Sentiment Analysis in Indian Sub-continent During COVID-19 Second Wave using Twitter Data

Meghana B[1*], Sanskriti Midha[2*], V Ramana Murthy Oruganti[3]
Department of Computer Science and Engineering[1,2] Department of Electrical and Electronics Engineering[3]
[1]cb.en.u4cse17521@cb.students.amrita.edu, [2]cb.en.u4cse17251@cb.students.amrita.edu, [3]ovr_murthy@cb.amrita.edu
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India

***Abstract*—In Indian sub-continent COVID-19 second wave started in early March 2021 and its effect was more lethal than the first wave, the confirmed cases and the death rate was higher than in the first wave. Unlike the national lockdown in 2020, this year different states have started imposing lockdown like restrictions spanning April-June 2021. This paper investigates the sentiments of the people using twitter messages during early period of the second wave. Two-weeks data is manually annotated and several machine learning models were built. The best performing models were used to predict sentiments for the next 2-3 weeks and analysis is presented. Predictions of public, commercial libraries were also analysed in the same context.**

***Index Terms*—COVID-19, lockdown, second-wave, sentiment analysis, n-grams, BERT, machine learning**

I. Introduction

Twitter is one of the popular microblogging platform like Facebook where its registered users can share freely their ideas, opinions or thoughts. These short messages (limited to 280 characters) are called as Tweets. Along with the textual content, a Tweet may also contain additional non-textual information including links, emoticons and hashtags to aid the user in enriching his/her opinion with sentiments. Very often, Government officials and political figures have used Twitter in many occasions to inform the general public about their activities or decisions.

Hence, researchers have mined twitter data for diverse issues including sentiment analysis [1], election trends, education, sports and so on. Sentiment analysis was used in the past to study the public opinion on 2015 Chennai floods in India [2], 2017 demonetization policy in India [3], 2018 issue over Jallikattu in India [4] and many more incidences. In this article, we intend to perform sentiment analysis of people during COVID-10 second wave amid the initial weeks of lockdown. So far we have not come across any work related COVID-19 second wave in the literature. In this context, the following contributions are made in this article

- Two datasets (Twitter IDs) of nearly 95K messages that will be made publicly available
- Analysis of Machine learning (ML) models trained by different COVID sentiments data
- Analysis of predictions by different ML models trained on indigenous data and public, commercial libraries.

The remainder of this article is organized as follows. Section II contains literature review on twitter based articles collected during COVID-19 (first wave). Section III contains the details of methodology used and contributions made in this work. Results and analysis followed by comparison with previous works are described in Section IV. Section V concludes this work and identifies future directions of research.

II. Related Works

Gupta *et al.* [5] used Twitter data from April 5, 2020 to April 17, 2020 having the keywords ”Indialockdown” to extract Tweets using Tweepy API. A dataset of 7284 was constructed with 3545 positive, 2097 neutral and 1642 negative polarity. Initially the Tweets were annotated using TextBlob and VADER libraries. Intersection of the results was used to consolidate the final sentiment polarities. Tweets were then labelled as positive, neutral, and negative. Pre-processing of Tweets was done using Python’s natural language tool kit. They reported a highest performance of 84.4% classification accuracy with unigrams features and Linear SVC classifier. They conclude that the majority of Indian citizens supported the decision by the Indian government on lockdown implemented during COVID first wave in 2020.

Cotfas *et al.* [6] used several machine learning and deep learning algorithms for stance analysis. Tweets were classified into three categories – in favor, against and neutral regarding COVID-19 vaccination. Tweets were collected over one-month period from the first announcement of a coronavirus vaccine after successful (limited) clinical trials. They used the Twitter Filtered Stream API (TweetInvi library) to extract Tweets. The highest performance of an accuracy 78.94% was obtained using BERT.

Safarnejad *et al.* [7] collected Tweets with keyword Zika and other related keywords during the entire year of 2016. Based on the number of reTweets received, they ranked the original Tweets from highest to lowest. Peer-reviewed journal and conference publications, government and health agencies including CDC and WHO reports and statistics were used to determine the Tweets information as real or misinformation. They identified 455 Tweets as ones containing real information and some other 264 Tweets, though very influential, as ones containing misinformation. using a novel data mining

[*] Equal contribution

technique (mis)information dissemination networks and signals were constructed. Dissemination features were extracted along with content-based features to build Zika misinformation classifier using random forest which yielded more than 85% accuracy and 90% AUC.

Raheja *et al.* [8] collected more than 370 Tweets from twitter using the three keywords (COVID, CORONA VIRUS, COVID – 19) during COVID-19 pandemic condition. 31%, 19% and 50% of the Tweets were identified as positive, nagative and neutral Tweets respectively. They performed the sentimental analysis to identify the opinion of people and conclude that the neutral sentiments were higher than positive and non-positive sentiments during COVID situation.

Tam *et al.* [9] proposed a novel CNN integrated with Bi-LSTM model for sentiment classification. Tweets were crawled from Chicago over a span of two months. With Word2Vec word vectorisation technique highest performance of 91.13% accuracy was reported.

Bibi *et al.* [10] investigated three hierarchical clustering techniques – single linkage (SL), complete linkage (CL) and average linkage (AL) for twitter sentiment analysis. A cooperative framework of the above three techniques was built to select the optimal cluster for Tweets. They reported that cooperative clustering based on majority voting provided best cluster quality amongst the three techniques.

Bechini *et al.* [11] investigated thoroughly COVID vaccination in Italy. Along with the public opinion (Tweets), user-related information was also used in their advanced analysis on temporal and spatial scales. They reported that the stance behaviour and number of Tweets posted is different for verified users accounts from that of unverified users. Regional-level stance was built based on the geospatial analysis of location of active users occasionally provided by the users.

III. Methodology

The overall framework is shown in Figure 1. Initially COVID-19 second wave related Tweets are collected from Twitter. These Tweets were manually annotated into three categories – negative, neutral, and positive. The input to our training framework is these Tweets. The dataset containing Tweet id and Tweet text are passed first through the pre-processing stage. In this stage the tweets are cleaned to remove any irrelevant information and make it more easier for the ML models to learn. The preprocessed Tweets are then randomly divided into the training and the testing set having 80% and 20% data respectively. Both the sets are then passed through feature extraction process – countvectorizer, TF-IDF and BBERT. The feature matrix of the training set is used to train different classifiers – Support Vector Machine (SVM), Linear Regression and Stochastic Gradient Descent . The trained ML models are evaluated against the testing set and other benchmark dataset available in the literature. More details of each stage and our experiments are described in the following sections.

A. Dataset Description

An Academic Research account in Twitter was used to collect the Tweets and create dataset for further experiments in this paper. The recent search endpoint (GET /2/Tweets/search/recent) to access filtered public Tweets posted over the last week was utilized and program was written in MATLAB 2021a to collect Tweets April 2, 2021 to May 6, 2021. Field query was one of the terms – “corona”, “lockdown”, “covid-19”, “second wave”. Due to the cap on Tweets retrieved, we executed the query code to collect a maximum of 100 Tweets every 3 hours in the initial dates (April 2, 2021 onwards) without any duplicates or reTweet option enabled. Later it was found that cap 100 was reached in most times. Hence we collected for every two hours and then later every hour towards. The number of Tweets for different days is shown in the Figure 2. We divided the collected data into two sets – SenTweetment11K and SenTweetment84K. SenTweetment11K contains 11792 samples collected from the dates April 2 to April 16, 2021. Annotations were done manually and used in training several ML models. SenTweetment84K contains around 84,000 samples collected from the dates April 17 to May 6, 2021. This dataset is used only for predictions. No annotations are available. Both the datasets (Tweet IDs only as per Twitter’s terms and conditions) will be made available for the research community along with this publication along with manual annotations (for SenTweetment11K).

B. Pre-Processing

Since twitter data is largely unstructured, preprocessing steps are extremely important. This will help us remove irrelevant information and make the data more structured for the ML models to learn better. The Tweets obtained using Twitter’s recent search endpoint were subject to following steps:

1) Hyperlinks, usernames, emails and tags were removed since they do not provide any extra information regrading the context
2) Punctuation and special characters such as commas, apostrophes, quotes , question marks , exclamation marks and full stops were removed to reduce noise and also since they do not add more information to the natural language.
3) The entire text was converted to lowercase to bring consistency and reduce the dimensions in the vector space model.
4) The emojis were removed and replaced with the corresponding textual descriptions using the Python’s emot library to understand the emotional stance better in the data.

C. Annotation

Only SenTweetment11K dataset (the Tweet data from dates April 2 -April 16, 2021) was annotated using different schemes as follows

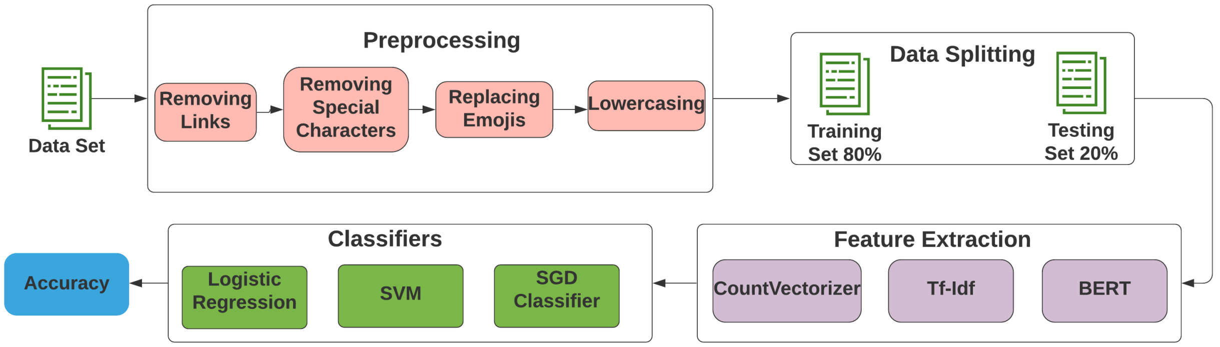

Fig. 1. Overall Framework

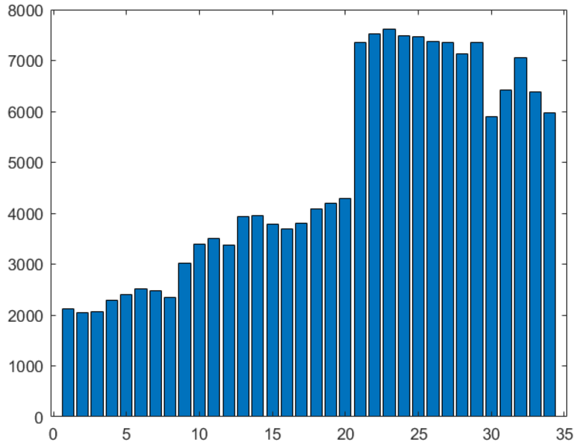

Fig. 2. Tweets collected on each day for the period April 2 – May 6, 2021

1) A regular Twitter user was requested to give label to the pre-processed Tweet. A total of four different Twitter users served as volunteer for our annotations task. For any Tweet, only one volunteer has given his/her label. Before labeling, all the four volunteers were trained using the guidelinesThis was to ensure uniformity in labeling the same sentiment. Through this technique the samples obtained in Negative, Neutral and Positive categories are 2996, 7773 and 1023 respectively. Tweets discarded as not belonging to any of the three sentiments have been discarded in the next two schemes also (below).
2) We use open source libraries to predict the annotations. Similar to Gupta *et al.* [5] we use TextBlob [1] and VADER [12] to estimate the polarity of the Tweet and then we take an intersection of the values, i.e we keep only those Tweets for which the polarity estimate is the same from both TextBlob and VADER. Through this technique the samples obtained in Negative, Neutral and Positive categories are 3171, 3480 and 5141 respectively.
3) We use commercial library AWS Comprehend [2] to estimate the polarity of the Tweets. Through this technique the samples obtained in Negative, Neutral and Positive categories are 2367, 8853 and 273 respectively.

The label distribution is shown in table I.

Reference	Number of Tweets (Neg., Neu., Pos.)	Availability
SenTweetment11K(ours)	11792 (2996,7773,1023)	*Public*
SenTweetment84K(ours) Not annotated	84,299 –	*Public*
Gupta *et al.* [5]	7284 (1642,2097,3545)	Not public
Cotfas *et al.* [6]	752,951 2751 (805,964,982)	Not public Public
Tam*et al.* [9]	797,324 70,042 (475266,0,322058)	Not public Public

TABLE I
LABEL DISTRIBUTION IN THE DATASET

D. *Feature Extraction and classifiers*

1) CountVectorizer [13] transforms the text into a numerical vector on the basis of the frequency of each word that occurs in the entire document.
2) We use TF-IDF [13] as this statistic reflects how important a word is in the document.TF-IDF associates each word in a document with a numerical value that describes its relevance in the document.

[1] https://textblob.readthedocs.io/en/dev/

[2] https://aws.amazon.com/comprehend/

3) BERT [14] is a bidirectional transformer model using an attention mechanism to contextual relations between words in a text. DistilBERT is a condensed version of BERT which uses a technique called distillation. It approximates the Google's BERT, retaining about 97% performance but using only half the number of parameters. The idea is that using a smaller neural network we can approximate the output distributions of a much larger neural network. We use the word embeddings generated by pre-trained DistilBERT directly for our classification tasks using classical ML models

We have used 3 different ML models namely Suppport Vector machine(SVM), Logistic Regression (LR), Stochastic Gradient descent(SGD)

1) **SVM:** SVM [13] is a supervised learning algorithm capable of performing non-linear classification by directly translating inputs into high-dimensional feature spaces.In the SVM algorithm, the value of each feature is the value of a particular coordinate and each data item is plotted as a point in n-dimensional space where n is number of features .Classification is performed by finding the hyper-plane that differentiates the classes very well.
2) **LR:**Logistic regression[13] is a predictive analysis algorithm which is used in classification problems.It is used to assign observations to a discrete set of classes based on the concept of probability.The hypothesis of logistic regression limits the cost function between 0 and 1.Sigmoid function is used to map predicted values to probabilities and it maps any real value into another value between 0 and 1.A threshold value is decided based on which the probabilities above a certain value are classified into different classes.
3) **SGD:** SGD [13] Classifier is a linear classifier which can be anything such as SVM, logistic regression or others. These classifiers are optimized by the SGD.Gradient descent is used to minimize a cost function.The default loss function is the 'squared loss' and it refers to the ordinary least squares fit.

IV. Results and Analysis

Several experiments are conducted to predict the sentiment (category) of people during COVID second wave. We begin with obtaining baseline performance on two datasets – [6] and our SenTweetment11K dataset

A. Baseline Performance using datasets

In this section we report results obtained by several combinations of features and classifiers, popular in NLP, on two datasets – SenTweetment11K and publicly available dataset [6]. The best performing feature and classifier performance will be treated as baseline performance for comparison with rest of the experiments.

The results obtained on our manually annotated dataset – SenTweetment11K and [6] have been summarized in Tables II and III respectively. The highest performing combination is highlighted in bold color.

Classifier	SVM	LR	SGD
Count Vectorizer(1gram)	68.41	70.79	68.03
TF-IDF	72.82	72.74	**72.86**
DistilBERT	71.67	72.08	71.30

TABLE II
Performance (accuracy) on SenTweetment11K dataset

Classifier	SVM	LR	SGD
Count Vectorizer (1gram)	69.56	70.76	69.87
TF-IDF	**72.05**	70.96	68.78
DistilBERT	65.24	70.32	62.36

TABLE III
Performance(accuracy) on benchmark dataset [6]

We now mix training data of [6] and SenTweetment11K and test on the testing data of publicly available [6]. The results obtained are shown in Table IV. The best performing features+classifier combination in Tables III and IV are compared because the testing data is same in both the cases i.e., testing data of [6]. There is decrease from 72.05% to 7155%. As this amount (0.5% absolute) is very small, we are unable to comment anything conclusively here.

Classifier	SVM	LR	SGD
Count Vectorizer(1gram)	66.84	71.37	69.20
TF-IDF	69.24	**71.55**	70.83
DistilBERT	68.65	67.65	69.20

TABLE IV
Performance(accuracy) of training Manually labelled dataset and testing on benchmark dataset [6]

Again we mix training data of [6] and SenTweetment11K and test on the testing data of SenTweetment11K. The results obtained are shown in Table V. The best performing features+classifier combination in Tables II and V are compared because the testing data is same in both the cases i.e., testing data of SenTweetment11K. There is an increase from 72.86% to 73.42%. As this amount is very small (0.56% absolute), here also we are unable to comment anything conclusively. However from both the comparisons performed we can infer that our manually collected dataset SenTweetment11K is homogeneous with the existing COVID sentiment analysis dataset. With size 11K compared to 3K [6] we conclude our manual dataset SenTweetment11K as significant contribution to the NLP research community.

Further, the best performing – TF-IDF+SVM combination is used in future experiments described in the following Sections.

B. Baseline Performance using libraries

Instead of manually annotations, we investigate open and commercial libraries for annotation on our SenTweetment11K

Classifier	SVM	LR	SGD
Count Vectorizer(1gram)	69.35	71.82	69.98
TF-IDF	**73.42**	73.37	73.08
DistilBERT	70.87	71.08	70.49

TABLE V
PERFORMANCE(ACCURACY) OF TRAINING MANUALLY LABELLED DATASET AND BENCHMARK DATASET [6] AND TESTING ON MANUAL DATASET

dataset. Highest performance in each table is highlighted using bold font. Results obtained by using Textblob and Vader agreed annotations are shown in VI.

Classifier	SVM	LR	SGD
Count Vectorizer (1gram)	83.81	**91.82**	88.81
TF-IDF	83.30	81.48	86.99
DistilBERT	72.57	76.13	74.35

TABLE VI
PERFORMANCE (ACCURACY) ON SENTWEETMENT11K DATASET WITH TEXTBLOB AND VADER AGREED ANNOTATIONS

Results obtained by using AWS sentiment annotations on our SenTweetment11K dataset are shown in VII

Classifier	SVM	LR	SGD
Count Vectorizer (1gram)	64.63	75.61	70.12
TF-IDF	67.07	66.46	75.00
DistilBERT	78.66	**81.71**	73.17

TABLE VII
PERFORMANCE (ACCURACY) ON SENTWEETMENT11K DATASET WITH LABELS GENERATED BY AWS SENTIMENT CLASSIFIER

Further, we use complete SenTweetment11K with Textblob and Vader agreed annotations to train the classifiers and investigate their performance on benchmark dataset [6]. The results obtained have been summarized in VIII

Classifier	SVM	LR	SGD
Count Vectorizer (1gram)	**68.42**	65.34	68.06
TF-IDF	64.25	62.25	65.70
DistilBERT	64.25	67.51	65.70

TABLE VIII
TRAINING ON A MERGED TBV AND BENCHMARK DATASET, TESTING ON BENCHMARK DATASET([6])

Also, we use complete SenTweetment11K with AWS sentiment annotations to train the classifiers and investigate their performance on benchmark dataset [6]. The results obtained have been summarized in IX. On comparison with highest performances of Table VIII i.e., 75.05% vs 68.42%, we infer that AWS annotations are more reliable than Textblob and Vader agreed annotations. Further support for this inference is reflected in the upcoming section.

Classifier	SVM	LR	SGD
Count Vectorizer (1gram)	71.32	72.96	73.50
TF-IDF	71.51	71.14	68.78
DistilBERT	71.51	**74.05**	71.32

TABLE IX
TRAINING ON A MERGED AWS AND BENCHMARK DATASET, TESTING ON BENCHMARK DATASET([6])

C. Annotations mismatch

Since manually labelling huge datasets is a difficult task and not always viable, we explore the various tools available for sentiment identification, both commercial and open source. For commercial we use AWS Comprehend's sentiment analysis. For open access we use TextBlob and Vader and take the intersection of their annotations. The label distribution for the manually labelled dataset, TextBlob Vader annotations; and AWS comprehend annotations are shown in Figure 3.

AWS Comprehend's sentiment analysis tool also assigns a label called ”MIXED” which is assigned when the emotion of the probability of the Tweet being positive, negative or neutral is similar, or the Tweet displays a mixed emotion.

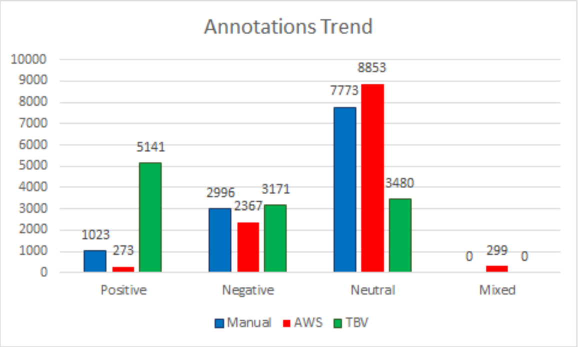

Fig. 3. Annotations on SenTweetment11K

As we can see from the figure 3, the annotations provided by AWS comprehend are more similar to the manual labels as compared to the open access TextBlob and Vader. In terms of accuracy, AWS gives a 67.9% match when compared to manual annotations while TextBlob and Vader only gives 39.2% match.

D. comparison with other works

In this section we intend to compare the public opinion of Indian people during COVID-19 first wave vs second wave. Gupta *et al.* [5] performed sentiment analysis of people towards lockdown during the first wave. As the data used by them is not available, we present an indirect analysis as preliminary attempt. Out of total 7284 samples, highest sentiment predicted was positive i.e., 48.67% of samples (people). In our manual annotation based SenTweetment11K dataset the neutral sentiment was found to be highest with 65.92%. This highest might be due to the people's preparedness of lockdown one year ago or the samples having many news articles, statements.

A ML model (TF-IDF+SVM) trained on different datasets was used to predict the sentiments on (not annotated) SenTweetment84K. As ground truth is not available for this dataset, the accuracy of the predictions cannot be validated. However, we just intend to see the highest predicted sentiment during COVID-19 second wave. From Figure 4, it is inferred that higher percentage of people were neutral followed by negative sentiments during the lockdown of COVID-19 second wave.

In another experiment, SenTweetment11K with annotations obtained from libraries – TextBlob, Vader and AWS was used to train ML model (TF-IDF+SVM) and predict the sentiments on (not annotated) SenTweetment84K. The results obtained are showin in Figure 5. It can be inferred from this figure that highest sentiment is Neutral, followed very closely by negative sentiment.

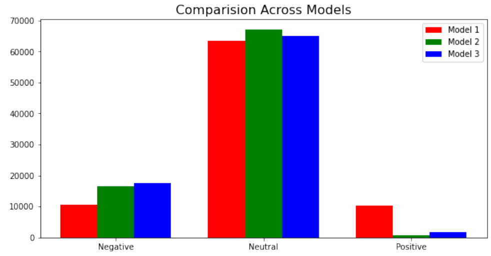

Fig. 4. Predictions on SenTweetment84K dataset. Model 1: Training on complete [6] dataset;Model 2: Training on complete SenTweetment11K; Model 3: Training on complete SenTweetment11K and complete [6] dataset

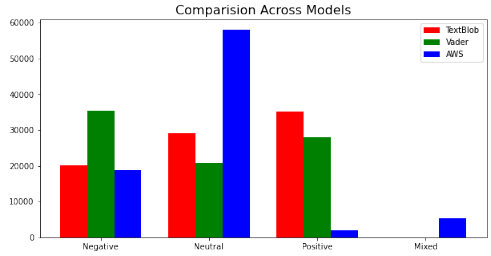

Fig. 5. Predictions on SenTweetment84K dataset. TextBlob: Training on complete SenTweetment11K with TextBlob annotations; Vader: Training on complete SenTweetment11K with Vader annotations; AWS: Training on complete SenTweetment11K with AWS annotations

V. Conclusions and Future Works

This is one of the foremost works performed on COVID-19 second wave in literature. Two datasets were collected from Twitter during early lockdown period of COVID-19 second wave in India. The baseline performances were established. Their alignment with benchmark datasets in literature was investigated and found to be homogeneous in sentiment analysis task. Usability of libraries – TextBlob, Vader and AWS was investigated and found that AWS yields annotations more closer to manual annotations; thus reliable than TextBlob or Vader libraries.

It is acknowledged that our manually contributed dataset is imbalanced with highest number of Neutral samples and lowest number of Positive sentiment samples. Future work can investigate techniques including oversampling and undersampling to address such imbalance nature. Further unsupervised techniques including clustering can be investigated to explore SenTweetment84K dataset for automatic labelling and then validate different ML models trained on SenTweetment11K and [6] dataset. or Semi-supervised techniques can be investigated on the mixed dataset containing 94K samples – SenTweetment11K and SenTweetment84K

References

[1] K. S. Naveenkumar, R. Vinayakumar, and K. P. Soman, “Amrita-cen-sentidb 1: Improved twitter dataset for sentimental analysis and application of deep learning,” in *2019 10th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2019*, 2019.

[2] M. R. Nair, G. R. Ramya, and P. B. Sivakumar, “Usage and analysis of twitter during 2015 chennai flood towards disaster management,” in *Procedia Computer Science*, vol. 115, 2017, pp. 350–358, cited By :20. [Online]. Available: www.scopus.com

[3] N. M. Dhanya and U. C. Harish, *Sentiment analysis of twitter data on demonetization using machine learning techniques*, ser. Lecture Notes in Computational Vision and Biomechanics, 2018, vol. 28, cited By :13.

[4] R. Archana Devi, K. Sooraj, S. V. Ghanapathy, and P. Dhileepan, “A comparative study on sentiment analysis for jallikattu protest,” *Journal of Advanced Research in Dynamical and Control Systems*, vol. 10, no. 5, pp. 1789–1794, 2018.

[5] P. Gupta, S. Kumar, R. R. Suman, and V. Kumar, “Sentiment analysis of lockdown in india during covid-19: A case study on twitter,” *IEEE Transactions on Computational Social Systems*, pp. 1–11, 2020.

[6] L. A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai, and F. Tajariol, “The longest month: Analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement,” *IEEE Access*, vol. 9, pp. 33 203–33 223, 2021.

[7] L. Safarnejad, Q. Xu, Y. Ge, and S. Chen, “A multiple feature category data mining and machine learning approach to characterize and detect health misinformation on social media,” *IEEE Internet Computing*, pp. 1–1, 2021.

[8] S. Raheja and A. Asthana, “Sentimental analysis of twitter comments on covid-19,” in *2021 11th International Conference on Cloud Computing, Data Science Engineering (Confluence)*, 2021, pp. 704–708.

[9] S. Tam, R. B. Said, and Tanriöver, “A convbilstm deep learning model-based approach for twitter sentiment classification,” *IEEE Access*, vol. 9, pp. 41 283–41 293, 2021.

[10] M. Bibi, W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem, and N. Habib, “A cooperative binary-clustering framework based on majority voting for twitter sentiment analysis,” *IEEE Access*, vol. 8, pp. 68 580–68 592, 2020.

[11] A. Bechini, P. Ducange, F. Marcelloni, and A. Renda, “Stance analysis of twitter users: the case of the vaccination topic in italy,” *IEEE Intelligent Systems*, pp. 1–1, 2020.

[12] C. Hutto and E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” 01 2015.

[13] 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, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” 2019.