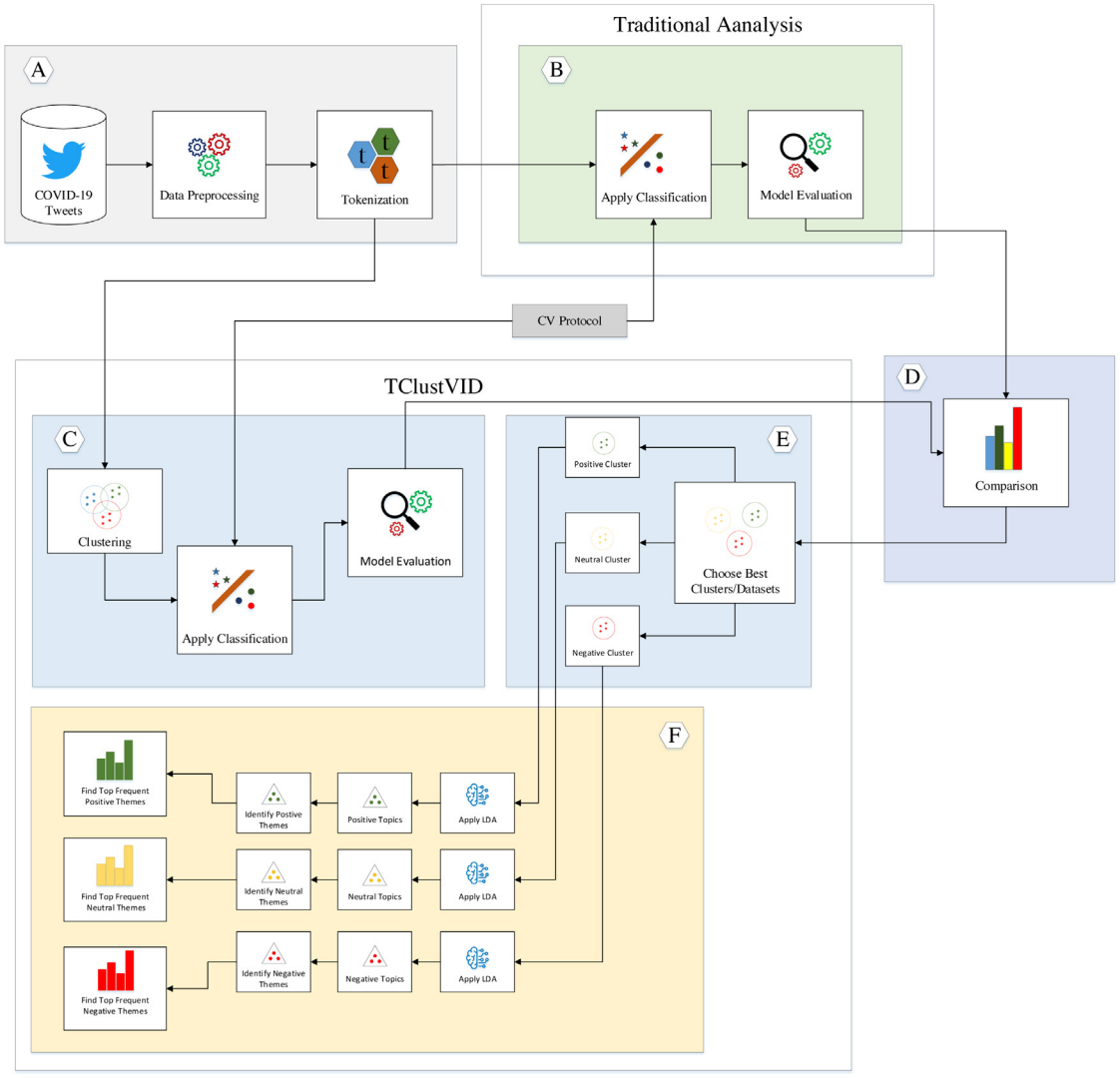

Fig. 1. Details of working methodology where A. Data preprocessing B. Traditional classification and evaluation C. Clustering, classification and evaluation D. Comparison the outcomes between traditional and TClustVID E. Select the best clusters/datasets and Identify positive, neutral and negative clusters F. Extract topics by LDA and represent top frequent topics from it.

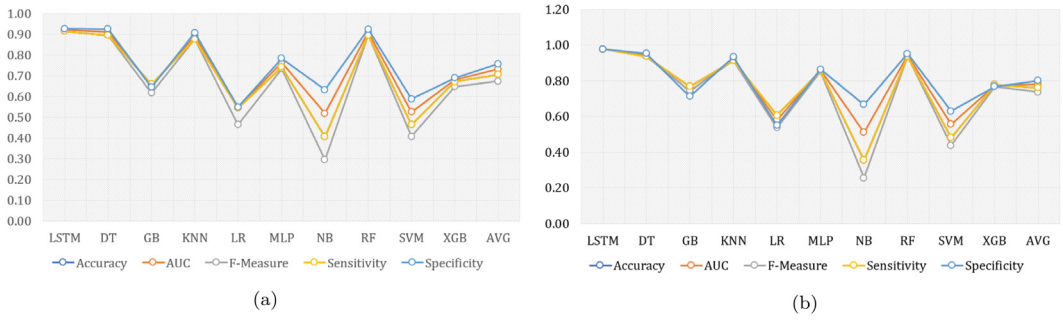

Fig. 2. Average performance of various classifiers for evaluating them using (a) traditional way (b) TClustVID corresponding to the nine twitter experimental datasets.

significant role in conducting effective human-to-human communication and provides major effort to take proper decisions [3]. Text is one of the essential components for affective computing as most of the people use text message/sms using computer to express their pinion [4]. During the COVID-19 pandemic, various social media have been used to communicate daily activities and thoughts, including many significant messages (texts) left by users sharing their general feelings about their personal situation, health status, tips to stay well, and other related information [5]. Such messages may provide large-scale insights into behavioral responses to the pandemic. However, it is not easy to judge whether various social media carries important information, not least because semantic abstruseness makes it hard to understand many messages. Nevertheless, machine learning and computational methods have increasingly been used to scrutinize social media data in the biomedical sector [6]. Content relating to

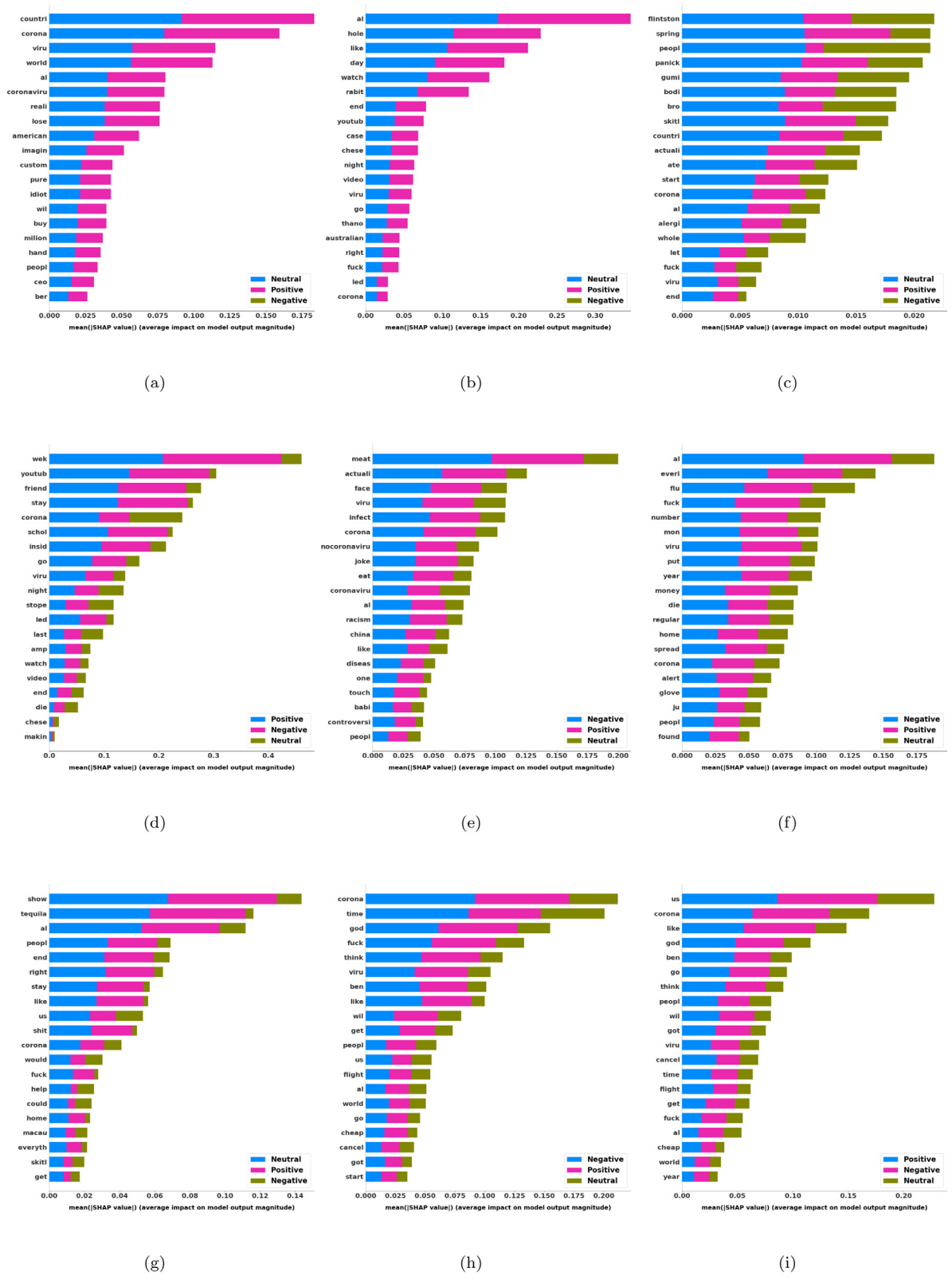

Fig. 3. Compute SHAP values to determine COVID-19 (a) Positive (b) Neutral (c) Negative topics.

COVID-19 may be useful to extract significant information for individuals and policy-makers. Twitter, in particular, is a popular micro-blogging and public networking service widely used for messaging and posting [7]. Automatic classification of tweets into particular classes is challenging, not least because these messages are short, 140 characters, or less [8]. In recent years, sentiment analysis is useful to process social media data like blogs, wikis, micro-blogging and other online collaborative media [9]. It is a branch of affective computing that classify as text either positive or negative. So, the analysis requires identification of

sentiments in Twitter messages (tweets) which contain abbreviations, spelling variations and ambiguous or informal language. The objective of this work is to investigate the type of tweets being communicated and to extract information on significant topics that are useful to understand the COVID-19 pandemic situation.

In this study, we collected several Twitter datasets and investigated sentiment topics related to COVID-19 by designing a novel machine learning model named TClustVID. This model was used to explore significant subsets using clustering methods and select them by verifying high classification performance. Each of these tweet clusters was split into the positive, negative and neutral group, and employed latent dirichlet allocation (LDA) to extract key topics. We then interpreted and identified more significant topics. This methodology can be used to generate relevant information on public and human social behavior dealing with COVID-19 issues for researchers and policymakers. The key contributions of this work are described briefly as follows:

- In TClustVID, we have incorporated clustering and classification to facilitate the extraction of significant topics concerning the pandemic.
- Multiple tweet datasets were used to verify the results of the proposed model in primary and different clusters.
- Significant topics were represented using various word clouds that render them more visible and understandable.
- The identification of the most frequently raised topics can make awareness of the underlying matters, particularly related to widespread concern.

2. Literature review

Affective computing and sentiment analysis is the key to the advancement of artificial intelligence. It has a great potentiality to become a sub-component technology for other systems [3]. Sentiment analysis is broadly categorized into symbolic and subsymbolic approaches [10]. Popular sources of affect words are created knowledge bases to identify polarity text e.g., WordNet-Affect, SentiWordNet, SenticNet. Therefore, the integration of logical reasoning was happened with deep learning in SenticNet6 to infer polarity of text [4,10]. Dragoni et al. [9] proposed commonsense ontology based on SenticNet that supports word embedding, domain information and polarity representation for sentiment analysis. Poria et al. [11] provided three deep learning based architectures where different facets of analysis to be considered for multimodal sentiment analysis. Chaturvedi et al. [12] introduced a convolutional fuzzy commonsense reasoning model which projects features into four dimensional space in order to increase classification performance. Jiang et al. [13] proposed joint-aspect level sentiment modification which trained aspect-specific sentiment words extraction and aspect-level sentiment transformation modules. Baired et al. [14] presented a lexical knowledge base approach where SenticNet was used to explore natural language concept and fine tune various feature types from the large scale multimodel dataset. Besides, several NLP works were performed based on the knowledge-based and statistical methods are combined for investigating short messaging, microblogging (e.g., Twitter) sentiment analysis. Khatua et al. [15] represented their work in the context of 2014 Ebola and 2016 Zika outbreaks where they suggest domain-specific word vectors are better than pre-trained Word2Vec (contrived from Google News) or Global Vector for Word Representation of Stanford NLP group (GloVe). Ahmed et al. [16] provided a query expansion model that accelerates the initial queries with expansion terms. In this case, various word embedding models such as Word2Vec, GloVe, and fastText are trained tweet corpus. Behera et al. [17] proposed a hybrid model combining convolutional neural network (CNN) and long short term memory (LSTM) called Co-LSTM, which is highly adaptable with big social data.

Alike recent relevant works of sentiment analysis, some recent studies have been attempted to scrutinize COVID-19 tweets in bulk for public health research purposes, although it is likely that they have been mined for more commercial purposes. Aljameel et al. [18] gathered 2,42,525 tweets from five regions in Saudi Arabia to analyze their sentiments using support vector machine (SVM), k-nearest neighbor (KNN) and Naïve Bayes (NB). Alomari et al. [19] investigated 14 million tweets where they extracted significant features using TF–IDF based correlation analysis and explored relevant topics using LDA. Al-rakmi et al. [20] gathered 4,00,000 tweets and implemented entropy and correlation based feature selection and ensemble methods using NB, Bayes Net, KNN, C4.5, random forest (RF) and SVM. Boot-Itt and Skunkan [21] explored 1,09,990 tweets to analyze their sentiments using NRC sentiments lexicon and LDA. Gencoglu et al. [22] investigated 26 million tweets using language agnostic BERT sentence embedding models and further classified sentiments using KNN, LR and Bayesian hyperparameter optimization. Kouzy et al. [23] explored tweets using 14 trending hashtags and keywords about COVID-19 and investigated the magnitude of misinformation by comparing terms and hashtags. Kaur et al. [24] translated 16,138 tweets into English and scrutinized sentiments and emotions using TextBlob and IBM Tone analyzer, respectively. Medford et al. [25] gathered all twitter user data from January 14th to 28th, 2020 and investigated sentiments and explored topics using LDA. Mackey et al. [26] collected 4,492,954 tweets from the United States, United Kingdom, India and Australia where they extracted topics using biterm topics model (BTM) with topics clusters. Nemes and Kiss [27] analyzed tweets using TextBlob and RNN. Samual et al. [28] investigated 9000 tweets and got non-textual variables using N-Gram and further analyzed sentiments using NB, Linear regression, LR and KNN. Xiang et al. [29] gathered 82,893 tweets for sentiment analysis and topics modeling using NRC Lexicon and LDA respectively. Xue et al. [30] extracted 4 million English language tweets using N-Gram, NRC Lexicon and LDA analysis. Also, they [31] scrutinized 1.9 million English language tweets using machine learning models and LDA. Yin et al. [32] utilized 13 million tweets by inspected them using VADER and dynamic LDA model. Zhang et al. [33] perused tweets by employing N-Gram model and TF–IDF as well as explored sentiments using DT, LR, KNN, RF and SVM respectively.

2.1. Drawbacks of previous works

There are few observed issues and potential pitfalls in interpreting recently published work. Most have not proposed a framework for the investigation of tweets and employ both sentiment analysis and topic modeling. In addition, many works have specified their analysis as specialized to particular regions or languages, and cannot easily generalize those approaches globally. For sentiment classification, a small number of machine learning methods have been implemented as well as verified their results with only a small number of evaluation metrics. Most times, they focused on the specific issues (e.g., psychological or human needs). However, they did not extract the most significant topics needed to realize this pandemic situation by individuals. In studies using topic modeling, positive, negative and neutral topics were not specified in their work. Hence it is difficult to gain an understanding of the current situation of pandemic according to this perspective. The details visualization of topics orientation was given in this work.

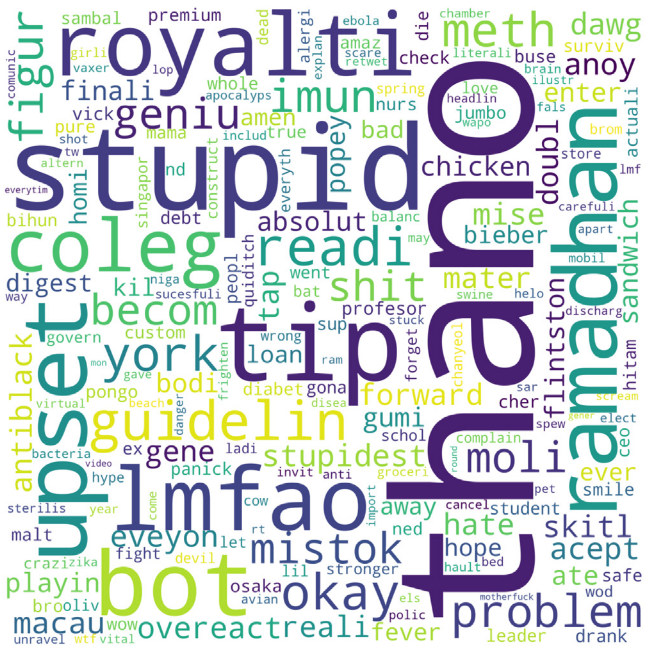

Fig. 4. Word cloud of various topics.

3. Materials and methods

We proposed a machine learning based COVID-19 tweet analytical model that can be used to explore significant topics from Twitter datasets. To process them, different natural language processing techniques are used along with machine learning methods as illustrated in Fig. 1. The working project is provided at the following link https://github.com/shahriariit/COVID-19-Twitter-Data-Analysis.

3.1. Data description

The COVID-19 Twitter datasets has been collected from the IEEE Data portal that originated from the LSTM model, developed by Rabindra Lamsal, who monitors the real-time feeds of COVID-19-related tweets [34]. It generates over 0.3 million requests every 24 h and its time-series graph is updated at every 30 s. Almost 16 million tweets were identified before March 20th 2020. Each database (*.db) contains three attributes where the first, second, and third columns have been indicated date and time, tweets, and sentiment scores, respectively. However, these sentiment scores have been manipulated within the range [0, 2] where the most negative, neutral, and positive sentiments are indicated as 0, 1 and 2, respectively. Eight twitter datasets (corona_tweets_1M.db, corona_tweets_1M_2, corona_tweets_1M, corona_tweets_2L, corona_tweets_2M.db, corona_tweets_2M_2, corona_tweets_2M_3 and corona-_tweets_3M) have been investigated and deemed suitable models to classify tweets in this study. Each dataset has been represented as the tweets related to COVID-19 of each day before March 20th 2020. We gathered datasets of a couple of days to understand and extract various topics everyday. The first seven of these datasets are denoted as dataset-1, dataset-2, dataset-3, dataset-4, dataset-5, dataset-6, and dataset-7. In this study, corona_tweets_3M was split into dataset-8 and dataset-9 because the computational cost is manipulated very high for the corona_tweets_3M.

3.2. Data preprocessing

In the preprocessing steps, different twitter datasets have been prepared for manipulation. These types of tweets contain various HTML tags, punctuation, numbers, single characters and multiple spaces. Several functions were used to clean datasets in this step. The symbols '⟨⟩' were replaced with empty spaces. Again, every single character which does not indicate any meaningful communication was replaced with space respectively. Finally, all multiple spaces were removed from these tweets. This process was employed in the nine twitter datasets and combined for further analysis. Table 1 represents the number of tweets before and after prepossessing steps.

3.3. Tokenization

After the pre-processing steps, tokenization procedures were used to generate a word-to-index dictionary whereby each word is created as a key in the corpus. Hence, the corresponding unique index has been indicated the value of the keys. In the training phase, each list is held on each sentence where the size is dissimilar. Thus, the maximum length of the list is fixed. If the length of any list is exceeded, it is truncated into the maximum permitted length. Zeros are added to the endpoint of a shortlist until it reaches a maximum length, a process is termed padding. Employing word embedding is useful to extract significant words and investigate similarity along with semantic relations more precisely. Pennington et al. [35] proposed a certain weighted least squares model that trains and counts global word-word co-appearance for efficient statistical usage. This is called GloVe that is also publicly available [15]. Thus, this embedding word vector has been used to create a dictionary that holds a word as a key and the corresponding list as values [36]. Finally, an embedding matrix is generated whereby each row number matches with the index of the word in the corpus. Raw tweets contain text instances which cannot handle by machine learning procedure. Therefore, we run data pre-processing and tokenization process to make it executable for clustering and classification computation.

3.4. Traditional analysis

In the traditional process, we have been manipulated by various data pre-processing, tokenization and implemented different baseline classifiers into twitter datasets. Therefore, various well known classifiers were applied in the primary datasets using 10 fold cross validation and compared the results with TClustVID. However, both traditional and TClustVID have been used the same baseline classifier which indicates at Section 3.6.

3.5. TClustVID: Clustered based classification and topics modeling approach

In the beginning, different preprocessing and tokenization process has been implemented into COVID-19 twitter datasets and split them into several groups applying k-means method. Clustering is an unsupervised technique to partition a set of the dataset into subsets/clusters. This procedure is helpful to improve the performance of machine learning methods by creating clusters. There are existing various algorithms such as k-means, k-medoids, fuzzy C-means, hierarchical, and density based clustering [37,38]. K-medoids is not the best choice for analyzing sparse data like tweets. Then, fuzzy C-means is useful to the sheer volumes of tweets and contains low scalability where human annotation really expensive. The performance of hierarchical clustering is slower than the k-means method. Density based

Table 1
Number of cleaned tweets COVID-19 after data preprocessing.

Primary dataset	# tweets (N = 19,797,541)	Denoted	# tweets (N = 19,712,979)
Before preprocessing		After preprocessing	
corona_tweets_1M.db	1,578,957	Dataset-1	1,569,619
corona_tweets_1M_2	1,889,781	Dataset-2	1,880,297
corona_tweets_1M	1,903,768	Dataset-3	1,894,526
corona_tweets_2L	2,80,304	Dataset-4	2,76,566
corona_tweets_2M.db	2,322,153	Dataset-5	2,312,104
corona_tweets_2M_2	2,268,634	Dataset-6	2,257,529
corona_tweets_2M_3	2,081,576	Dataset-7	2,072,575
corona_tweets_3M	7,472,368	Dataset-8	3,724,882
		Dataset-9	3,724,881

Algorithm 1 TClustVID: Clustered Based Proposed Classification and Topics modeling Approach

Input: Set of twitter dataset D_s, set of classifier C, the number of dataset s, the number of tokens *tokens*, set of cluster $Clust_s$, derived cluster $Clust_{jm}$, set of evaluation metrics P_{jm},the number of topics T_N
Output: Find out the significant topics to COVID-19.

Begin
Cleaning Dataset $D_s[review]$ by removing tags, punctuation, characters and multiple spaces
$K \leftarrow 5$, $T_N \leftarrow 20$
for each Dataset $D_i \in D_s$ **do**
  $D_i[tokens] \leftarrow Tokenize(D_i[reviews])$
  $Clust_i[tokens] \leftarrow kmeans(D_i[tokens], K)$
  Replace $Clust_{ji}[tokens]$ with $D_i[reviews]$
  $Clust_i[tokens] \leftarrow Tokenize(Clust_i[tokens])$
end for
$m \leftarrow 0$
while $m! = s$ **do**
  for each Classifier $C_i \in C$ **do**
    for each $Cluster\_num_j \in K$ **do**
      $P_{jm} \leftarrow Classification_{CV}(Clust_{jm})$
    end for
  end for
  $P_{jm}^{max} \leftarrow maximum(P_{jm})$
  Compare P_{jm}^{max} with traditional classification
  Find out $Clust_{jm}^{max}$ by considering P_{jm}^{max}
  Divide $Clust_{jm}^{max}$ into $Clust_{pos}^{max}$, $Clust_{neu}^{max}$ and $Clust_{neg}^{max}$
  $Topic_{pos} \leftarrow LDA(Clust_{pos}^{max}, T_N)$
  $Topic_{neu} \leftarrow LDA(Clust_{neu}^{max}, T_N)$
  $Topic_{neg} \leftarrow LDA(Clust_{neg}^{max}, T_N)$
  Interpret $Topic_{pos}$, $Topic_{neu}$ and $Topic_{neg}$
  Calculate top frequent topics from $Topic_{pos}$, $Topic_{neu}$ and $Topic_{neg}$
  $m \leftarrow m + 1$
end while

techniques are highly efficient for clustering unstructured data and less prone to outliers and noise. In this work, we processed a large amount of tweet data where K-means defines the mean point within the cluster by optimizing the Euclidean distance between each instance in less time [38,39]. The default values of k = 5 are mainly used in this work. Each cluster has been contained positive, negative and neutral tweets where generated tokens were replaced by primary tweets and re-tokenized in each cluster. Baseline classifiers have then been used to investigate the performance of individual clusters using 10 fold cross-validation. Different evaluation metrics such as accuracy, the area under the curve (AUC), f-measure, g-mean, sensitivity and specificity were used to assess these results. The detailed working steps of TClustVID is represented briefly in the Algorithm summary 1.

Compared the classification results of traditional approach and TClustVID, the best performing clusters can be used to extract more frequent topics. These clusters are divided into positive, neutral and negative sentiments for further analysis. Therefore, LDA has been used to explore significant positive, neutral and negative topics from the high performing nine clusters. 20 topics were extracted from each cluster. We represented individual topics in a word cloud where each contains different words/tokens. According to the weights of tokens, this cloud represents different word. However, LDA cannot interpret these topics so we manually analyzed the words/tokens of each topics to define them.

3.6. Baseline classification

In previous studies, the various classifiers such as decision tree (DT), Gradient Boosting (GB), K-Nearest Neighbor (KNN), Logistic Regression (LR), Multi-Layer Perceptron (MLP), Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM) and XGBoost (XGB) have been commonly used to investigate different twitter datasets for sentiment analysis. These classifiers were used in similar kinds of tasks such as C5.0 (DT), KNN, SVM, LR and ZeroR [40], personality prediction using KNN, NB, SVM, and XGB [41,42], spam detection using RF, NB, SMO and Ibk (KNN equivalent) [43], sentiment analysis using NB, SVM, and MLP of top colleges [44], prediction of alternation price fluctuation using GB [45]. Following this literature, we selected them to investigate COVID-19 twitter dataset and explore the best clusters.

3.7. Evaluation metrics

A confusion matrix is needed to estimate the performance of the classifier that indicates the number of correct and incorrect predictions by considering known true values. Based on positive and negative classes, this shows True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values for the data fitting.

- **Accuracy:** represents the efficiency of the algorithm in terms of predicting true values.

$$\text{Accuracy} = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (1)$$

- **AUC:** is used to explore machine learning models considering the TP and TN rates represent how well positive classes are isolated from negative classes.

$$AUC = \frac{\text{TPrate} + \text{TNrate}}{2} \quad (2)$$

- **F-measure:** represents the harmonic mean of precision and recall.

$$F-\text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{(\text{precision} + \text{recall})} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (3)$$

Fig. 5. Positive topics of Cluster-3.

- **Geometric-mean (G-mean):** specifies the root of the class-specific sensitivity product and makes a trade-off between the expansion of accuracy on each class and balancing accuracy.

$$\text{GMean} = \sqrt{(\text{TPrate} \times \text{TNrate})} \tag{4}$$

- **Sensitivity:** The portion of appropriately detected actual positives is indicated as sensitivity.

$$\text{Sensitivity} = \frac{TP}{(TP + FN)} \tag{5}$$

- **Specificity:** The portion of correctly identified actual negatives is denoted as specificity.

$$\text{Specificity} = \frac{TN}{(FP + TN)} \tag{6}$$

4. Experimental result

4.1. Sentiment analysis through classification approach

In this study, our proposed TClustVID has detected positive, negative, and neutral tweets more accurately using a clustering based classification and explored more significant thematic topics. However, primary datasets were cleaned using different data preprocessing procedures and Word-to-index dictionaries were then created using GloVe embedding tokenization. Several classification algorithms such as DT, GB, KNN, LR, MLP, NB, RF, SVM and XGB were analyzed sentiments of the COVID-19 datasets using the sci-kit-learn machine learning python library [46,47]. The results of individual classifiers for nine COVID-19 twitter datasets are represented at Table 2.

In traditional analysis, a number of classifiers such as LSTM, DT, RF, GB, KNN, MLP, NB, RF, SVM and XGB have been implemented. Therefore, LSTM gave the highest accuracy, f-measure and sensitivity and DT provided maximum AUC and specificity for dataset-1. Also, this classifier outperformed other classifiers in all evaluation metrics for dataset-2, 5 and 8, respectively. In addition, LSTM provided the highest accuracy, f-measure and sensitivity as well as RF provided the best AUC and specificity for dataset-3 and 4, individually. However, DT generated the maximum accuracy and sensitivity while RF gave the highest AUC, f-measure and specificity for dataset-6. Again, RF outperformed other classifiers in all metrics for dataset-7. LSTM showed the highest AUC, f-measure and sensitivity and RF provided the best accuracy and specificity for dataset-9. In contrast, individual classifiers were employed into different datasets using TClustVID where their results have been improved over the traditional analysis. However, the same classification methods that have been used in a general way were employed into Twitter datasets using TClustVID.

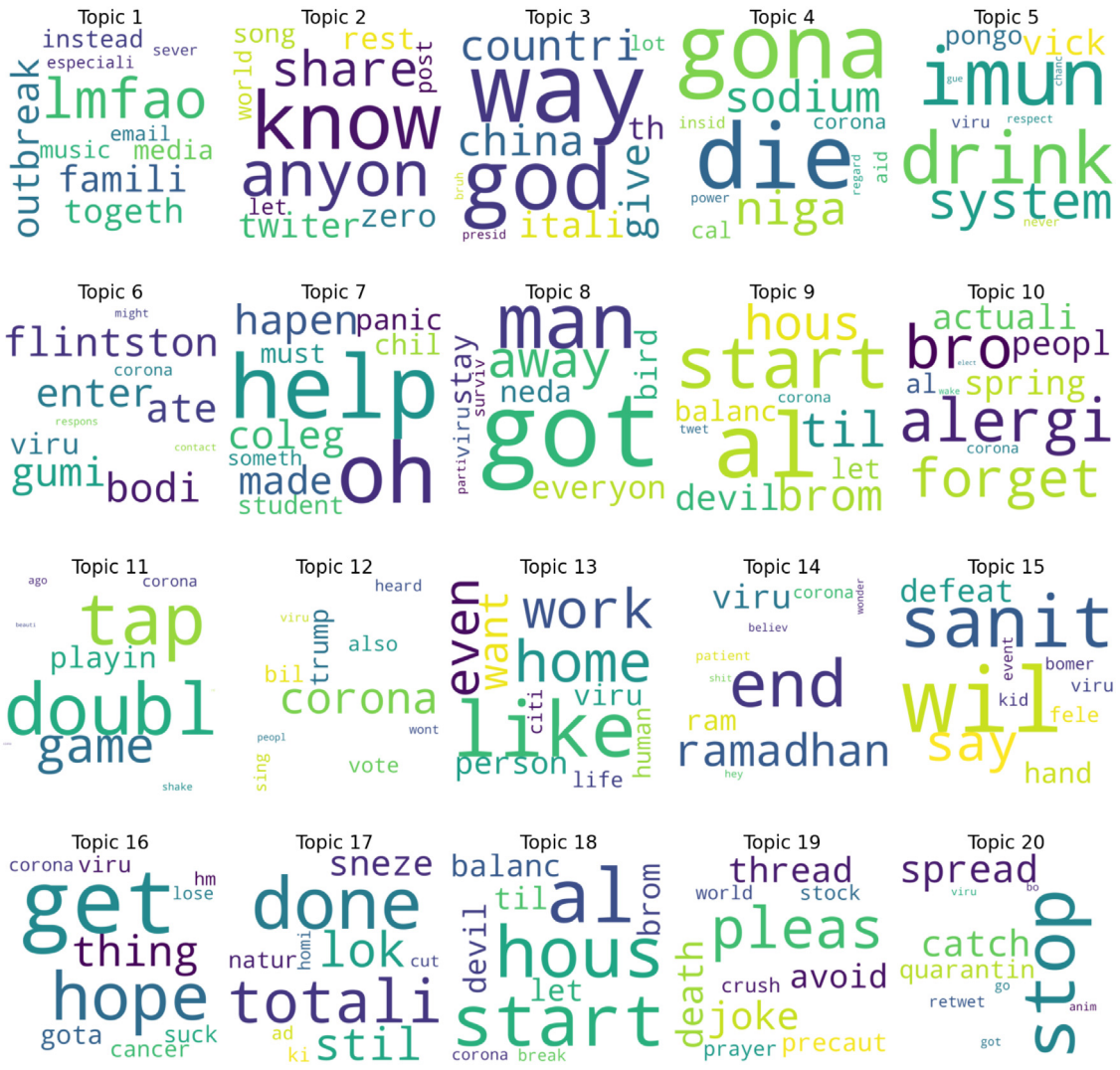

Fig. 6. Neutral topics of Cluster-3.

Several clusters have been produced that were used to generate classification results where TClustVID has been identified those clusters whose were given the best classification results among them. In this case, LSTM outperforms other classifiers with all evaluation metrics for all datasets.

In Fig. 2(a), the average outcomes of different classifiers such as LSTM, DT, KNN, MLP, XGB, GB, SVM and LR are represented using a traditional approach. Similarly, TClustVID manipulated average results of the same classifiers used by TClustVID and compared its findings with traditional procedure (see Fig. 2(b)). In this case, LSTM provided the highest average accuracy, AUC, f-measure, sensitivity and specificity for both traditional way and TClustVID. In addition, TClustVID showed better results compared to more traditional approaches (see Fig. 2).

However, we measured shapley additive explanations (SHAP) values of various tokens to determine positive, neutral and negative sentiments more effectively. SHAP is a game theoretic technique to interpret the findings of any machine learning model. Therefore, the result of TClustVID for LSTM has been evaluated in each cluster and explored which tokens are responsible to classify positive, neutral and negative sentiments. Fig. 3 shows the probability of SHAP values for different tokens in different nine clusters.

Along with observing the performance of various classifiers, we noted that TClustVID shows better performance than traditional analysis. Hence, a topic modeling approach is used to produce high performing clusters for the extraction of significant topics in the next section (see Fig. 4).

4.2. Topic modeling approach

4.2.1. Extraction of clusters using TClustVID

A comprehensive analysis of different classifiers in traditional and TClustVID analyses indicated that TClustVID is the best model to identify significant groups of tweets from large COVID-19 Twitter datasets. The data obtained from the identification of groups/clusters were significant because they showed the highest classification accuracy were achieved compared to traditional analysis in primary data. In the TClustVID analysis, we generated significant clusters from each of these twitter datasets (for positive neutral, and negative categories) that showed greatly improved results for the different classifiers. These clusters have been denoted as Cluster-1, Cluster-2, Cluster-3, Cluster-4, Cluster-5, Cluster-6, Cluster-7, Cluster-8, and Cluster-9, respectively.

4.2.2. Topics exploration using LDA

A number of topics were then extracted from these clusters where within nine clusters seven of them produced positive,

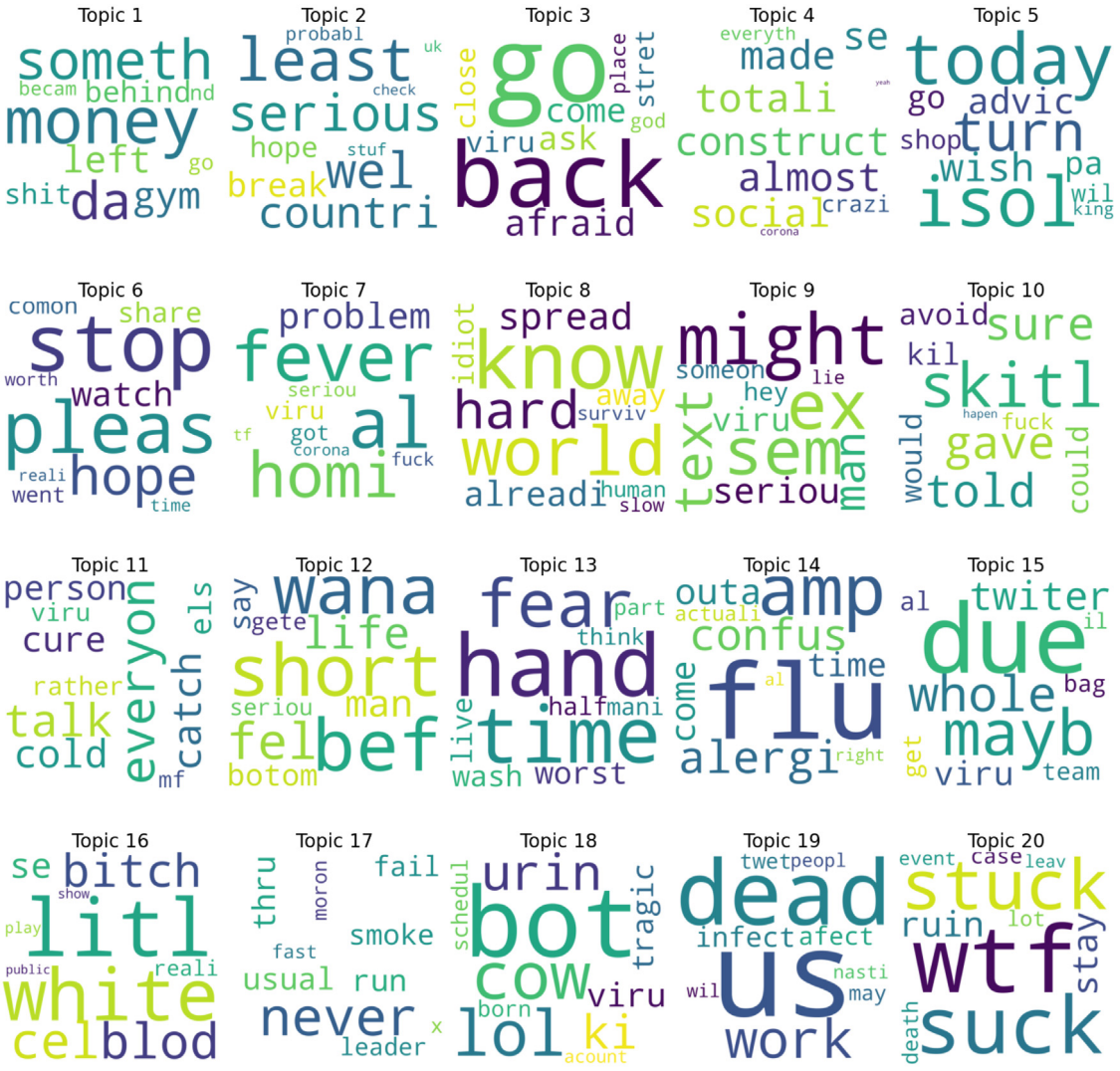

Fig. 7. Negative topics of Cluster-3.

neutral and negative topics and two of them extracted positive and neutral topics using LDA. Each topic contains 10 tokens along with related weights and they can be used to prioritize each token. 20 topics were identified from each of the categories (positive, neutral and negative) in these clusters. Therefore, all topics of individual clusters are represented as a word cloud in the supplementary section. In this paper, extracted positive, neutral and negative topics of cluster-3 are visualized with word cloud in Figs. 5–7 individually.

4.2.3. Qualitative analysis

As LDA cannot interpret the meaning of topics, we defined their themes by determining the meaning and weight values in different groups manually. The themes of positive, neutral and negative topics are indicated in Tables 3–5 respectively. These tasks are not simple because many pre-processed words do not have any semantic meaning. However, it can be hard to understand the association between the different words/tokens in these topics and these interpretations may slightly differ from that used in other types of reviewing.

In the different categories of tweets, we manipulated the frequency of different topics that appears several times. Positive, neutral and negative topics have been identified what activities are generated in the context. To understand individual topics into different themes, we considered the best themes which are appeared more than 1 times (see Fig. 8). The examples of positive topics of cluster-3 are shown as the word cloud in Fig. 5. In addition, The themes of positive topics within different clusters are shown in Table 3 and the top frequent positive themes are shown in Fig. 8(a). For the positive cases, awareness and situation are the most frequent themes that appear many times in different clusters. Both of these appear 17 times in different significant clusters. Awareness has specified those actions whose are taken by individuals and situation symbolizes the general situation of particular places/incidents where pandemic news indicates a generic situation relating to COVID-19. Wishes appear 8 and new appears 7 times in this study. Furthermore, caring, coronavirus, right and treatment are gathered 5 times, and message, and social distance are found 4 times this effort. Subsequently, cases, prevention, testing and tourism are obtained 3 times in the COVID-19 situation. In addition, other precaution related themes such as affect, annoying, blaming, closing, crisis, effect, facts, financial help, help, infectious, lockdown, medicine, need, panic, quarantine, risk and scaring are represented their frequency 2 times in different clusters. They are appeared regularly and specifies how we can improve this condition. However, some of negative themes, for instance blaming, crisis, infectious, panic, risk appeared in positive cases but their frequencies are not greater. More upcoming positive issues are also addressed in this

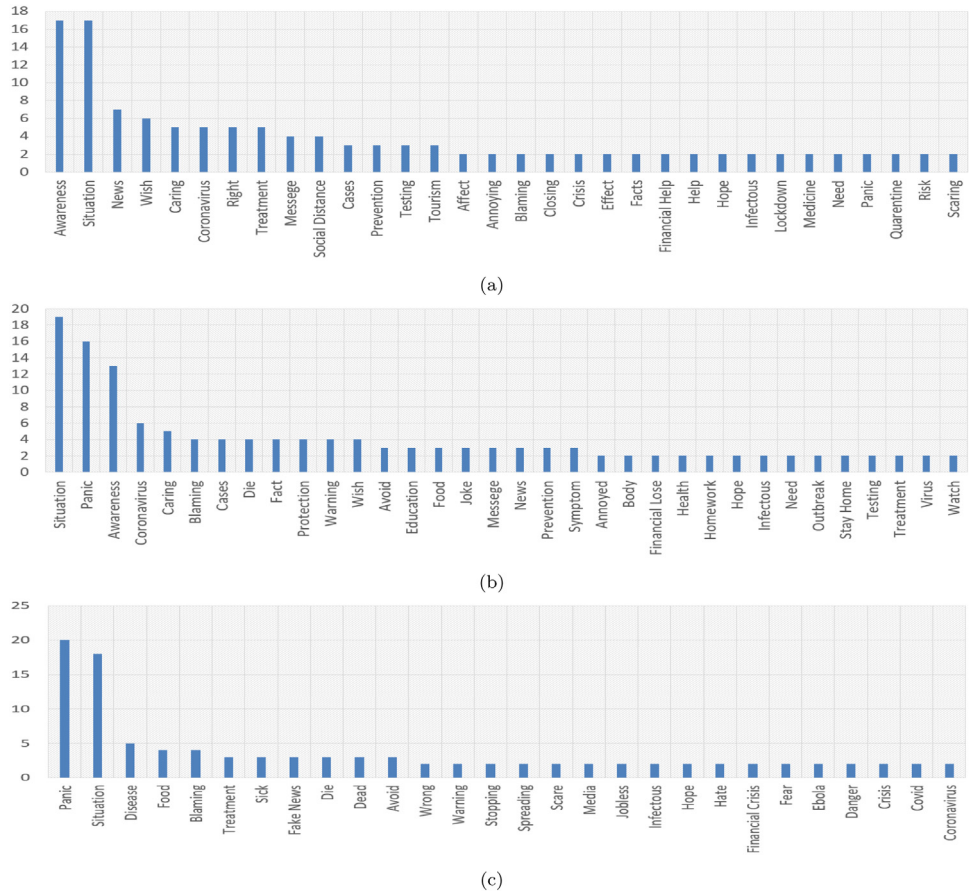

Fig. 8. Top frequency of (a) Positive (b) Neutral (c) Negative COVID-19 associated topics.

analysis included financial help, help, lockdown, quarantine and medicine.

In the neutral category, there are appeared the mixture of positive and negative topics which indicates the most frequent topics in recent timeframes. For example, we have represented an example of neutral topics as a world cloud in Fig. 6. Besides this, neutral themes of different clusters are provided in Table 4 and top frequent themes are shown in Fig. 8(b). Therefore, situation, panic and awareness are found 19, 16 and 13 times in the following list of twitter topics. Panic is a related theme to explain epidemic conditions and news. In addition, wish and coronavirus appear 6 times as well as caring which appears 5 times at negative tweets. Consequently, blaming, cases, die, warning and protection appear 4 times while education, food, joke, message, news, prevention, and symptom appear 3 times in this condition. The rest of the themes perform 2 times to represent neutral topics. The issues such as those related to before and after the COVID-19 pandemic like Financial, lose, crisis, food, education also arose in this analysis.

The negative topics using the word cloud are represented in Fig. 7. Thus, the themes of negative topics have been provided in Table 5 and topmost frequent themes are shown in Fig. 8(c). In this category, panic and situation appear most of the times than other topics. Both of them appear 20 and 18 times respectively. Dead and disease appear 6 and 5 times enabling estimation of its influence. Thus, food and blaming occur 4 times and treatment, sick, fake news and avoid represent 3 times to represent significant topics. Some cases like food and treatment indicate the level of crisis perceived. The rest of the themes are provided with a frequency of 2 in this work. Therefore, they are shown in the top list of feelings or perceptions relating to COVID-19 that are negative.

5. Discussion

5.1. Comparison of TClustVID with recent published work

Proposed TClustVID is overcame many of the pitfalls that are evident in many recent work. In current work, we present a well-organized machine learning model that has been employed into common COVID-19 oriented tweets where different regions are not specified like previous studies [18,28,48,49]. Both sentiment analysis and topics modeling were used to explore COVID-19 related themes than many works [18,20,22,24,27,48–51]. However, many machine learning classifiers have been implemented in which we compared our proposed model with more traditional analyses to evaluate performance. However, most previous studies [18,28,31] used only a small number of classifiers to verify their tasks. Our work was also able to extract reliable themes of positive, negative and neutral topic to explore clusters and realize the condition of COVID-19.

5.2. Implications

Twitter refer to a reasonable and proficient platform to validate the efficiency of public health communication. Real-time epidemiological data are required to properly and comprehensively characterize user discussion, self-reporting capabilities and rapid evaluation of pandemic situation. In this study, we developed a machine learning based framework named TClustVID and investigated various types of public tweets related to COVID-19, identifying related sentiments and extracted associated topics from a number of localities. This efficiently provides significant

Table 2
The results of sentiment classification for individual datasets.

Dataset	Classifier	Accuracy	AUC	F-Measure	Sensitivity	Specificity	Dataset	Accuracy	AUC	F-Measure	Sensitivity	Specificity
		Traditional Analysis						TClustVID				
Dataset-01	LSTM	**0.927**	0.897	**0.926**	**0.927**	0.868	Cluster-01	**0.983**	**0.979**	**0.983**	**0.983**	**0.974**
	DT	0.915	**0.901**	0.916	0.915	**0.886**		0.952	0.945	0.952	0.952	0.938
	GB	0.788	0.653	0.746	0.788	0.518		0.816	0.713	0.790	0.816	0.610
	KNN	0.910	0.880	0.909	0.910	0.850		0.946	0.930	0.945	0.946	0.914
	LR	0.695	0.502	0.576	0.695	0.308		0.679	0.500	0.552	0.679	0.322
	MLP	0.840	0.766	0.830	0.840	0.692		0.901	0.869	0.899	0.901	0.836
	NB	0.654	0.503	0.597	0.654	0.352		0.644	0.502	0.577	0.644	0.359
	RF	0.924	0.898	0.923	0.924	0.873		0.957	0.944	0.957	0.957	0.932
	SVM	0.757	0.600	0.694	0.757	0.442		0.803	0.693	0.772	0.803	0.583
	XGB	0.787	0.657	0.750	0.787	0.527		0.854	0.774	0.840	0.854	0.695
Dataset-02	LSTM	**0.968**	**0.949**	**0.968**	**0.968**	**0.929**	Cluster-02	**0.988**	**0.977**	**0.988**	**0.988**	**0.967**
	DT	0.931	0.899	0.931	0.931	0.866		0.964	0.943	0.964	0.964	0.921
	GB	0.816	0.567	0.755	0.816	0.318		0.856	0.626	0.819	0.856	0.396
	KNN	0.924	0.865	0.922	0.924	0.806		0.958	0.918	0.957	0.958	0.878
	LR	0.787	0.501	0.695	0.787	0.216		0.807	0.505	0.727	0.807	0.203
	MLP	0.867	0.730	0.854	0.867	0.592		0.925	0.841	0.921	0.925	0.756
	NB	0.213	0.500	0.076	0.213	0.787		0.192	0.500	0.063	0.192	0.808
	RF	0.937	0.888	0.936	0.937	0.838		0.968	0.936	0.967	0.968	0.905
	SVM	0.787	0.501	0.695	0.787	0.215		0.806	0.502	0.724	0.806	0.197
	XGB	0.820	0.578	0.764	0.820	0.336		0.875	0.694	0.855	0.875	0.513
Dataset-03	LSTM	**0.915**	0.922	**0.915**	**0.915**	0.929	Cluster-03	**0.985**	**0.987**	**0.985**	**0.985**	**0.988**
	DT	0.911	0.930	0.911	0.911	0.950		0.960	0.967	0.960	0.960	0.973
	GB	0.699	0.717	0.674	0.699	0.735		0.846	0.838	0.836	0.846	0.830
	KNN	0.893	0.918	0.893	0.893	0.942		0.950	0.958	0.950	0.950	0.965
	LR	0.514	0.548	0.426	0.514	0.582		0.668	0.651	0.628	0.668	0.635
	MLP	0.793	0.827	0.788	0.793	0.860		0.909	0.914	0.908	0.909	0.918
	NB	0.485	0.551	0.441	0.485	0.616		0.212	0.504	0.178	0.212	0.796
	RF	0.911	**0.933**	0.911	0.911	**0.955**		0.959	0.966	0.959	0.959	0.973
	SVM	0.344	0.520	0.308	0.344	0.696		0.463	0.617	0.482	0.463	0.770
	XGB	0.722	0.766	0.710	0.722	0.809		0.847	0.846	0.838	0.847	0.845
Dataset-04	LSTM	**0.904**	0.915	**0.903**	**0.904**	0.926	Cluster-04	**0.957**	**0.957**	**0.956**	**0.957**	**0.956**
	DT	0.892	0.915	0.892	0.892	0.937		0.943	0.949	0.943	0.943	0.956
	GB	0.621	0.614	0.553	0.621	0.607		0.818	0.788	0.806	0.818	0.758
	KNN	0.873	0.901	0.873	0.873	0.929		0.930	0.939	0.930	0.930	0.948
	LR	0.547	0.556	0.457	0.547	0.565		0.747	0.722	0.728	0.747	0.697
	MLP	0.765	0.797	0.758	0.765	0.829		0.882	0.877	0.878	0.882	0.872
	NB	0.533	0.536	0.422	0.533	0.539		0.274	0.506	0.139	0.274	0.737
	RF	0.892	**0.918**	0.892	0.892	**0.943**		0.943	0.950	0.942	0.943	0.958
	SVM	0.397	0.519	0.398	0.397	0.641		0.326	0.523	0.340	0.326	0.720
	XGB	0.683	0.691	0.648	0.683	0.699		0.825	0.809	0.817	0.825	0.792
Dataset-05	LSTM	**0.904**	**0.927**	**0.903**	**0.904**	**0.951**	Cluster-05	**0.968**	**0.975**	**0.968**	**0.968**	**0.983**
	DT	0.866	0.899	0.866	0.866	0.932		0.902	0.925	0.902	0.902	0.949
	GB	0.534	0.625	0.494	0.534	0.715		0.624	0.684	0.587	0.624	0.744
	KNN	0.841	0.880	0.841	0.841	0.920		0.878	0.907	0.878	0.878	0.937
	LR	0.431	0.552	0.367	0.431	0.673		0.454	0.557	0.386	0.454	0.659
	MLP	0.624	0.712	0.622	0.624	0.801		0.749	0.800	0.744	0.749	0.851
	NB	0.419	0.529	0.305	0.419	0.639		0.429	0.524	0.344	0.429	0.619
	RF	0.865	0.900	0.865	0.865	0.934		0.900	0.924	0.900	0.900	0.949
	SVM	0.338	0.525	0.258	0.338	0.711		0.424	0.537	0.362	0.424	0.650
	XGB	0.548	0.647	0.532	0.548	0.745		0.645	0.717	0.639	0.645	0.789
Dataset-06	LSTM	0.876	0.908	0.877	0.876	0.941	Cluster-06	**0.977**	**0.982**	**0.977**	**0.977**	**0.987**
	DT	**0.879**	0.909	0.879	**0.879**	0.938		0.932	0.948	0.932	0.932	0.963
	GB	0.602	0.659	0.562	0.602	0.715		0.763	0.785	0.748	0.763	0.807
	KNN	0.858	0.893	0.859	0.858	0.929		0.917	0.936	0.917	0.917	0.955
	LR	0.474	0.561	0.400	0.474	0.648		0.526	0.581	0.465	0.526	0.636
	MLP	0.714	0.778	0.712	0.714	0.842		0.846	0.874	0.845	0.846	0.902
	NB	0.450	0.522	0.315	0.450	0.594		0.475	0.515	0.328	0.475	0.554
	RF	0.879	**0.910**	**0.879**	0.879	**0.942**		0.931	0.948	0.931	0.931	0.964
	SVM	0.418	0.530	0.341	0.418	0.643		0.536	0.568	0.433	0.536	0.600
	XGB	0.642	0.719	0.637	0.642	0.796		0.774	0.813	0.772	0.774	0.851

(*continued on next page*)

insights on how people interpret mixed around COVID-19 messages. There are numerous theoretical and practical implications about this model which is described as follows.

5.2.1. Theoretical implications of the study

- This proposed method has extracted positive, negative and neutral topics to scrutinize its contents and extract significant values to give various information about related issues.

Table 2 (*continued*).

Dataset	Classifier	Accuracy	AUC	F-Measure	Sensitivity	Specificity	Dataset	Accuracy	AUC	F-Measure	Sensitivity	Specificity
		Traditional Analysis						TClustVID				
Dataset-07	LSTM	0.903	0.919	0.903	0.903	0.936	Cluster-07	**0.983**	**0.986**	**0.983**	**0.983**	**0.990**
	DT	0.908	0.929	0.908	0.908	0.951		0.955	0.965	0.955	0.955	0.975
	GB	0.664	0.718	0.656	0.664	0.773		0.810	0.830	0.806	0.810	0.850
	KNN	0.889	0.915	0.889	0.889	0.942		0.941	0.954	0.941	0.941	0.967
	LR	0.451	0.538	0.380	0.451	0.624		0.548	0.598	0.501	0.548	0.647
	MLP	0.768	0.813	0.764	0.768	0.859		0.885	0.905	0.885	0.885	0.925
	NB	0.219	0.501	0.083	0.219	0.783		0.220	0.503	0.094	0.220	0.787
	RF	**0.909**	**0.931**	**0.909**	**0.909**	**0.954**		0.954	0.964	0.954	0.954	0.975
	SVM	0.353	0.517	0.353	0.353	0.681		0.299	0.539	0.251	0.299	0.780
	XGB	0.635	0.705	0.632	0.635	0.774		0.815	0.843	0.814	0.815	0.871
Dataset-08	LSTM	**0.908**	**0.921**	**0.907**	**0.908**	**0.935**	Cluster-08	**0.976**	**0.981**	**0.976**	**0.976**	**0.985**
	DT	0.870	0.901	0.870	0.870	0.931		0.910	0.929	0.910	0.910	0.948
	GB	0.600	0.654	0.557	0.600	0.709		0.687	0.698	0.655	0.687	0.708
	KNN	0.847	0.884	0.847	0.847	0.921		0.853	0.884	0.853	0.853	0.915
	LR	0.501	0.582	0.440	0.501	0.663		0.516	0.547	0.428	0.516	0.578
	MLP	0.650	0.722	0.635	0.650	0.794		0.795	0.825	0.790	0.795	0.856
	NB	0.460	0.529	0.332	0.460	0.599		0.489	0.536	0.379	0.489	0.582
	RF	0.870	0.903	0.870	0.870	0.936		0.909	0.930	0.909	0.909	0.951
	SVM	0.440	0.513	0.326	0.440	0.585		0.409	0.505	0.337	0.409	0.601
	XGB	0.597	0.678	0.577	0.597	0.759		0.678	0.724	0.669	0.678	0.770
Dataset-09	LSTM	0.897	**0.912**	**0.896**	**0.897**	0.928	Cluster-09	**0.976**	**0.981**	**0.976**	**0.976**	**0.986**
	DT	0.870	0.900	0.870	0.870	0.931		0.911	0.930	0.911	0.911	0.949
	GB	0.600	0.654	0.557	0.600	0.709		0.686	0.698	0.651	0.686	0.711
	KNN	0.847	0.884	0.847	0.847	0.921		0.856	0.886	0.856	0.856	0.917
	LR	0.498	0.579	0.437	0.498	0.660		0.508	0.541	0.420	0.508	0.574
	MLP	0.650	0.715	0.633	0.650	0.780		0.802	0.830	0.797	0.802	0.859
	NB	0.221	0.500	0.083	0.221	0.780		0.250	0.507	0.191	0.250	0.764
	RF	**0.869**	0.902	0.870	0.869	**0.936**		0.910	0.931	0.910	0.910	0.952
	SVM	0.345	0.508	0.300	0.345	0.671		0.270	0.515	0.243	0.270	0.759
	XGB	0.599	0.680	0.579	0.599	0.760		0.676	0.726	0.668	0.676	0.776

- TClustVID has been designed to focus on particular types of analyses such as psychological and emotional analysis. However, it can easily be generalized and adapted to analyze any specific topics of interest.
- This study is very useful to verify these kinds of analysis in various perspective. However, demographic analysis, comparison and discussion can give a concrete idea about various source.
- Theoretical understanding gained from this work can be used for addressing similar types of problems but also doing so at a lower cost.
- From the limitations and suggestions, researchers can take numerous new challenges in future work.

5.2.2. Practical implications

- Users of this model can isolate individuals one from another by giving relaxation and support via social media. It safeguards people interest and needs in the society.
- This analytical approach can be used for comprehensive contact tracing, unidentified hot spots of COVID-19 infection and increase the accuracy, predictability to find out COVID-19 cases.
- This model can be employed to explore how to improve public health campaigns on the leading topics featuring in twitter conversations to give timely responses and improve initiatives taken by agencies.
- This work has mainly focused on a number of particular common concerns relating to working conditions. Many tweets have been posted about working from home during this outbreak.
- It can be explored an opportunity to follow patterns of vaccine acceptance and failure or criticism against it. Also, it allows assessment of real-time trends for COVID-19 treatment, medical equipment, diagnosis, cross correlating its information with medical information and other factors.
- A new surveillance system can be built to examine web-based contents using this model for better understanding of public emotions and concerns.
- This works can be generalized to analyze other social media data such as Instagram, Facebook and YouTube
- The scientific community can also be studied to determine for their the similarity and dissimilarity from public comments using this model.
- Our work has generated useful data for agencies, local leaders, health providers and municipalities. This can enable governments to coordinate the flow of information and combat misinformation about the pandemic.

5.3. Limitations of the sudy

Twitter gives the community interaction and its user profiles represent a relatively small demographic data for further analysis. We only gathered tweets using a few numbers of keywords from one social media platform. This study has only investigated English language tweets. In addition, machine and deep learning methods have not been implemented into a large amount of COVID-19 oriented tweets. Again, the interpretation of topics is a challenging task, hence some manual interpretation of topics may misinterpret in the topics modeling.

5.4. Challenges and future suggestions

A number of challenges can be considered for investigating COVID-19 tweets for sentiment analysis and topics modeling. In different social media such as Twitter, many cases of showing irrelevant, fake, misinformed and insufficient data has been found.

Table 3
Positive themes of all significant clusters.

	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5
Theme-1	Culture	Prevention	Kids	Wish	Sunny
Theme-2	Nationality	Situation	Wish	News	Watch
Theme-3	Prevention	Situation	Testing	Situation	Affect
Theme-4	Caring	Homework	Treatment	Help	Situation
Theme-5	Blaming	News	Testing	Help	Treatment
Theme-6	Believe	News	Caring	Facts	Awareness
Theme-7	Die	News	Feeling	Control	Medicine
Theme-8	Caring	Wish	Situation	Infectious	Treatment
Theme-9	Discrimination	Awareness	Scaring	Right	Medicine
Theme-10	Situation	Financial state	Buying	Awareness	Awareness
Theme-11	Crisis	News	Fun	Wish	Prevention
Theme-12	Financial Help	Avoidness	Right	News	Situation
Theme-13	Condition	Crisis	Panic	Situation	Awareness
Theme-14	Wish	Food	Protection	Distance & Treatment	Treatment
Theme-15	Lockdown	Blaming	Health	Annoying	Awareness
Theme-16	Closing	Situation	Awareness	Situation	Humor
Theme-17	Closing	Lockdown	Panic	Job	Situation
Theme-18	Awareness	Awareness	Effect	Stay Safe	Risk
Theme-19	Financial help	Annoying	Micro-Organism	Awareness	Situation
Theme-20	Caring	Awareness	News	Wish	Risk
	Cluster-6	Cluster-7	Cluster-8	Cluster-9	
Theme-1	Right	Testing & Treatment	Survive	Shut	
Theme-2	Need	Interest	Flu	Honest	
Theme-3	Covid	Need	Move	Media	
Theme-4	Social media	Social distance	Overreact	Right	
Theme-5	Awareness	Social distance	Situation	Testing	
Theme-6	Flight	Epidemic	Rumor	Caring	
Theme-7	Messege	Social distance	Fight & Caring	Isolation	
Theme-8	Right	Symptoms	Cases	Survive	
Theme-9	Treatment	Effect	Disease	Home	
Theme-10	Wish	Confirmed	Cases	Wish	
Theme-11	Situation	Coronavirus	Awareness	Worried	
Theme-12	Warning	Message	Infectious	Situation	
Theme-13	Testing & Treatment	Coronavirus	Social guys	Quarantine	
Theme-14	Cases	Social distance	Situation	Love	
Theme-15	Message	Tourism	Quarantine	Scaring	
Theme-16	Message	Tourism	Awareness	Do not Move	
Theme-17	Situation	Coronavirus	Facts	Affect	
Theme-18	Tourism	Outbreak	Schools	Wind	
Theme-19	Coronavirus	Coronavirus	Crisis & Prevention	Awareness	
Theme-20	Awareness	Awareness	Financial enrichment	Fuck	

In addition, these tweets needed to be collected from different domains of the social media. It is difficult for researchers to work with this dataset as processing such dataset requires a high degree of technical skill. It is often hard to define which keywords are appropriate to gather COVID-19 related tweets and identify desired data. Moreover, decision makers face troubles to identify people's sentiments on a subject or to characterize their beliefs. Also, there remain a lack of scientific studies, to gather knowledge for designing a new model.

In these difficult circumstances, we will need to face these challenges. Along with Twitter, the records of other social media (such as Facebook, YouTube, Instagram and Reddit) need to be investigated to explore knowledge about COVID-19 pandemic from users. To overcome the general lack of published literature on the subject, most relevant previous works about pandemic situation can be useful for getting solutions from them. However, COVID-19 related hashtags and keywords need to be explored using to recently developed academic literature and sentiment and opinion mining tasks. New developments such as TClustVID can also be used with modifications to analyze more similar but heterogeneous records of various sources.

6. Conclusion

In this work, we have proposed a clustered based machine learning model named TClustVID that has given the best performance outcomes in sentiment analysis and topics modeling by analyzing COVID-19 twitter datasets compared to other methods. TClustVID first extracted various clusters from individual datasets using k-means algorithm [38], then the proposed model was used to separate different classifiers into clusters and one of them represents the highest classification accuracy in each dataset. We subsequently compared the topmost clustering result of each dataset with traditional analysis with TClustVID showing the maximum outcomes for each case. Furthermore, the best clusters identified provided more significant topics in each dataset and represents public opinions on Twitter. It also explored more significant information that can be abstracted from very large numbers of tweets by extracting commonly occurring topics and interpreting their themes. This model is helpful to identify important themes about the situation at the time the tweets were sent, and can enable designing better strategies to counter the pandemic that take human responses and behavior into account. This knowledge was extracted from positive, neutral and negative tweets and identified high frequency information features transmitted and commented as the response to the epidemic condition.

As noted in the Study Limitations (Section 5.3) and future guidelines of this work (Section 5.4), more COVID-19 oriented social media data from different sources can in future be collected and investigated using TClustVID (and improved versions of TClustVID) and other techniques currently being used, which will enable efficient extraction and analysis of significant information about COVID-19 and other health emergencies.

Table 4
Neutral themes of all significant clusters.

	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5
Theme-1	Financial lose	Warning	Outbreak	Situation	Awareness
Theme-2	Fact	Food	Sharing	Panic	Infectious
Theme-3	Warning	Situation	Wish	Situation	Situation
Theme-4	Estimate	Situation	Gonna	Entertainment	Need
Theme-5	Blaming	Testing	Caring	Protection	Wish
Theme-6	Pleased	Rumor	Caring	Dead	Food
Theme-7	Financial lose	Warning	Panic	Health	Break
Theme-8	Pandemic warning	Visiting	Survive	Stay Home	Treatment
Theme-9	Awareness	Joke	Awareness	Avoid	Want
Theme-10	Disease	Panic	Treatment	Fact	Prevention
Theme-11	Warning	Situation	Playing game	Awareness	Awareness
Theme-12	Caring	Panic	Coronavirus	Protection	Panic
Theme-13	Panic	Closing	Homework	Awareness	Situation
Theme-14	Panic	Panic	Ramadhan news	Situation	Awareness
Theme-15	Awareness	Panic	Sanitation	Fact	Prevention
Theme-16	Panic	Situation	Wish	Panic	Coronavirus
Theme-17	Blaming	Homework	Situation	Wish	Avoid
Theme-18	Joke	Blaming	Coronavirus	Update	Food
Theme-19	Joke	Panic	Avoid	Cases	Situation
Theme-20	Annoyed	Annoyed	Stop spreading	Hospitalize	Coronavirus
	Cluster-6	Cluster-7	Cluster-8	Cluster-9	
Theme-1	Vacine	Ruin	Situation	Tourism	
Theme-2	News	Cases	Watch	Outbreak	
Theme-3	Message	Coronavirus	Virus	Situation	
Theme-4	Prevention	Awareness	Touch	Situation	
Theme-5	Dead	Wait & Things	Symptom	Quarantine	
Theme-6	News	Crisis	Problem	Education	
Theme-7	Panic	Symptom	Shot	Education	
Theme-8	Protection	News	Like	Virus	
Theme-9	Awareness	Symptom	Situation	Pandemic	
Theme-10	Situation	Infectious	Sick	Dead	
Theme-11	Thread	Expose	Dead	Education	
Theme-12	Wish	Caring	Body	Awareness	
Theme-13	Situation	Help & Need	Flu	Body	
Theme-14	Awareness	Protection	Wish	Need	
Theme-15	Message	Testing	Panic	Caring	
Theme-16	Situation	Blaming	Watch	Panic	
Theme-17	Media	Cure	Time	Fact	
Theme-18	Coronavirus	Message	Panic	Cases	
Theme-19	Cases	Stay Home	Contract	Public	
Theme-20	Health	Situation	Awareness	Exhibit	

Table 5
Negative themes of all significant clusters.

	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5	Cluster-6	Cluster-7
Theme-1	Financial crisis	Panic	Anxiety	Warning	Serious	Financial crisis	Worry
Theme-2	Panic	Media	Die	Avoid	Blaming	Hope	Excuse
Theme-3	Panic	Food	Panic	Warning	Message	Panic	Fake News
Theme-4	Situation	Jobless	Panic	Sick	Buy	Dead	Sad
Theme-5	Isolation	Restriction	Incur	Blaming	Hate	Situation	Situation
Theme-6	Stopping	Food	Panic	Situation	Avoid	Fever	Coronavirus
Theme-7	Disease	Situation	Panic	Covid	Stopping	Awareness	Media
Theme-8	Spreading	Food	Situation	Afraid	Infectious	Situation	Catch & Game
Theme-9	Situation	Jobless	Situation	Situation	Scare	Food	Ebola
Theme-10	Avoid	Situation	Panic	Blaming	Erazi	Lack of protection	Worst
Theme-11	Treatment	Panic	Situation	Crisis	Crisis	Need	Sick
Theme-12	Panic	News	Sick	Panic	Panic	Lockdown	Quarantine
Theme-13	Fear	Closing	Coronavirus	Die	Long lasting	Fear	Disease
Theme-14	Disease	Blaming	Situation	Spreading	Propaganda	Wrong	Scare
Theme-15	Situation	Social distance	Suffer	Treatment	Fake	Toilet	Panic
Theme-16	Situation	Panic	Situation	Danger	Lock	Hate	Covid
Theme-17	Habitual Fact	Non-Realiable	Panic	Fake News	Panic	Dead	Disease
Theme-18	Humor	Infectous	Situation	Wrong	Outbreak	Danger	Situation
Theme-19	Panic	Disease	Die	Treatment	Accept	Cold	Panic
Theme-20	Panic	Care	Fake News	Dead	Hope	Ebola	Annoy

CRediT authorship contribution statement

Md. Shahriare Satu: Conceptualization, Methodology, Resources, Data curation, Writing—original draft preparation, Visualization. **Md. Imran Khan:** Conceptualization, Methodology, Software, Data curation, Visualization. **Mufti Mahmud:** Methodology, Formal analysis, validation. **Shahadat Uddin:** Formal analysis, validation. **Matthew A. Summers:** Formal analysis, validation. **Julian M.W. Quinn:** Writing—review and editing. **Mohammad Ali Moni:** Writing—review and editing, supervision.

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.

References

[1] G. Lippi, M. Plebani, Procalcitonin in patients with severe coronavirus disease 2019 (covid-19): A meta-analysis, Clin. Chim. Acta; Int. J. Clin. Chem. (2020).

[2] R.-H. Xu, J.-F. He, M.R. Evans, G.-W. Peng, H.E. Field, D.-W. Yu, C.-K. Lee, H.-M. Luo, W.-S. Lin, P. Lin, et al., Epidemiologic clues to sars origin in China, Emerg. Infect. Diseases 10 (2004) 1030.

[3] E. Cambria, Affective computing and sentiment analysis, IEEE Intell. Syst. 31 (2016) 102–107, http://dx.doi.org/10.1109/MIS.2016.31.

[4] E. Cambria, A. Hussain, C. Havasi, C. Eckl, Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems, in: A. Esposito, N. Campbell, C. Vogel, A. Hussain, A. Nijholt (Eds.), Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March (2009) 23-27, Revised Selected Papers, in: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2010, pp. 148–156, http://dx.doi.org/10.1007/978-3-642-12397-9_12.

[5] H. Zhang, C. Wheldon, A.G. Dunn, C. Tao, J. Huo, R. Zhang, M. Prosperi, Y. Guo, J. Bian, Mining twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States, J. Am. Med. Inform. Assoc. 27 (2020) 225–235.

[6] A. Akay, A. Dragomir, B. Erlandsson, Network-based modeling and intelligent data mining of social media for improving care, IEEE J. Biomed. Health Inf. 19 (2015) 210–218.

[7] D.J. Fiander, Social media for academic libraries, in: Social Media for Academics, Elsevier, 2012, pp. 193–210.

[8] D.T. Nguyen, K.A.A. Mannai, S. Joty, H. Sajjad, M. Imran, P. Mitra, Rapid classification of crisis-related data on social networks using convolutional neural networks, 2016, arXiv preprint arXiv:1608.03902.

[9] M. Dragoni, S. Poria, E. Cambria, Ontosenticnet: A commonsense ontology for sentiment analysis, IEEE Intell. Syst. 33 (2018) 77–85, http://dx.doi.org/10.1109/MIS.2018.033001419, conference Name: IEEE Intelligent Systems.

[10] E. Cambria, Y. Li, F.Z. Xing, S. Poria, K. Kwok, Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 105–114, http://dx.doi.org/10.1145/3340531.3412003.

[11] S. Poria, N. Majumder, D. Hazarika, E. Cambria, A. Gelbukh, A. Hussain, Multimodal sentiment analysis: Addressing key issues and setting up the baselines, IEEE Intell. Syst. 33 (2018) 17–25, http://dx.doi.org/10.1109/MIS.2018.2882362, conference Name: IEEE Intelligent Systems.

[12] I. Chaturvedi, R. Satapathy, S. Cavallari, E. Cambria, Fuzzy commonsense reasoning for multimodal sentiment analysis, Pattern Recognit. Lett. 125 (2019) 264–270, http://dx.doi.org/10.1016/j.patrec.2019.04.024, URL: https://www.sciencedirect.com/science/article/pii/S0167865519301394.

[13] Q. Jiang, L. Chen, W. Zhao, M. Yang, Towards aspect-level sentiment modification without parallel data, IEEE Intell. Syst. 36 (2021) 75–81.

[14] A. Baird, E. Cambria, B.W. Schuller, Sentiment analysis and topic recognition in video transcriptions, IEEE Intell. Syst. 36 (2021).

[15] A. Khatua, A. Khatua, E. Cambria, A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks, Inf. Process. Manage. 56 (2019) 247–257, http://dx.doi.org/10.1016/j.ipm.2018.10.010, URL: https://www.sciencedirect.com/science/article/pii/S0306457317307495.

[16] S. Ahmed, A.N. Chy, M.Z. Ullah, Exploiting various word embedding models for query expansion in microblog, in: 2020 IEEE 8th R10 Humanitarian Technology Conference (R10-HTC), (ISSN: 2572-7621) 2020, pp. 1–6, http://dx.doi.org/10.1109/R10-HTC49770.2020.9357016.

[17] R.K. Behera, M. Jena, S.K. Rath, S. Misra, Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data, Inf. Process. Manage. 58 (2021) 102435, http://dx.doi.org/10.1016/j.ipm.2020.102435, URL: https://www.sciencedirect.com/science/article/pii/S0306457320309286.

[18] S.S. Aljameel, D.A. Alabbad, N.A. Alzahrani, S.M. Alqarni, F.A. Alamoudi, L.M. Babili, S.K. Aljaafary, F.M. Alshamrani, A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia, Int. J. Environ. Res. Public Health 18 (2021) 218, http://dx.doi.org/10.3390/ijerph18010218, URL: https://www.mdpi.com/1660-4601/18/1/218, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[19] E. Alomari, I. Katib, A. Albeshri, R. Mehmood, COVID-19: Detecting government pandemic measures and public concerns from Twitter arabic data using distributed machine learning, Int. J. Environ. Res. Public Health 18 (2021) 282, http://dx.doi.org/10.3390/ijerph18010282, URL: https://www.mdpi.com/1660-4601/18/1/282, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[20] M.S. Al-Rakhami, A.M. Al-Amri, Lies kill facts save: Detecting COVID-19 misinformation in Twitter, IEEE Access 8 (2020) 155961–155970, http://dx.doi.org/10.1109/ACCESS.2020.3019600, conference Name: IEEE Access.

[21] S. Boon-Itt, Y. Skunkan, Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study, JMIR Public Health Surveill. 6 (2020) e21978, http://dx.doi.org/10.2196/21978, URL: https://publichealth.jmir.org/2020/4/e21978/, company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Label: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc. Toronto, Canada.

[22] O. Gencoglu, Large-scale, language-agnostic discourse classification of tweets during COVID-19, Mach. Learn. Knowl. Extraction 2 (2020) 603–616, http://dx.doi.org/10.3390/make2040032, URL: https://www.mdpi.com/2504-4990/2/4/32, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[23] R. Kouzy, J. Abi Jaoude, A. Kraitem, M.B. El Alam, B. Karam, E. Adib, J. Zarka, C. Traboulsi, E.W. Akl, K. Baddour, Coronavirus goes viral: quantifying the covid-19 misinformation epidemic on twitter, Cureus 12 (2020).

[24] S. Kaur, P. Kaul, P.M. Zadeh, Monitoring the dynamics of emotions during COVID-19 using Twitter data, Procedia Comput. Sci. 177 (2020) 423–430, URL: http://www.sciencedirect.com/science/article/pii/S1877050920323243.

[25] R.J. Medford, S.N. Saleh, A. Sumarsono, T.M. Perl, C.U. Lehmann, An infodemic: Leveraging high-volume twitter data to understand public sentiment for the covid-19 outbreak, medRxiv (2020).

[26] T. Mackey, V. Purushothaman, J. Li, N. Shah, M. Nali, C. Bardier, B. Liang, M. Cai, R. Cuomo, Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on Twitter: Retrospective big data infoveillance study, JMIR Public Health Surveill. 6 (2020) e19509, http://dx.doi.org/10.2196/19509, URL: https://publichealth.jmir.org/2020/2/e19509/, company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Label: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc. Toronto, Canada.

[27] L. Nemes, A. Kiss, Social media sentiment analysis based on COVID-19, J. Inf. Telecommun. (2020) 1–15, http://dx.doi.org/10.1080/24751839.2020.1790793, publisher: Taylor & Francis _eprint.

[28] J. Samuel, G.G.M.N. Ali, M.M. Rahman, E. Esawi, Y. Samuel, COVID-19 public sentiment insights and machine learning for tweets classification, Information 11 (2020) 314, http://dx.doi.org/10.3390/info11060314, URL: https://www.mdpi.com/2078-2489/11/6/314, number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[29] X. Xiang, X. Lu, A. Halavanau, J. Xue, Y. Sun, P.H.L. Lai, Z. Wu, Modern senicide in the face of a pandemic: An examination of public discourse and sentiment about older adults and COVID-19 using machine learning, J. Gerontol: Ser. B (2020) http://dx.doi.org/10.1093/geronb/gbaa128.

[30] J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, T. Zhu, Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter, PLoS One 15 (2020) e0239441.

[31] J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, T. Zhu, Twitter discussions and emotions about the COVID-19 pandemic: Machine learning approach, J. Med. Internet Res. 22 (2020) e20550, http://dx.doi.org/10.2196/20550, URL: https://www.jmir.org/2020/11/e20550/, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc. Toronto, Canada.

[32] H. Yin, S. Yang, J. Li, Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media, in: X. Yang, C.-D. Wang, M.S. Islam, Z. Zhang (Eds.), Advanced Data Mining and Applications, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2020, pp. 610–623, http://dx.doi.org/10.1007/978-3-030-65390-3_46.

[33] X. Zhang, H. Saleh, E.M.G. Younis, R. Sahal, A.A. Ali, Predicting coronavirus pandemic in real-time using machine learning and big data streaming system, 2020, URL: https://www.hindawi.com/journals/complexity/2020/6688912/.

[34] R. Lamsal, Corona virus (covid-19) tweets dataset, 2020, http://dx.doi.org/10.21227/781w-ef42.

[35] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[36] K. Sangeetha, D. Prabha, Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for lstm, J. Ambient Intell. Humanized Comput. (2020) 1–10.

[37] K. Crockett, D. Mclean, A. Latham, N. Alnajran, Cluster analysis of twitter data: A review of algorithms, in: Proceedings of the 9th International Conference on Agents and Artificial Intelligence, Vol. 2, Science and Technology Publications (SCITEPRESS)/Springer Books, 2017, pp. 239–249.
[38] S. Ahuja, G. Dubey, Clustering and sentiment analysis on twitter data, in: 2017 2nd International Conference on Telecommunication and Networks, TEL-NET, IEEE, 2017, pp. 1–5.
[39] D. Godfrey, C. Johns, C. Meyer, S. Race, C. Sadek, A case study in text mining: Interpreting twitter data from world cup tweets, 2014, arXiv preprint arXiv:1408.5427.
[40] K. Lee, D. Palsetia, R. Narayanan, M.M.A. Patwary, A. Agrawal, A. Choudhary, Twitter trending topic classification, in: 2011 IEEE 11th International Conference on Data Mining Workshops, IEEE, 2011, pp. 251–258.
[41] V. Ong, A.D. Rahmanto, D. Suhartono, A.E. Nugroho, E.W. Andangsari, M.N. Suprayogi, et al., Personality prediction based on twitter information in Bahasa Indonesia, in: 2017 Federated Conference on Computer Science and Information Systems, FedCSIS, IEEE, 2017, pp. 367–372.
[42] B.Y. Pratama, R. Sarno, Personality classification based on twitter text using Naive Bayes, KNN and SVM, in: 2015 International Conference on Data and Software Engineering, ICoDSE, IEEE, 2015, pp. 170–174.
[43] M. Mccord, M. Chuah, Spam detection on twitter using traditional classifiers, in: International Conference on Autonomic and Trusted Computing, Springer, 2011, pp. 175–186.
[44] N. Mamgain, E. Mehta, A. Mittal, G. Bhatt, Sentiment analysis of top colleges in india using twitter data, in: 2016 International Conference on Computational Techniques in Information and Communication Technologies, ICCTICT, IEEE, 2016, pp. 525–530.
[45] T.R. Li, A. Chamrajnagar, X. Fong, N. Rizik, F. Fu, Sentiment-based prediction of alternative cryptocurrency price fluctuations using gradient boosting tree model, Front. Phys. 7 (2019) 98.
[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830, URL: http://jmlr.org/papers/v12/pedregosa11a.html.
[47] M.S. Satu, K.C. Howlader, M. Mahmud, M.S. Kaiser, S.M.S. Islam, J.M. Quinn, M.A. Moni, Short-term prediction of COVID-19 cases using machine learning models, Appl. Sci. (2021) [Online First article].
[48] M. Hung, E. Lauren, E.S. Hon, W.C. Birmingham, J. Xu, S. Su, S.D. Hon, J. Park, P. Dang, M.S. Lipsky, Social network analysis of COVID-19 sentiments: Application of artificial intelligence, J. Med. Internet Res. 22 (2020) e22590, http://dx.doi.org/10.2196/22590, URL: https://www.jmir.org/2020/8/e22590/, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc. Toronto, Canada.
[49] Z. Long, R. Alharthi, A.E. Saddik, Needfull – A tweet analysis platform to study human needs during the COVID-19 pandemic in New York state, IEEE Access 8 (2020) 136046–136055, http://dx.doi.org/10.1109/ACCESS.2020.3011123, conference Name: IEEE Access.
[50] A.S. Imran, S.M. Daudpota, Z. Kastrati, R. Batra, Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related tweets, IEEE Access 8 (2020) 181074–181090, http://dx.doi.org/10.1109/ACCESS.2020.3027350, conference Name: IEEE Access.
[51] M. Sethi, S. Pandey, P. Trar, P. Soni, Sentiment identification in COVID-19 specific tweets, in: 2020 International Conference on Electronics and Sustainable Communication Systems, ICESC, 2020, pp. 509–516, http://dx.doi.org/10.1109/ICESC48915.2020.9155674.