Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
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 Bishwavidyalay, 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

A R T I C L E   I N F O

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

A B S T R A C T

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
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. 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 TCustVID 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) TCustVID corresponding to the nine twitter experimental datasets.
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 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 non-symbolic 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. Baird 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 multimodal 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. Khattab 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-It 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 bitemper 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.
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_1, corona_tweets_1M, corona_tweets_2L, corona_tweets_2L.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 preprocessing 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
Find out the significant topics to COVID-19.

Input: Set of twitter dataset $D_s$, set of classifier $C$, the number of dataset $s$, the number of tokens $tokens$, set of cluster $Clust_{st}$, derived cluster $Clust_{sm}$, set of evaluation metrics $P_{jm}$, the number of topics $T_n$

Output: Find out the significant topics to COVID-19.

1: Begin
2: Cleaning Dataset $D_i[review]$ by removing tags, punctuation, characters and multiple spaces
3: $K ← 5$, $T_N ← 20$
4: for each Dataset $D_i ∈ D_s$, do
5: $D_i[tokens] ← Tokenize(D_i[reviews])$
6: $Clust[i][tokens] ← kmens[D_i[tokens], K]$
7: Replace $Clust[i][tokens]$ with $D_i[reviews]$
8: $Clust[i][tokens] ← Tokenize(Clust[i][tokens])$
9: end for
10: $m ← 0$
11: while $m = s$ do
12: for each Classifier $C_i ∈ C do$
13: for each Cluster $num_i ∈ K$ do
14: $P_{jm} ← Classification_{CV}(Clust_{sm})$
15: end for
16: end for
17: $P_{jm}^{max} ← maximum(P_{jm})$
18: Compare $P_{jm}^{max}$ with traditional classification
19: Find out $Clust_{jm}^{max}$ by considering $P_{jm}^{max}$
20: Divide $Clust_{jm}^{max}$ into $Clust_{pos}^{max}$, $Clust_{neu}^{max}$ and $Clust_{neg}^{max}$
21: $Topic_{pos} ← LDA(Clust_{pos}^{max}, T_N)$
22: $Topic_{neu} ← LDA(Clust_{neu}^{max}, T_N)$
23: $Topic_{neg} ← LDA(Clust_{neg}^{max}, T_N)$
24: Interpret $Topic_{pos}$, $Topic_{neu}$ and $Topic_{neg}$
25: Calculate top frequent topics from $Topic_{pos}$, $Topic_{neu}$ and $Topic_{neg}$
26: $m ← m + 1$
27: end while

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

Table 1

| 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 | 7,248,882 |
| corona_tweets_1M.db | 1,578,957 | Dataset-9 | 7,248,881 |

The 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.

For each $m$, $m ∈ [1, s]$

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

$$\text{AUC} = \frac{TP \text{Rate} + TN \text{Rate}}{2}$$

$$\text{F-measure} = 2 × \frac{\text{precision} × \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$

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.

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

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

  $$\text{AUC} = \frac{TP \text{Rate} + TN \text{Rate}}{2}$$

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

  $$\text{F-measure} = \frac{2 × \text{precision} × \text{recall}}{\text{precision} + \text{recall}} = \frac{2TP}{2TP + FP + FN}$$
• **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.

\[
G\text{Mean} = \sqrt{\text{TPRate} \times \text{TNRate}}
\]  

(4)

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

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}
\]  

(5)

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

\[
\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}
\]  

(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.
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,
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
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 overcome 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
about this model which is described as follows.

| Dataset  | Classifier | Accuracy | AUC | F-Measure | Sensitivity | Specificity | Cluster-01 | Cluster-02 | Cluster-03 | Cluster-04 | Cluster-05 | Cluster-06 |
|----------|------------|----------|-----|-----------|-------------|-------------|------------|------------|------------|------------|------------|------------|
| LSTM     | 0.927      | 0.907    | 0.926 | 0.927      | 0.927       | 0.868       | 0.983      | 0.979      | 0.983      | 0.983      | 0.974      |
| DT       | 0.915      | 0.901    | 0.916 | 0.915      | 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       | 0.816      | 0.739      | 0.816      | 0.816      | 0.610      |
| XGB      | 0.910      | 0.880    | 0.909 | 0.910      | 0.850       | 0.946       | 0.946      | 0.946      | 0.946      | 0.946      | 0.914      |
| LR       | 0.695      | 0.502    | 0.576 | 0.695      | 0.308       | 0.679       | 0.901      | 0.901      | 0.901      | 0.901      | 0.887      |
| MLP      | 0.940      | 0.901    | 0.901 | 0.940      | 0.940       | 0.901       | 0.901      | 0.901      | 0.901      | 0.901      | 0.901      |
| NB       | 0.654      | 0.503    | 0.597 | 0.654      | 0.352       | 0.644       | 0.644      | 0.644      | 0.644      | 0.644      | 0.644      |
| RF       | 0.924      | 0.898    | 0.924 | 0.924      | 0.873       | 0.957       | 0.957      | 0.957      | 0.957      | 0.957      | 0.932      |
| SVM      | 0.757      | 0.600    | 0.694 | 0.757      | 0.442       | 0.803       | 0.772      | 0.803      | 0.803      | 0.803      | 0.803      |
| XGB      | 0.787      | 0.657    | 0.750 | 0.787      | 0.527       | 0.854       | 0.840      | 0.854      | 0.854      | 0.854      | 0.854      |

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.

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.
5.2.2. Practical implications

- **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 analyses in various perspectives. Although, demographic analysis, comparison and discussion can give a concrete idea about various sources.
- 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.

### 5.3. Limitations of the study

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 2 (continued)**

| Dataset | Classifier | Accuracy | AUC | F-Measure | Sensitivity | Specificity | Dataset | Classifier | Accuracy | AUC | F-Measure | Sensitivity | Specificity |
|---------|------------|----------|-----|-----------|-------------|-------------|---------|------------|----------|-----|-----------|-------------|-------------|
| LSTM    | LSTM       | 0.903    | 0.919 | 0.903 | 0.903 | 0.936 | LSTM    | LSTM       | 0.908    | 0.929 | 0.908 | 0.908 | 0.951 | 0.955 | 0.955 | 0.955 | 0.975 |
| DT      | DT         | 0.908    | 0.929 | 0.908 | 0.908 | 0.951 | DT      | DT         | 0.908    | 0.929 | 0.908 | 0.908 | 0.951 | 0.955 | 0.955 | 0.955 | 0.975 |
| GB      | GB         | 0.664    | 0.718 | 0.666 | 0.666 | 0.773 | GB      | GB         | 0.664    | 0.718 | 0.666 | 0.666 | 0.773 | 0.810 | 0.830 | 0.806 | 0.810 |
| KNN     | KNN        | 0.889    | 0.915 | 0.889 | 0.889 | 0.942 | KNN     | KNN        | 0.889    | 0.915 | 0.889 | 0.889 | 0.942 | 0.941 | 0.954 | 0.941 | 0.967 |
| LR      | LR         | 0.451    | 0.538 | 0.380 | 0.451 | 0.624 | LR      | LR         | 0.451    | 0.538 | 0.380 | 0.451 | 0.624 | 0.548 | 0.598 | 0.501 | 0.548 |
| MLP     | MLP        | 0.768    | 0.813 | 0.768 | 0.768 | 0.859 | MLP     | MLP        | 0.768    | 0.813 | 0.768 | 0.768 | 0.859 | 0.885 | 0.885 | 0.885 | 0.925 |
| NB      | NB         | 0.219    | 0.501 | 0.083 | 0.219 | 0.783 | NB      | NB         | 0.219    | 0.501 | 0.083 | 0.219 | 0.783 | 0.220 | 0.503 | 0.094 | 0.220 |
| RF      | RF         | 0.909    | 0.931 | 0.909 | 0.909 | 0.954 | RF      | RF         | 0.909    | 0.931 | 0.909 | 0.909 | 0.954 | 0.954 | 0.954 | 0.954 | 0.975 |
| SVM     | SVM        | 0.353    | 0.517 | 0.353 | 0.353 | 0.681 | SVM     | SVM        | 0.353    | 0.517 | 0.353 | 0.353 | 0.681 | 0.299 | 0.539 | 0.251 | 0.299 |
| XGB     | XGB        | 0.635    | 0.705 | 0.632 | 0.635 | 0.774 | XGB     | XGB        | 0.635    | 0.705 | 0.632 | 0.635 | 0.774 | 0.815 | 0.843 | 0.814 | 0.871 |

- **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 perspectives. Although, demographic analysis, comparison and discussion can give a concrete idea about various sources.
- 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.

### 5.4. 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.
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 TCltustVID 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 TCltustVID that has given the best performance outcomes in sentiment analysis and topics modeling by analyzing COVID-19 twitter datasets compared to other methods. TCltustVID 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 TCltustVID 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 TCltustVID (and improved versions of TCltustVID) 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       |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Financial lose  | Warning         | Outbreak        | Situation       | Awareness       |
| Fact            | Food            | Sharing         | Panic           | Infectious      |
| Warning         | Situation       | Wish            | Situation       | Situation       |
| Estimate        | Situation       | Gonna           | Entertainment   | Need            |
| Blaming         | Testing         | Caring          | Protection      | Wish            |
| Pleased         | Rumor           | Caring          | Dead            | Food            |
| Financial lose  | Warning         | Panic           | Health          | Break           |
| Disease         | Visitors        | Survive         | Stay Home       | Treatment       |
| Awareness       | Joe             | Awareness       | Avoid           | Want            |
| Disease         | Panic           | Treatment       | Fact            | Prevention      |
| Warning         | Situation       | Playing game    | Awareness       | Awareness       |
| Caring          | Panic           | Coronavirus     | Protection      | Panic           |
| Panic           | Closing         | Home work       | Awareness       | Situation       |
| Panic           | Panic           | Ramadan news    | Situation       | Awareness       |
| Awareness       | Panic           | Sanitation      | Fact            | Prevention      |
| Panic           | Situation       | Wish            | Panic           | Coronavirus     |
| Blaming         | Homework        | Situation       | Wish            | Avoid           |
| Joke            | Blaming         | Coronavirus     | Update          | Food            |
| Joke            | Panic           | Avoid           | Cases           | Situation       |
| Annoyed         | Annoyed         | Stop spreading  | Hospitalize     | Coronavirus     |

| Cluster-6       | Cluster-7       | Cluster-8       | Cluster-9       |
|-----------------|-----------------|-----------------|-----------------|
| Vaccine         | Ruin            | Situation       | Tourism         |
| News            | Cases           | Watch           | Outbreak        |
| Message         | Coronavirus     | Virus           | Situation       |
| Prevention      | Awareness       | Touch           | Situation       |
| Dead            | Wait & Things   | Symptom         | Quarantine      |
| News            | Crisis          | Problem         | Education       |
| Panic           | Symptom         | Shot            | Education       |
| Protection      | News            | Like            | Virus           |
| Awareness       | Symptom         | Situation       | Pandemic        |
| Situation       | Infectious      | Sick            | Death           |
| Thread          | Expose          | Dead            | Education       |
| Wish            | Caring          | Body            | Awareness       |
| Situation       | Help & Need     | Flu              | Body            |
| Awareness       | Protection      | Wish            | Need            |
| Message         | Testing         | Panic           | Caring          |
| Blaming         | Situation       | Blaming         | Watch           |
| Media           | Cure            | Time            | Fact            |
| Coronavirus     | Message         | Panic           | Cases           |
| Cases           | Stay Home       | Contract        | Public          |
| 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       |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| Financial crisis| Panic           | Anxiety         | Warning         | Serious         | Financial crisis| Worry           |
| Panic           | Media           | Die             | Avoid           | Blaming         | Hope            | Excuse          |
| Panic           | Food            | Warning         | Message         | Panic           | Fake News       |
| Panic           | Food            | Panic           | Buy             | Dead            | Sd              |
| Panic           | Jobless         | Sick            | Blaming         | Hate            | Situation       | Situation       |
| Isolation       | Restriction     | Incur           | Blaming         | Hate            | Situation       |
| Stoping         | Food            | Panic           | Situation       | Avoid           | Fever           | Coronavirus     |
| Disease         | Situation       | Panic           | Covid           | Stopping        | Awareness       | Media           |
| Disease         | Situation       | Afraid          | Infectious      | Situation       | Catch & Game    |
| Spread           | Food            | Situation       | Situation       | Scare           | Food            | Ebola           |
| Jobless         | Situation       | Panic           | Blaming         | Eazi            | Lack of protection| Worst          |
| Avoid           | Situation       | Panic           | Blaming         | Eazi            | Lack of protection| Worst          |
| Treatment       | Panic           | Sick            | Panic           | Lockdown         | Quarantine      |
| Panic           | News            | Sick            | Long lasting    | Fear            | Disease         |
| Panic           | Panic           | Coronavirus     | Die             | Long lasting    | Fear            | Disease         |
| Panic           | Blaming         | Situation       | Spreading       | Propaganda      | Wrong           | Scare           |
| Panic           | Social distance | Suffer          | Treatment       | Fake            | Toilet          | Panic           |
| Panic           | Situation       | Panic           | Danger          | Lock            | Hate            | Covid           |
| Panic           | Habitual Fact   | Non-Realiable   | Panic           | Fake News       | Panic           | Deed            | Disease         |
| Panic           | Infection       | Situation       | Wrong           | Outbreak        | Danger           | Situation       |
| Panic           | Disease         | Die             | Treatment       | Accept           | Cold            | Panic           |
| Panic           | Care            | Fake News       | Dead            | Hope            | Ebola           | Annoy           |

CRediT authorship contribution statement

Md. Shahriar 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: Methodology, 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/10.1007/978-3-642-1397-9_12.

[5] H. Zhang, C. Wheldon, F. Mohammad, A. Nijholt, J. Bunt, 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. 25 (2018) 1163–1171, http://dx.doi.org/10.1136/amiajnl-2018-013957.

[6] S. Aljameel, D.A. Alabbad, N.A. Alzahrani, S.M. Alqarni, F.A. Alamoudi, L.M. Alshammari, F.K. Alhashmi, A. Almarhoon, 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) 2184, http://dx.doi.org/10.3390/ijerph18010218, URL: https://www.mdpi.com/1660-4601/18/1/218, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[7] E. Alomari, I. Karib, A. Albeshr, 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) e202100550, http://dx.doi.org/10.3390/ijerph180100550, URL: https://www.mdpi.com/1660-4601/18/1/218, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[8] M.S. Al-Rahami, 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.2020.3019460, conference Name: IEEE Access.

[9] S. Boon-Itt, Y. Sunkanun, Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study, JMIR Public Health Surveil. 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 Publisher: JMIR Publications Inc Toronto, Canada.

[10] 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/make20040032, URL: https://www.mdpi.com/2594-4902/2/4/32, number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

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

[12] S. Kaur, P. Kaushik, P.M. Zadeh, Monitoring the dynamics of emotions during COVID-19 using the Twitter data, Comput. Sci. 177 (2020) 423–430, URL: http://dx.doi.org/10.1016/j.csic.2020.03.021.

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

[14] 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 Surveil. 6 (2020) e19509, http://dx.doi.org/10.2196/19509, company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc Toronto, Canada.

[15] L. Nemes, A. Kiss, Social media sentiment analysis based on COVID-19, J. Inf. Technol. Commun. (2020) 1–15, http://dx.doi.org/10.3390/information11030218, number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[16] 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, number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[17] J. S. Kaur, P. Kaul, P.M. Zadeh, Monitoring the dynamics of emotions during COVID-19 using the Twitter data, Comput. Sci. 177 (2020) 423–430, URL: http://dx.doi.org/10.1016/j.csic.2020.03.021.

[18] 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, R. Yang, R. Li, J. Yao, J. Wu, J. Li, J. Xue, J. Chen, J. Chen, Z. 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.

[19] J. Xue, J. Chen, R. Hu, C. Chen, Z. 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, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Publisher: JMIR Publications Inc Toronto, Canada.

[20] 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 Surveil. 6 (2020) e19509, http://dx.doi.org/10.2196/19509, company: JMIR Public Health and Surveillance Distributor: JMIR Public Health and Surveillance Institution: JMIR Public Health and Surveillance Publisher: JMIR Publications Inc Toronto, Canada.

[21] S. Boon-Itt, Y. Sunkanun, Public perception of the COVID-19 pandemic on Twitter: Sentiment analysis and topic modeling study, JMIR Public Health Surveil. 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 Publisher: JMIR Publications Inc Toronto, Canada.

[22] S. Aljameel, D.A. Alabbad, N.A. Alzahrani, S.M. Alqarni, F.A. Alamoudi, L.M. Alshammari, F.K. Alhashmi, A. Almarhoon, 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) 2184, http://dx.doi.org/10.3390/ijerph18010218, URL: https://www.mdpi.com/1660-4601/18/1/218, number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
[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. Abuja, 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. Andangsi, 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, ICOSE, 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, ICTICT, 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) [OnlineFirst 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. Daadpota, 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.