Textual analysis of Covid-19: A Review

Omswroop Thakur[1], Sri Khetwat Saritha[1], Sweta Jain[1]

{thakuromswroop@gmail.com, sksarithamanit@gmail.com, shweta_j82@yahoo.co.in}

Department of CSE, Maulana Azad National Institute of Technology Bhopal, India[1]

Abstract. The evaluation of the sentiments from COVID-19 text data has received a lot of importance seeing the current situation of pandemics. This research focuses on the evaluation of sentiments of COVID-19 text data, which is effective for analyzing information in tweets where opinions are either negative, neutral, or positive. Social media platforms are conveying a variety of sentiments as well as a variety of emotions in different events of the outbreak. The intention of this paper is to review the related research in the field of textual analysis such as topic modeling, emotion detection, sentiment analysis, and text summarization of COVID-19 text data. And also, along with this, we differentiate earlier techniques of mining opinions/sentiments, such as machine learning, deep learning, and lexicon-based approaches, as well as their evaluation methodologies. After investigating and reviewing the various methods, in totality, the BERT based model gives better performance than other approaches and BernoulliNB based model gave the least performance among the methodologies. Since, manual evaluation or judgment of the sentiments or emotions of the huge amount of textual content of COVID-19 text can be very challenging and infeasible due to the enormous outspread of the COVID-19 disease and thus, the need for comprehensive and systematic analysis of vast text data is the research motivation of the work.

Keywords: NLP, Sentiment Analysis,COVID-19,LSTM,RoBERTa,LDA,Top2vec

1 Introduction

The COVID-19 (coronavirus disease 2019) pandemic has affected the day-to-day life of individuals around the world. During lockdown as well as general phases, people around the globe use social media networks to convey their thoughts, feelings, and viewpoints about the pandemic that has impeded their daily lives. There is the usage of online websites such as Google plus, Twitter as well as websites such as Reddit, and Facebook. To share and express their ideas on a variety of issues, participate in online discussions, and send messages all over the world. Also, the information can be variable and unstructured as well. The novel coronavirus named SARS-CoV-2 gave rise to the COVID-19 pandemic which struck China in the month of December 2019. Millions of confirmed cases and a large number of deaths have been reported globally. Technology has transformed the approaches in which people express and distribute the

THEETAS 2022, April 16-17, Jabalpur, India

DOI 10.4108/eai.16-4-2022.2318113

conceptions and beliefs that they have in today's world. These are currently done by means of various ways such as blogging, sharing opinions in forums present on the internet, websites, and allowing users to share their thoughts. These communities which are present online acquire a view, where the user impacts other users with these online communities through tweets and updates that they post on social media. These social networking platforms create very large amounts of data that are filled with opinions. For this purpose, several methods of analysis of sentiments are utilized

The novel coronavirus named SARS-CoV-2 gave rise to the COVID-19 pandemic which struck China in the month of December 2019. As per the WHO, millions of numbers of confirmed cases of COVID-19 and a large number of deaths have been reported globally. Large dispersal of COVID-19 news is taking place across social media and news websites. As a result, the social media platforms are encountering as well as conveying a variety of sentiments, viewpoints as well as a variety of emotions in different events of the outbreak. These big data are very crucial resources for the researchers as well as the computer scientists along with organizations that are trying to overcome the pandemic, for studying the sentiment of people related to current incidents, particularly those incidents related to the COVID-19 pandemic. Thus, notable results can be yielded by the analysis of these sentiments. This review will give insights into the various studies which are done on the COVID-19 text data and will give an overview of the findings.

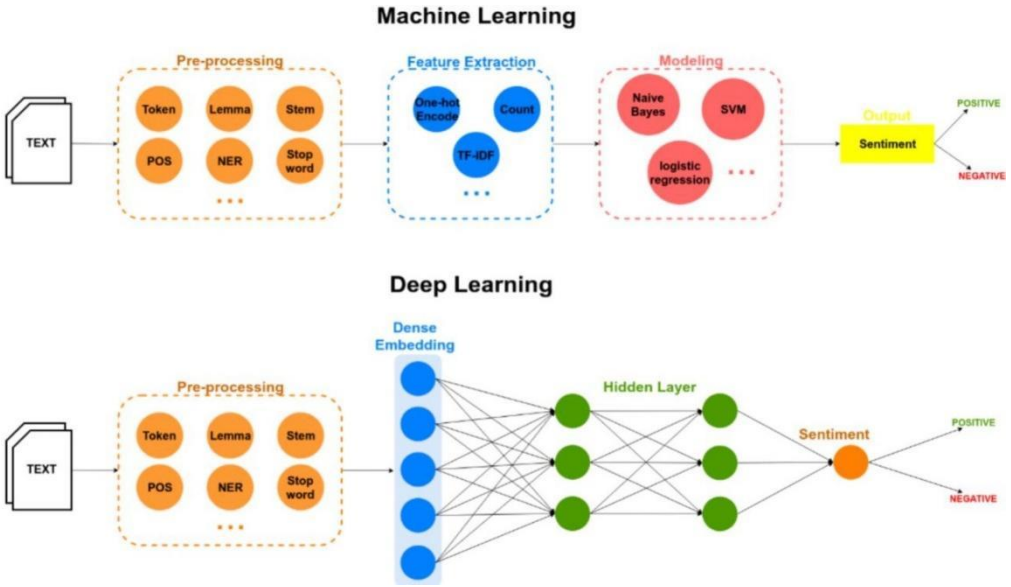


Fig. 1. Sentiment Analysis

2 Sentiment Analysis

Sentiment analysis is a process used for evaluating and obtaining the view, opinions, emotions as well as perspectives by using the Natural language Processing technique from database

sources, tweets, text as well as speech. In the Sentiment analysis procedure, the task of classifying the views into different classes like “negative", "positive”, and “neutral”.

Sentiment Analysis Extracts opinions and trends from data through four processes:

. The first process is pre-processing as shown in Fig 2, in this step task of cleaning the data, then transforming the data, and lastly, data reduction [1] is performed.

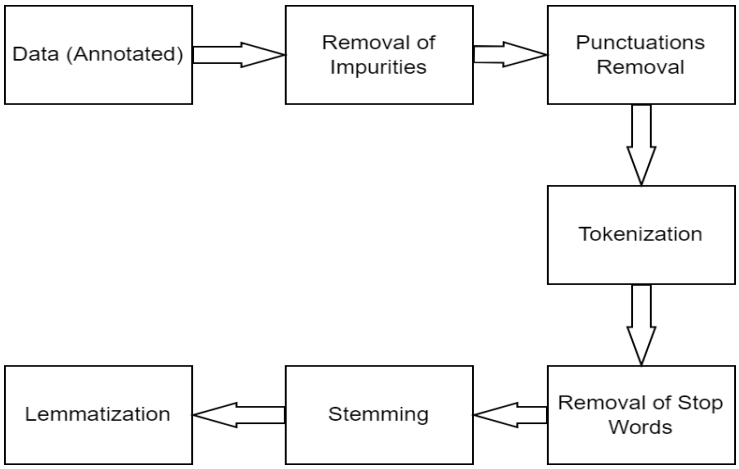


Fig. 2. Data preprocessing.

These steps ensure that the raw data is converted into a format of data that is machine-readable. The deletion of symbols such as hashtags (“#”) and various other special characters along with the removal of all the URLs takes place in the data cleaning step. In order to save space as well as time, the stop words are removed in the data preprocessing step. Since the stop words are basically the commonly utilized words like-a, in, the, an, etc[1]. These words can be excluded by the search engine when retrieving them as result.

The second process is feature extraction; it aims to extract important features for training. The more extracted features the more accurate the classification results [2]. Several features can be extracted, including the following: Sentiment features (SENF), which are related to the positivity and negativity of words and emotions (e.g., number of positive and negative words or emotions in a text), Syntax-based features (SYNF): The syntax features are the features that deals with features related to syntactical features such as quotation marks, and other features like question mark along with features such an exclamations mark. Also, along with these their number of occurrences is also considered. Semantic features (SEMF): The semantic features deal with the logic of the sentences. These logics include the forms such as passive, and active forms. Unigram-based features (UGF): It includes hypernyms (i.e., more generals) and hyponyms (i.e., more specific) features. These features act as seed words of the data or the input which is given by the user. For example, the color is a hypernym of red. N-gram features (NGF): These are the features that are obtained by putting N consecutive words together. Bigram features (BGF) are created by combining two consecutive words (BGF) [3] Top words feature (TWF): extract terms that occur frequently in the text. Pattern-based features (PTF): These features to procure the patterns utilizes the Part-of-Speech tags such as positive and negative verbs, adjectives, pronouns, etc.

The third process is feature selection (filtering), which is used to minimize the size of features so that the classification model's accuracy and speed can be improved. The most extensively used feature selection method is known as Frequency-Inverse Document Frequency (TF-IDF) [4]. It estimates majorly the commonly used terms (TF) as well as the frequency with which the phrase appears in a text (IDF). The terms with the highest TF-IDF (i.e., the product of TF and IDE) scores are those that appear the most frequently and contain the most important information on a certain topic.

The fourth process is classification, it is the process of dividing the text into distinct categories. Examples of various classifiers consist of a Lexicon classifier [5]: a lexicon can be defined as a group of words where the polarity of the word is predefined. Naïve Bayes (NB) [6]: The Naïve Bayes classifier can be described as a classifier that evaluates a chunk of probability by the process of counting the frequency and the combination of the data set values. SVM (Support Vector Machine) [7] can be described as a tool for data mining for the work of novelty detection along with works such as classification as well as regression. Support Vector machines have been used successfully in a variety of applications, including text categorization along with other applications such as face detection as well as text categorization as well as particle identification. The classification has 3 types [8]. The first category is binary classification, it is categorizing emotions into two basic polarities: positive (high positive scores) and negative (low positive scores) (i.e., high negative scores). The second category is ternary classification, in this classification, sentiment is classified as neutral as well along with negative and positive. It means that the sentiment is neither positive nor negative. The third category is multiclass classification, it is the classification of sentiments into multiple predefined classes to extract not only positivity or negativity but also to extract feelings and opinions. Furthermore. The classes may be defined to classify texts with happiness, enjoy, hate, etc.

3 Sentiment Classification Evaluation

The following equations which are basically indexes [9,10] can be utilized to assess the performance of the classification of the sentiments:

$$\text{Recall} = TP/(TP+FN) \tag{1}$$

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \tag{2}$$

$$F1 = (2\times\text{Precision}\times\text{Recall})/(\text{Precision}+\text{Recall}) \tag{3}$$

$$\text{Precision} = TP/(TP+FP) \tag{4}$$

Here the term TP is the total instances that are true positive, and FN is the total instances which are false negative, it is the total instances which are false-positive along with this TN is the total instances which are true Negative.

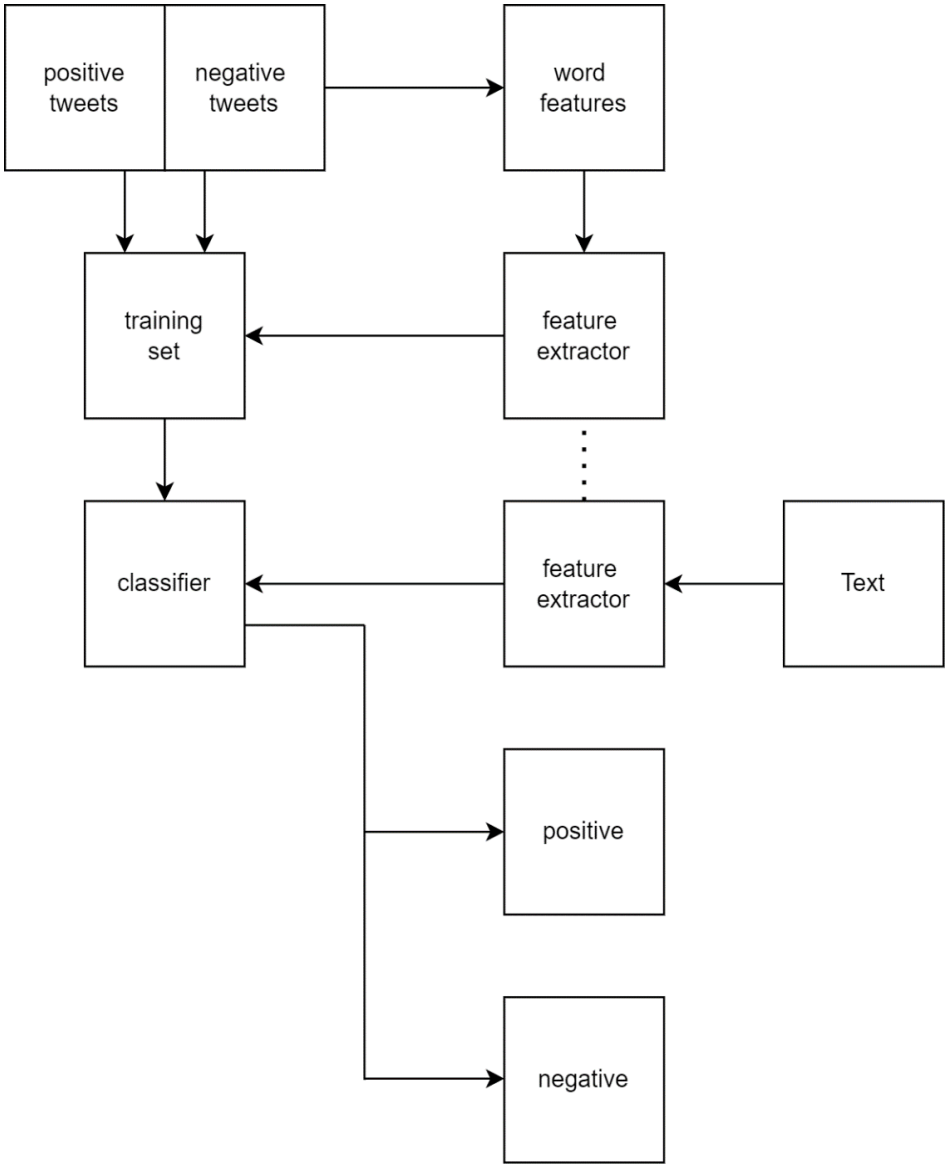


Fig. 3. Sentiment Analysis Architecture.

4 Related Works

This section involves various works which have been done in the field of COVID –19 textual analysis such as topic modeling, emotion detection, and Sentiment classification where various methodologies have been used. Also, a table (Table 1) is included in this review which further describes and compares related works. Table 1 gives insights into various datasets found in the

related works, Table 1 consists of datasets, classification methods on the basis of feature extraction methods used, classification algorithms used, and achieved accuracy on the respective datasets.

Ali Shariq Imran [11] on the Kaggle dataset and Trending hashtag # data, proposed a Deep learning LSTM technique as shown in Fig 4. The pre-training step comprises of weight initialization on the input data in an unsupervised manner using generative deep belief networks (DBN), followed by greedy network training using a restricted Boltzmann machine (RBM), described as:

$$E(v,h) = -\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{v_k}{\sigma_k}h_l w_{kl} - \sum_{k=1}^{K}\frac{(v_k - a_k)^2}{2\sigma_k^2} - \sum_{l=1}^{L}h_l b_l \quad (5)$$

where σ_k is the standard deviation, w_{kl} is the weight value connecting visible units v_k , and the hidden units h_l , a_k , and b_l are the biases for visible and hidden units, respectively. To classify the sentiment polarity of data along with the emotions an LSTM model having multiple layers have been utilized and it has an accuracy of 82.4 percent. They gave significant insights into the reactions of people on the social platforms related to the COVID-19 pandemic. Also, they also validated the emotions that are generated through emoticons. The goal of this study is to look at how folks from various cultures reacted to the new Coronavirus and how they felt about various governments' subsequent responses. They wanted to see if there was a link between people's feelings and emotions from within. Jim Samuel [12] created a paradigm called PSS which stands for polarity-based public sentiment scenarios. This paradigm will also be useful for calamities that may occur in the coming times ahead. That is, this paradigm will be useful after the COVID pandemic ends. The authors used n-grams word associations, descriptive tweet analysis, illustrative tweets, sentiment analysis, and statistical sentiment analysis. Word frequency and N-grams analysis were utilized to study the text corpus of all the tweets, to discover dominant patterns. They identified 1794 Twitter iPhone users, as well as 621 Twitter for Android users, in their dataset and ignored smaller classes, such as users of web client technologies. According to this study, tweets about the reopening comprised mainly three emotions, the sentiments of negative polarity comprise of 36.82 % of the total sentiments, whereas the sentiments of positive polarity consisted of 48.27 %, and with the neutral polarity of 14.98% respectively.

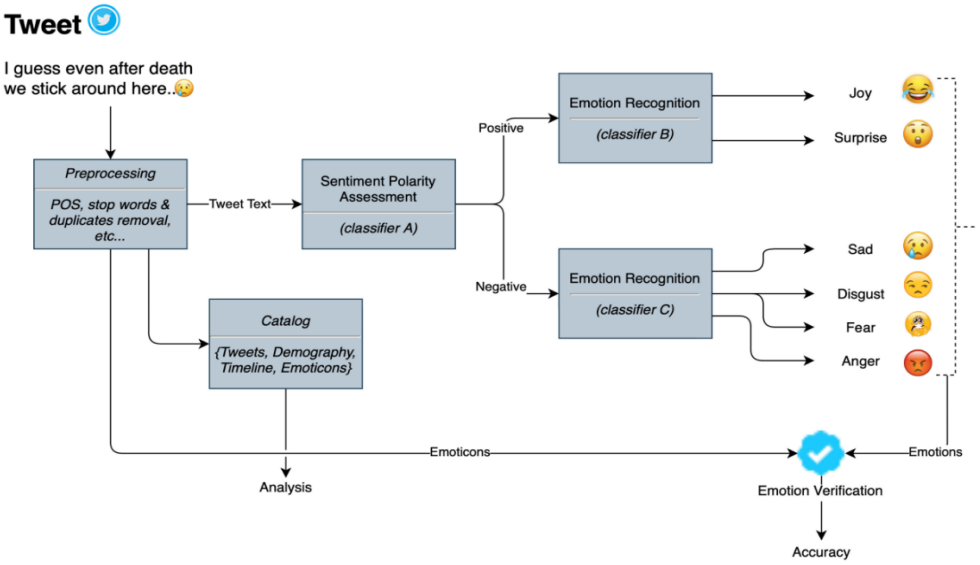
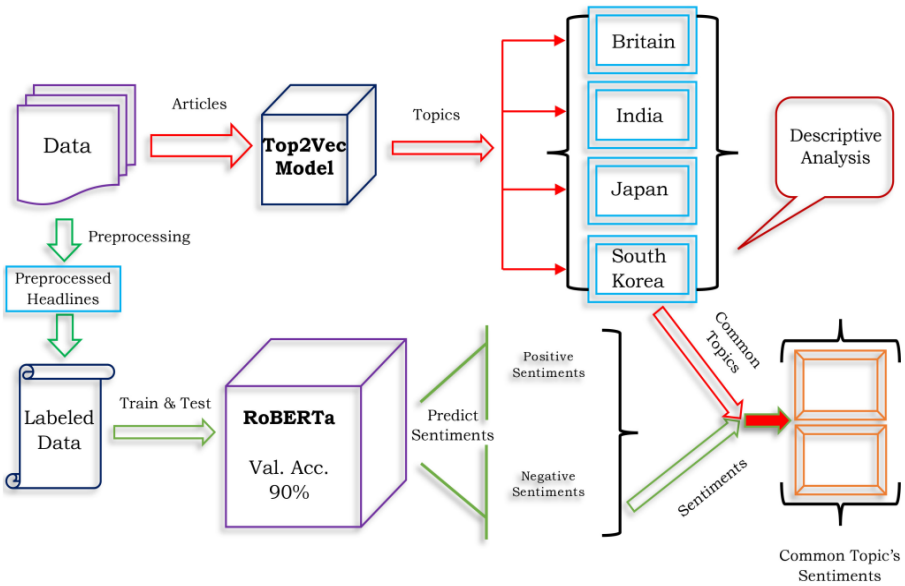


Fig. 4. Tweets' Sentiment and Emotion Analyser [11].

The goal of this study is to provide the detection of critical and exigent insights which will help in providing the solutions to the problems that people are facing during the pandemic, also it will help in the recovery of many places in the USA and world are moving towards reopening phase.To obtain the needs of individuals during the coronavirus pandemic who are in New York State, Zijian Long [13] Proposed a Platform for the analysis of tweets by utilizing the Support vector machine algorithm. In order to achieve this, they created 4 modules for the platform. The first module collects the data, the second module does the task of storing data, the third module analyzes the data and the fourth module is used for the presentation of data. Based on the inquest of the users, these four modules give the user a thorough analysis of their queries and their needs. The hashtags #virus was utilized along with #COVID-19 as well as hashtags such as #outbreak and #coronavirus along with #pandemic are utilized. Also, in the module which performs the task of collection of data the area to search was given as New York State in order to get as many tweets linked to this area as possible. According to the findings, the most prevalent human need in tweets during the pandemic time is relatedness (47.42 percent), followed by autonomy (16.63 percent), and competence (14.51 percent). This demonstrates that people's basic needs were not met during this time, and dissatisfaction levels were at an all-time high. Piyush G [14] proposed a technique in which they collected the data of headlines and news articles from the news websites and for determining the topics on these datasets, they applied the model named top2vec as shown in Fig 5. After that, they performed sentiment analysis by applying two steps which includes the creation of a dataset which is labeled by the method of unsupervised machine learning, and after that, this labeled dataset was trained and tested by applying the algorithm of Roberta. And this methodology attained 90 percent accuracy. The issues that were majorly reported in the four countries were Economy along with topics such as Education as well as Sports, also the topic modeling experiments show that the United Kingdom has the most negative news that was related to Coronavirus. The sectors which were badly hit by the COVID-19 pandemic were sports along with other sectors such as economy as well as study. News articles greater than 100,000 which were linked to Coronavirus were gathered by them from these 4 countries of 11 months duration.



. **Fig. 5.** News articles topics and sentiment analyzer [14].

Hamed Jelodar [15] proposed an approach which utilizes Long Short-Term Memory (also called LSTM) RNN which is based on Natural language Processing to do Sentiment Classification which is based on Deep learning as shown in Fig 6. Along with this Discovery of topics on online discussions which are based on COVID-19 as well as Coronavirus. Accuracy of 81.15 percent was accomplished by this methodology. In this study data from healthcare forums/websites like Reddit in which there are subreddits have been utilized. On these subreddits, the LSTM model which is based on Natural language processing has been applied and the detection of hidden topics that possess some meaning and classification of sentiment-comment based on the subreddits was the aim of the research. The limitations of this research are that they have collected comments from a limited duration which is between 20/01/2020 to 19/02/2020.For the process of semantic extraction and discovery of topics that are hidden, they have used latent Dirichlet allocation topic model along with Gibbs sampling on comments which are related to coronavirus. They used the model named latent Dirichlet allocation to consider a set of documents as topics (K), such as remarks related to coronavirus and words, with discrete distributions of topics chosen from a symmetric Dirichlet distribution. The words which were most highlighted were risk word-weights of 0.0218%, covid with word-weights of 0.0299%, and young with word-weights of 0.0222% and these topics were having negative sentiment polarity. Along with these terms, the terms "coronavirus" with word-weights of 0.0353% and "quarantine" with word-weights of 0.0346% were also the words that were majorly highlighted. M Bahja [16] created a methodology that examined COVID-19 tweets tying the epidemic to 5G in order to detect reoccurring themes and subjects. The task of collection of tweets was performed by utilizing the API (application programming interface) of Twitter, the date from which the task of collection of tweets took place is 28/01/2020.. Social network analysis along with other techniques such as latent Dirichlet allocation and techniques such as sentiment analysis was utilized for tweets analysis and detection of topics. Each of the eight Latent Dirichlet allocation (LDA) analytic investigations revealed 20 subjects, each represented by a

20-word distribution. After each LDA attempt had identified the subjects, the distribution of words in each of the topics was manually labeled as well as evaluated. The dataset indexes almost 50 million COVID-19-related tweets.

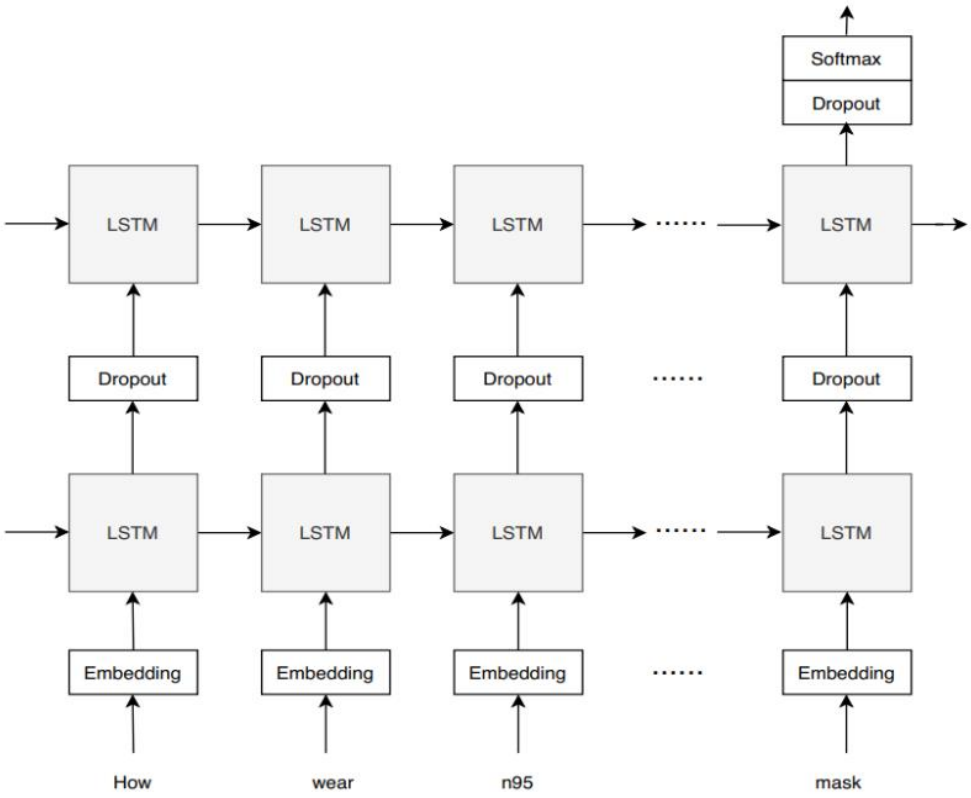


Fig. 6. Structure of the LSTM designed for COVID-19 sentiment classification [15]

.Here the implementation of the algorithm Latent Dirichlet allocation was restricted to twenty topics because after the twenty topics, word distribution does not give insights that are relevant and word distribution turns repetitive. Although the optimal number was N_topic = 35, it was still restricted to twenty topics. Using the LDA model, Richard F. Sear [17] presented a model that evaluated the opinion war between anti-vaccination communities, that is the communities that are against the vaccination, and pro-vaccination communities, that is the communities that are in favor of vaccination on the social media site Facebook. The dataset was divided into two equal-length durations: Time duration 17 January 2020 to 7 February 2020 which is T1, along with time duration 7 February 2020 to 28 February 2020 which is T2.T1 basically correlates to when COVID-19 was mostly considered a problem in the continent of Asia, whereas T2 approximately correlates to when it became a severe problem in Europe. The snowball method has been used to collect the content which is present in the web of these groups on the internet by starting a collection from a seed of pages that are identified manually related to the discussion on the pro-vaccination communities as well as anti-vaccination communities' discussion along with other pages related to the regulation of vaccinations and also pages related to vaccines. The research was conducted worldwide, not just in one location. There was separate packing of the

contents for the pro-vaccination communities as well as the anti-vaccination communities as clusters and machine learning was used to assess the two sets of content. The pro-vaccination community is involved in less discussion as compared to the anti-vaccination community related to coronavirus, this is overserved in this research. X.Wan [18] proposed a system that filters articles from numerous digital sources using the concept of NLP and converts them into information that can be understood by the user. They also modeled the topics from various sources that are frequently debated to help the users to achieve opinions about the serious such as coronavirus pandemic. They implemented these by utilizing the given steps. At first, they gathered a large dataset of articles that are linked with the coronavirus pandemic. After that, they summarized the articles by applying techniques such as unsupervised learning as well as supervised learning, then they utilized techniques for community detection to cluster the data based on similarities present among them. Then, they made use of an algorithm called the BART algorithm to detect every cluster's topic. And achieved the highest ROUGE score of 50.45 on New York Times articles. R. L. Rosa et al [9] proposed a system that performs the task of detection of events at the very initial phase of that particular event. There are basically 5 modules in this proposed system, the first module performs the task of identifying the location of the individual. After that, the second module performs the task of deriving the messages which are present in Online Social Networks. Then the third modules utilize the Deep Belief Network which is also called DBN, to identify the topics by utilizing the NLP techniques. To distinguish the parameters of the DBN, the function parameter is defined as sθ. Thus, the output of the softmax regression (hsθ) is represented by:

$$h_{s\theta}=\begin{bmatrix} p(y^{(i)}=1)|m^{(i)}; & s\theta \\ p(y^{(i)}=2)|m^{(i)}; & s\theta \\ \vdots & \\ p(y^{(i)}=l)|m^{(i)}; & s\theta \end{bmatrix}=\frac{1}{\sum_{t=1}^{l} e^{s\theta_t^T m^{(i)}}}\begin{bmatrix} e^{s\theta_1^T m^{(i)}} \\ e^{s\theta_2^T m^{(i)}} \\ \vdots \\ e^{s\theta_l^T m^{(i)}} \end{bmatrix} \tag{6}$$

After that, for the optimization of cost function, the gradient descent algorithm and the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm were used which is represented by :

$$J(s\theta)=\frac{-1}{m}\sum_{i=1}^{m}\sum_{t=1}^{k} y_t^i \log(t_{s\theta t})+\frac{\lambda}{2}\sum_{i=1}^{l}\sum_{t=0}^{n}(s\theta_{it})^2+\frac{\lambda}{2}\sum_{p=1}^{q-1}\theta_p^2 \tag{7}$$

where k is the number of subtopics, m denotes the number of previous computations to be stored, $y_t^{(i)}$ is the label of t class, $t_{s\theta t}$ is the output of the softmax regression, $s_{\theta it}$ represents the parameters of the model, λ is the penalty factor, and θ_p is the weight of the RBMs in the DBN. Then the fourth module performs the task of detecting changes in the behavior of the person in the Online Social Networks. At last, the fifth module performs the task of recognizing emotion and performs an effective analysis of it using the tree-CNN. This methodology achieved an accuracy of 87% for the US election dataset, the accuracy of 86% for the US Manhattan dataset, and accuracy of 93% for the COVID-19 symptoms dataset. A. Mourad [19] proposed a heuristic study that was conducted during the coronavirus pandemic, to provide a quantitative estimation

of the infodemic caused by the social media platforms. This research studied the good and bad impact of social media platforms on beating the coronavirus pandemic. They proposed a methodology that is based on mixed ontology data analytics techniques and experiments that target the tweet contexts as well as profiles of individuals by following two steps methodology. The first step is the identification of tweets that are making use of coronavirus to propagate the information that is misleading to the user. The second step is the evaluation of the authenticity of the coronavirus information which is being circulated by identification of the tweets for their source of the user specialty. These works also provided detailed descriptions of computing as well non-computing techniques. The results conveyed that around 16.1 percent of tweets about coronavirus are being diverted to things that are not in the context of coronavirus or for the purpose of advertisement. Also, around 93.7 percent of the coronavirus tweets are propagating data that are from unverified sources. T. Da and L. Yang [20] proposed a model with three steps that sequentially performs the following steps. The first step includes the task of labeling the tweets of Sina Weibo by utilizing the Sentiment Knowledge Enhanced pre-training (SKEP) which is a state-of-the-art pre-trained model. This SKEP makes rational predictions and delivers results in a timely manner. SKEP extracts sentiment knowledge using a straightforward and obvious method known as point mutual information (PMI), which is frequently used in information retrieval. PMI determines whether a pair of words are seen together more frequently than if they are seen separately. The PMI score is determined by the following formula:

$$PMI(x1, x2) = log\ log\ \frac{p(x1,x2)}{p(x1).p(x2)} \quad (8)$$

The second step includes the usage of random forest (RF) as well as linear probit model (LP) to demonstrate individual word influence. These two models are used because the SKEP model is a neural network and it doesn't provide us information on each feature (that is a word) that contributes to the assignment of labels. The third step includes the usage of a linear regression algorithm to create the inverse relationship between the mean sentiment responses and the locally confirmed cases of coronavirus. This was done by aggregation of tweets and labels at the province level. The results obtained in this step confirmed that an area will suffer a drastic decline in the polarity of sentiment towards pessimism when the number of coronavirus cases in that area. This proposed methodology achieved a precision of 81.39 %, recall of 81.69 %, F1-score of 81.07 %. F. Es-Sabery [10] proposed a MapReduce enhanced weighted ID3 decision tree classification strategy for the purpose of opinion mining and primarily comprised the 3 aspects. In the first step, they have utilized feature extractors such as TF-IDF, word embedding (Word2Vec and Glove), along with feature extractors such as FastText as well as N-grams to identify and capture the data which is relevant efficiently. After that, they have utilized various feature selectors such as Gini Index, Gain Ratio along with other feature selectors like Chi-square as well as Information in order to decrease the high feature's dimensionality. After that the work of classification using an enhanced ID3 decision tree classifier was by utilizing the collected features, the classifier seeks to evaluate the weighted information gain rather than the usual id3 information gain. To put it another way, there are two processes to measure the weighted information gain for the present conditioned feature. The first step is that they evaluate the weighted correlation function of the feature which is conditioned currently. In the second step, the task of multiplication of the weighted correlation function by the information gain of the current conditioned feature is performed. The Hadoop framework, along with its programming framework which is, MapReduce, and distributed file system which is HDFS, is used to implement this job in a distributed environment. The project's main purpose is to improve the accuracy as well as execution time, along with the ability to handle large datasets

of a well-known ID3 classifier. On the COVID-19_sentiment dataset, the proposed classifier outperformed other classifiers on basis of false-positive rate achieved is 13.49 percent, classification rate achieved is 88.82 percent, specificity achieved is 86.51 percent, F1-score achieved is 85.54 percent, kappa statistic achieved is 87.69 percent, the precision rate achieved is 86.67 percent, and error rate achieved is 11.18 percent, convergence rate achieved is 86.67 percent, false-negative rate achieved is 14.28 percent, Recall achieved is 85.72 percent, execution time achieved is 15.95s.T. Wang [21] consider the social platform, Sina Weibo, in which they deal only with the posts which have sentiments with only negative polarity. They utilized 999,978 coronavirus posts which were arbitrarily chosen from the Sina Weibo platform between the time duration of 01/01/2020 and 18/02/2020.

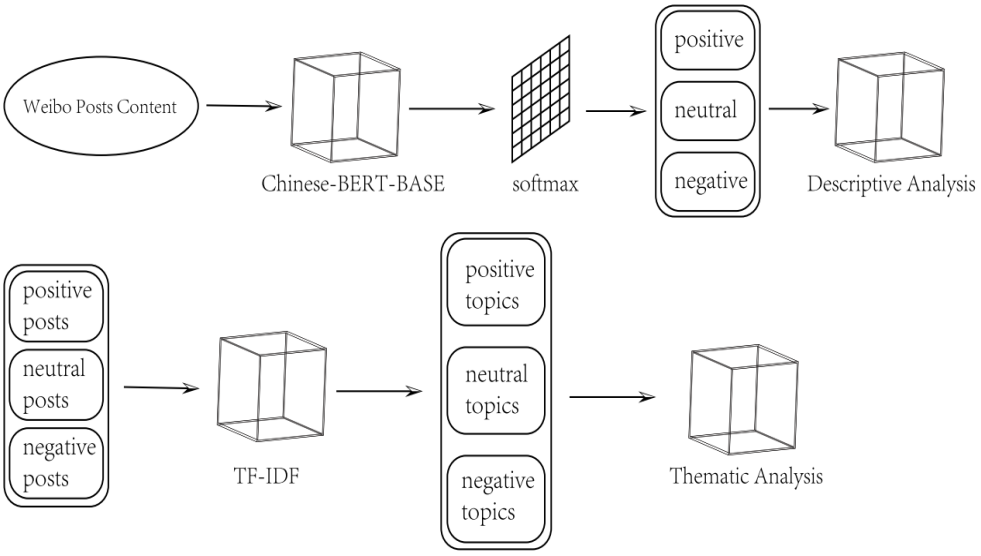


Fig. 7. Sentiment analysis model [21].

To perform the task of classification of sentiments into classes of negative, as well as positive and neutral they utilized the unsupervised BERT algorithm as shown in Fig 7. And for the summarization of the topics of the Sina Weibo posts, they utilized TF-IDF which is also known as the term frequency-inverse document frequency model. For the discovery of sentiments of negative polarity, they have used thematic analysis as well as trend analysis. They discovered that individuals are worried about the 4 sides of coronavirus. The first is the public health control with City Shutdown at 1.09 percent, Temperature Taking at 1.39 percent, and Coronavirus Cover-up at 1.26 percent. The Second is the virus's origin with; Conspiracy Theory at 1.43 percent, Gamey Food at 3.08 percent, and; Bat at 2.70 percent. The third is symptoms with Cough at 1.19 percent, Fever at 2.13 percent, and the fourth is production activity with School New Semester Beginning at 1.06 percent, Resume Work at 1.12 percent, and Go to Work at 1.94 percent. The methodology achieved an accuracy of 75.65 %.

M. K. Elhadad [22] presented a methodology for detecting false information that uses data from the United Nations, WHO, as well as UNICEF as shown in Fig 8. Along with this, they utilized the epidemiological data gathered from a variety of websites that performs the task of fact-

checking. The veracity of information should be ensured by gathering it from sources that are trusted.

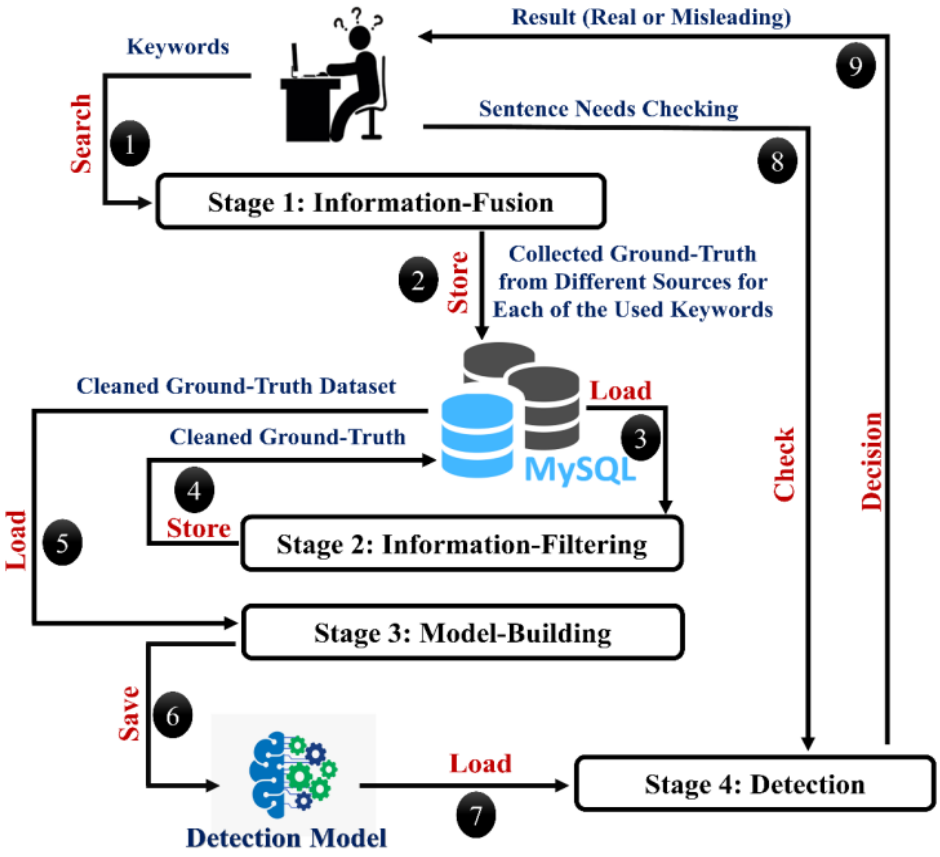


Fig. 8. Misleading-information detection framework [22]

They created a system for detection that employs machine learning techniques to spot misinformation by utilizing the gathered ground-truth data. They built a classifier that is based on voting ensemble machine learning by using 10 machine learning models which are Extreme Gradient Boosting classifiers, Neural Network, Bernoulli Naïve Bayes, Linear Support Vector Machines, k-Nearest Neighbor, Ensemble Random Forest along with algorithms such as Perceptron, Multinomial Naïve Bayes, Logistic Regression, Decision Tree and feature extraction techniques such as TF-IDF, word embeddings. To ensure the authenticity of the acquired data, we use a five-fold cross-validation technique and report on the assessment of 12 performance measures. They achieved the highest accuracy of 99.68 %.P. Gupta [23] proposed a methodology that aims to find out how the people of India feel about the government's nationwide lockdown, which was implemented to slow the spread of COVID-19. The sentiment analysis of tweets produced by the people of India was carried out utilizing Natural Language Processing as well as machine learning classifiers in this research. A total of 12 741 tweets with the keywords "Indialockdown" were retrieved from 05/04/2020 to 17/04/2020. They have used

the Tweepy API, which is, annotated with the TextBlob as well as Valence Aware Dictionary and sentiment Reasoner lexicons and which are preprocessed with the Python natural language toolkit to gather the data. The data were classified using 8 distinct classifiers. By using the LinearSVC classifier with unigrams, the experiment attained the greatest accuracy of 84.4 percent. They have used the various classifiers, Perceptron, RidgeClassifier, LinearSVC, BernoulliNB, PassiveAggressive Classifier, AdaBoostClassifier, Logistic Regression, MultinomialNB classifiers. According to the findings, the majority of citizens of India accepted the Indian government's choice to impose a lockdown during the Corona outbreak.

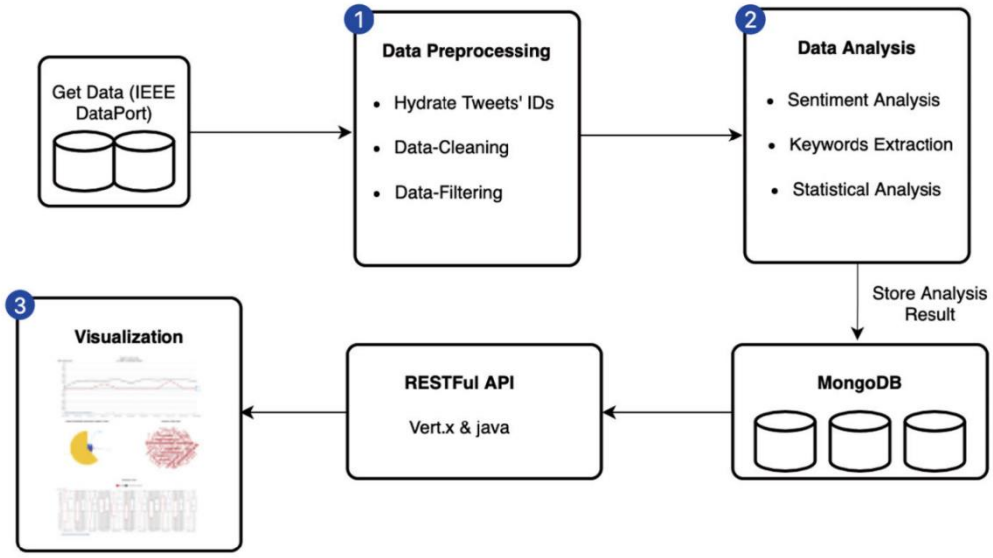


Fig. 9. Senti-COVID19 system architecture [24]

X. Yu [24] Proposed a methodology which is a visual analytic system which is also interactive named Senti-COVID-19 as shown in Fig 9. The system analyses the sentiment of people and also identifies the fluctuations triggers of sentiments on social platforms. The system utilizes sentiment analysis which is based on lexicons to reveal the perceptions of coronavirus events. They performed this by utilizing the libraries to draw out keywords as well as statistics for producing comprehensive data. The system also includes visuals for displaying the findings, helping users to find data that is relevant efficiently. The dataset utilized was Coronavirus (COVID-19) Geo-tagged Tweets Dataset, (IEEE DataPort) [25].

Table 1. Comparative analysis of existing techniques for sentiment analysis of COVID-19.

Research	**Model/Technique**	**Dataset used**	**Result**
Jelodar H et al. [15]	LDA, LSTM Recurrent Neural Network	Sub-reddits linked to Coronavirus	81.15% accuracy
R. F. Sear et al. [17]	LDA model	Facebook dataset	Anti-vax community: coherence score:0.46

			Pro vax community: coherence score: 0.498
M A. S. Imran et al. [11]	LSTM approach	Trending hashtag # data and the Kaggle dataset	82.4 % accuracy
J. Samuel et al. [12]	Naïve Bayes method	Twitter tweets	91% accuracy
Z. Long et al. [13]	SVM, Psychological need recognition models (NCR model, NTI model, NSM model, SCE model, LAI model)	New York state tweet hashtag dataset.	93.57 % accuracy by NSM model. Human need pronounced percentage relatedness: 47.42 percent, autonomy :16.63 percent competence :14.51 percent.
M. Bahja et al. [16]	LDA, VADER model, SNA approach	Twitter COVID-19tweets linking the pandemic to 5G	"Coronavirus" has the highest degree centrality of 0.1989 and betweenness centrality of 0.2098
R. L. Rosa et al. [9]	Tree-CNN, DBN (Deep Belief Network)	Twitter ,Sina Weibo dataset.Duration :15/11/2019 to 15/03/2020	Accuracy :87% for US election dataset, Accuracy: 86% for US Manhattan dataset, Accuracy: 93% for COVID-19 symptoms dataset.
A. Mourad et al. [19]	Mixed ontology data analytics	Twitter dataset	16.1% COVID-19 tweets redirect users to out of the context, advertisement 93.7% coronavirus tweets may be transmitting misleading or unverified medical information
F. Es-Sabery et al.[10]	Improved ID3 decision tree classifier, Hadoop framework Word2Vec and Glove word embedding	sentiment140 Dataset (Twitter dataset from Kaggle)COVID 19_Sentiments (Twitter dataset from Kaggle)	false-positive rate:13.49 percent, classification rate :88.82 percent, specificity :86.51 percent, F1-score :85.54 percent, kappa statistic: 87.69 percent, precision rate 86.67 percent, and error rate 11.18 percent,

			convergence rate 86.67 percent, false-negative rate 14.28 percent, Recall:85.72 percent, execution time :15.95s.
T. Da Yang et al. [20]	Linear probit model (LP), random forest model (RF), Linear regression model	Sina Weibo dataset	Precision :81.39 %, recall :81.69 %, F1: 81.07 %.
P. Ghasiya et al. [14]	RoBERTa, top2vec	COVID-19 news headlines, articles	90 % accuracy
X. Wan et al . [18]	TextRank and BART for text summarization, Louvain method, Leiden approach	US media Articles related to coronavirus	Highest ROUGE score of 50.45 on New York Times articles.
T. Wang et al. [21]	BERT, TF-IDF model	Sina Weibo dataset Duration:01/01/ 2020 to 18/02/2020	75.65 % accuracy
M. K. Elhadad et al. [22]	Voting ensemble machine learning on 10 algorithms: Extreme Gradient Boosting classifiers, Neural Network, Bernoulli Naïve Bayes, Linear Support Vector Machines, k-Nearest Neighbor, Ensemble Random Forest along with algorithm such as Perceptron, Multinomial Naïve Bayes, Logistic Regression, Decision Tree	Coronavirus data from WHO, UNICEF, and UN websites	Highest accuracy :99.68 %.
P. Gupta et al. [23]	Perceptron,RidgeClassifier ,LinearSVC,BernoulliNB,Passive Aggressive Classifier AdaBoostClassifier,Logistic Regression,Multinomi	Twitter dataset.	Accuracy: 84.4 percent

	alNB classifiers, VADER, TextBlob		
X. Yu et al. [24]	VADER lexicon-based sentiment analysis (Yake): unsupervised automatic keyword extraction algorithm	Coronavirus (COVID-19) Geo-tagged Tweets Dataset, (IEEE DataPort) [Z 25].	Weighted Average Sentiment score, Canada:0.09435, UK: 0.21439, US: 0.194132

5 Conclusion

As per the intention of the paper, there are many types of research that have been reviewed and certain flaws have been detected. The various research includes data from Reddit, Facebook , and other platforms like Twitter , Sina Weibo and news articles and headlines. The different methods for the analysis of the sentiments reviewed include various datasets like Twitter, Reddit, Facebook datasets as well as news headlines and articles. The topic techniques such as LDA, top2vec have been used extensively. Along with the sentiment analysis of the textual data, social network analysis has also been applied in some datasets. Deep learning approaches, such as LSTM and CNN, as well as BERT, have been utilized extensively by researchers. Along with this, lexicon-based methods, which take less effort in human-labeled documents, are quite effective in some circumstances. Amongst the various researches reviewed, in totality, the RoBERT classifier performed better than other approaches and BernoulliNB methods gave the least performance. We also looked into the impact of different features on the classifier. This methodical review sermonized the methodologies that were utilized to classify the sentiments as well techniques such as topic modeling and text summarization on the dataset that was utilized by the researchers. Along with this, the results that were obtained by the researchers are illustrated in this paper. These results are usually expressed in terms of Accuracy, Recall, Precision, and F1-score. Some studies also expressed the results in terms of degree centrality measures.

Most of the proposed methodologies were vulnerable to spam posts and bot activity. And the techniques were not able to detect the sarcasm efficiently. Future works can be performed in this direction to overcome these limitations. The data that are available currently are necessary for handling outbreaks of similar nature in the time ahead and tackling these adversities. We all stand together in the faith that advances in technology along with science will help avert diseases like COVID-19 from resurfacing.

References

[1]J. Akilandeswari and G. Jothi, “Sentiment classification of tweets with non-language features,” Procedia Comput. Sci., vol. 143, pp. 426–433, 2018

[2]M. Bouazizi and T. Ohtsuki, “Sentiment analysis in twitter: From classification to quantification of sentiments within tweets,” 2016 IEEE Glob. Commun. Conf. GLOBECOM 2016 - Proc., 2016.

[3]P. Barnaghi, P. Ghaffari, and J. G. Breslin, “Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment,” Proc. - 2016 IEEE 2nd Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2016, pp. 52–57, 2016.

[4]A. Prabhat and V. Khullar, “Sentiment classification on big data using Naïve bayes and logistic regression,” 2017 Int. Conf. Comput. Commun. Informatics, ICCCI 2017, 2017
[5]A. P. Rodrigues and N. N. Chiplunkar, “A new big data approach for topic classification and sentiment analysis of Twitter data,” Evol. Intell., no. 0123456789, 2019
[6]F. Paquin, J. Rivnay, A. Salleo, N. Stingelin, and C. Silva, “Multi-phase semicrystalline microstructures drive exciton dissociation in neat plastic semiconductors,” J. Mater. Chem. C, vol. 3, pp. 10715–10722, 2015.
[7]K. P. Bennett and C. Campbell, “Support Vector Machines: Hype or Hallelujah?,” SIGKDD Explor. Newsl., vol. 2, no. 2, pp. 1–13, Dec. 2000.
[8]M. Byrkjeland, F. Gørvell de Lichtenberg, and B. Gambäck, “Ternary Twitter Sentiment Classification with Distant Supervision and Sentiment-Specific Word Embeddings,” pp. 97–106, 2019.
[9]R. L. Rosa et al., "Event Detection System Based on User Behavior Changes in Online Social Networks: Case of the COVID-19 Pandemic," in IEEE Access, vol. 8, pp. 158806-158825, 2020, doi: 10.1109/ACCESS.2020.3020391.
[10]F. Es-Sabery et al., "A MapReduce Opinion Mining for COVID-19-Related Tweets Classification Using Enhanced ID3 Decision Tree Classifier," in IEEE Access, vol. 9, pp. 58706-58739, 2021, doi: 10.1109/ACCESS.2021.3073215.
[11] M A. S. Imran, S. M. Daudpota, Z. Kastrati and R. Batra, "Cross-Cultural Polarity and Emotion Detection Using Sentiment Analysis and Deep Learning on COVID-19 Related Tweets," in IEEE Access, vol. 8, pp. 181074-181090, 2020, doi: 10.1109/ACCESS.2020.3027350.
[12]J. Samuel et al., "Feeling Positive About Reopening? New Normal Scenarios From COVID-19 US Reopen Sentiment Analytics," in IEEE Access, vol. 8, pp. 142173-142190, 2020, doi: 10.1109/ACCESS.2020.3013933.
[13]Z. Long, R. Alharthi and A. E. Saddik, "NeedFull – a Tweet Analysis Platform to Study Human Needs During the COVID-19 Pandemic in New York State," in IEEE Access, vol. 8, pp. 136046-136055, 2020, doi: 10.1109/ACCESS.2020.3011123.
[14]P. Ghasiya and K. Okamura, "Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach," in IEEE Access, vol. 9, pp. 36645-36656, 2021, doi: 10.1109/ACCESS.2021.3062875.
[15] Jelodar H, Wang Y, Orji R, Huang S. Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach. IEEE J Biomed Health Inform. 2020 Oct;24(10):2733-2742. doi: 10.1109/JBHI.2020.3001216. Epub 2020 Jun 9. PMID: 32750931.
[16]M. Bahja and G. A. Safdar, "Unlink the Link Between COVID-19 and 5G Networks: An NLP and SNA Based Approach," in IEEE Access, vol. 8, pp. 209127-209137, 2020, doi: 10.1109/ACCESS.2020.3039168.
[17]R. F. Sear et al., "Quantifying COVID-19 Content in the Online Health Opinion War Using Machine Learning," in IEEE Access, vol. 8, pp. 91886-91893, 2020, doi: 10.1109/ACCESS.2020.2993967.
[18] X. Wan, M. C. Lucic, H. Ghazzai and Y. Massoud, "Topic Modeling and Progression of American Digital News Media During the Onset of the COVID-19 Pandemic," in IEEE Transactions on Technology and Society, doi: 10.1109/TTS.2021.3088800.
[19]A. Mourad, A. Srour, H. Harmanani, C. Jenainati and M. Arafeh, "Critical Impact of Social Networks Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-Based Study and Research Directions," in IEEE Transactions on Network and Service Management, vol. 17, no. 4, pp. 2145-2155, Dec. 2020, doi: 10.1109/TNSM.2020.3031034.
[20]T. Da and L. Yang, "Local COVID-19 Severity and Social Media Responses: Evidence From China," in IEEE Access, vol. 8, pp. 204684-204694, 2020, doi: 10.1109/ACCESS.2020.3037248.
[21]T. Wang, K. Lu, K. P. Chow and Q. Zhu, "COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model," in IEEE Access, vol. 8, pp. 138162-138169, 2020, doi: 10.1109/ACCESS.2020.3012595.
[22]M. K. Elhadad, K. F. Li and F. Gebali, "Detecting Misleading Information on COVID-19," in IEEE Access, vol. 8, pp. 165201-165215, 2020, doi: 10.1109/ACCESS.2020.3022867.

[23]P. Gupta, S. Kumar, R. R. Suman and V. Kumar, "Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter," in IEEE Transactions on Computational Social Systems, vol. 8, no. 4, pp. 992-1002, Aug. 2021, doi: 10.1109/TCSS.2020.3042446.
[24]X. Yu, M. D. Ferreira and F. V. Paulovich, "Senti-COVID19: An Interactive Visual Analytics System for Detecting Public Sentiment and Insights Regarding COVID-19 From Social Media," in IEEE Access, vol. 9, pp. 126684-126697, 2021, doi: 10.1109/ACCESS.2021.3111833.
[25]R. Lamsal. (2020). Coronavirus (COVID-19) Tweets Dataset. [Online]. Available: https://dx.doi.org/10.21227/781w-ef42.