How to Cite:
Vasukidevi, G., Anita, C. S., Rani, P. S., Kumar, V. M. N., & Jaithunbi, A. K. (2022). COVID tweet analysis using NLP. *International Journal of Health Sciences*, *6*(S3), 9457–9466. https://doi.org/10.53730/ijhs.v6nS3.8314

COVID tweet analysis using NLP

Vasukidevi G.
Assistant Professor, Department of Science & Humanities, R.M.K.College of Engineering and Technology, Kavaraipettai

C. S. Anita
Professor, Department of AIML, R.M.D. Engineering College, Kavaraipettai

P. Shobha Rani
Associate Professor, Department of CSE, R.M.D. Engineering College, Kavaraipettai

Vimal Kumar M. N.
Associate Professor, Department of Mechatronics Engineering, Sona College of Technology, Salem

A. K. Jaithunbi
Assistant Professor, Department of CSE, R.M.D. Engineering College, Kavaraipettai

***Abstract*---**The pandemic has taken the world by storm. Almost the entire world went into lockdown to save the people from the deadly COVID-19. With the progression of time, news and mindfulness about COVID-19 spread like the actual pandemic, with a blast of messages, updates, recordings, and posts. Widespread panic manifest as one more worry not withstanding the well-being risk that COVID-19 introduced. Typically, for the most part because of misinterpretations, an absence of data, or now and again by and large deception about COVID- 19 and its effects. General people however have been expressing their feelings about the safety and effectiveness of the vaccines on social media like Twitter. In this study, such tweets are being extracted from Twitter using a Twitter API authentication token. The raw tweets are stored and processed using NLP. The processed data is then classified using a CNN classification algorithm. The algorithm classifies the data into three classes, positive, negative, and neutral. These classes refer to the sentiment of the general people whose Tweets are extracted for analysis. From the analysis it is seen that Our review upholds the view that there is a need to foster a proactive and general well-being presence to battle the spread of negative opinion via web- based entertainment following a pandemic.

International Journal of Health Sciences ISSN 2550-6978 E-ISSN 2550-696X © 2022.
Manuscript submitted: 27 March 2022, Manuscript revised: 9 April 2022, Accepted for publication: 18 May 2022

***Keywords*---**COVID19, social media, tweets, Natural Language processing, recurrent neural network.

Introduction

Virtual entertainment (and the world in general) have been inundated with fresh insight about the COVID-19 pandemic. With the progression of time, news and mindfulness about COVID-19 spread like the actual pandemic, with a blast of messages, updates, recordings, and posts. Widespread panic manifest as onemore worry not withstanding the well- being risk that COVID-19 introduced. Typically, for the most part because of misinterpretations, an absence of data, or now and again by and large deceptionabout COVID-19 and its effects. It is inthis way ideal and essential to direct appraisal of the early data streams during the pandemic via web-based entertainment as well as a contextual analysis of developing popular assessments via online entertainment which is of general interest. This study expects to illuminate strategy that can be applied to web- based entertainment stages; for instance, figuring out what level of balance isimportant to diminish falsehood via online entertainment. This concentrate additionally dissects sees concerning COVID-19 by zeroing in on individualswho interface and offer web-based entertainment on Twitter. As a stage for our trials, we present another huge scope feeling informational index COVIDSENTI, which comprises of 41,000 COVID-19-related tweets gathered in the beginning phases of the pandemic, from February to March 2020. The tweets have been named into positive, negative, and neutral classes. We examined the gathered tweets for opinion order utilizing variousarrangements of classifiers. Pessimistic assessment assumed a significant part in molding public opinion, for example, we saw that individuals inclined toward lockdown before in the pandemic; in any case, true to form, feeling moved bymid March.

This study expects to distinguish the points and the local are opinion elements communicated on Twitter about COVID-19. The examination questions we address are: 1) how to automatically distinguish individuals' feelings communicated on Twitter because of COVID-19 and 2) what themes are for the most part talked about by the Twitter clients while communicating opinions about COVID- 19? To respond to these inquiries, we gathered a novel marked dataset on COVID-19-related tweets and present an exploratory investigation of the informational collection and subject disclosure and feeling discovery utilizing NLP. Streaming Twitter information were gathered utilizing the Twitter API from February to March 2020, and opinion distinguished from the gathered tweets was named into three classes: those containing positive, negative, and unbiased tweets. Different algorithmic models are utilized to prepare and approve the informational index to give the baselines to recognizing feeling connected with forthcoming COVID-19 medicines spread on Twitter. Through a last check examination, the best performing model is chosen to upgrade and advance. We reason that future work should better record for setting and the heterogeneity in opinion connected with COVID-19 medicines. The principle commitments of this studyare as per the following:

- Plan a Transformation-based Multi- Profundity Distil BERT model for feeling examination of tweets to distinguish opinions concerningCovid from tweets.
- Remove feeling related succinct data from tweets to learn highlights without human mediationconsequently.
- Present a wide correlation between existing ML and DL message characterization strategies and examine the given pattern results. The proposed model beat on genuine datasets contrasted with allrecently utilized techniques.

Related Work

Word embedding is the process of feature extraction from text for NLP tasks such as sentiment analysis. Word embedding can be obtained using methods where words or phrases from the vocabulary are mapped to vectors ofreal numbers. The process generally involves a mathematical embedding from a large corpus with many dimensions per word to a vector space with a lower dimension that is useful for Machine Learning or Deep Learning models for text classification tasks Basic word embedding methods such as *bag of words* and *term frequency inverse document frequency* do not have context awareness and semantic information in embedding. This is also a problem for skip-grams that use n-grams (such as bi-gram and tri-gram) to develop word embedding, and in addition allow adjacent sequences of words tokens to be “skipped”

Over the last decade, there has been phenomenal progress in the area of world embedding and language models. Mikolov et al. proposed *word2vec* embedding which uses a feed forward neural network model to learn word associations from a text dataset which can detect synonymous words or suggest additional words given a partialsentence. It uses Continuous Bag-Of- Words (CBOW) or continuous skip- gram model architectures to produce a distributed representation of words. Themethod is used to create a large vector which represent each unique word in the corpus where semantic information and relation between the words are preserved. It has been shown that for two sentences that do not have much words in common, their semantic similarity can be captured using word2vec The limitation of word2vec is that it does not well represent the context of a word. Pennington et al. for obtaining vector representations for words by mapping words into a meaningful space where the distance between words is related to semantic similarity. GloVe uses matrix factorization to constructs a large matrix of co-occurrence information to obtain representation that showcase linear substructures of the word vector space. The embedding feature vectors with top list words that match with certain distance measures. GloVe can be used to find relations between words such as synonyms, company-product relations. Due to the awareness in ethics in Machine Learning, there has been a major focus on ethical issues in NLP. Arecent study showed that GloVe can have gender biased information; hence, a gender neutral GloVe method has beenproposed. There are some studies that review the effectiveness of word embedding methods. Ghannay et al. provided an evaluation of word embedding methods such as GloVe skip-gram, and continuous space language models (CSLM) [. The authors reported that skip-gram and GloVe outperformedCSLM in all the language tasks. Wang et al. evaluated word embeddingmethods such as GloVe for applications of biomedical text analysis where it was found that word embedding trained from clinical notes and literature better captured word semantics.

Existing System

They proposed for the issue of Twitter sentiment on COVID-19-related Twitter posts. We benchmark sentiment analysis methods in the analysis of COVID-19-related sentiment. This gives rise to the need to create analytic methods that could be rapidly deployed to understand information flows and to interpret how mass sentiment among thepopulation develops in pandemic scenarios. They tweet has been labelled into positive, negative, and neutral sentiment classes. We analysed the collected tweets for sentiment classification using different sets of features and classifiers. Negative opinion played an important role in conditioning public sentiment, forinstance, we observed that people favoured lockdown earlier in the pandemic They are not using time series RNN with LSTM for text classification They are not using any model deployment process. Not provide detailed analysis. Covid tweets dataset from different sources would becombined to form a generalized dataset,and then different Deep Learning algorithms would be applied to extract patterns and to obtain results with maximum accuracy. In this section of the report will load in the data, check forcleanliness, and then trim and clean given dataset for analysis. Make sure that the document steps carefully and justify for cleaning decisions.

Methodology

The proposed approach is isolated into four stages: 1) pre- Processing, 2) Keyword pattern analysis, 3) Word embeddings for feature extraction, and 4) Classification techniques. The CovidSenti dataset is isolated into two pieces, preparing and testing. The primary target of this review is to assess the characterization execution of cutting edge classifiers on the COVIDSenti dataset and afterward endeavor to further develop execution by separating key highlights of tweets. Figure 1 demonstrates ourproposed approach with each of the method explained in the following figure.

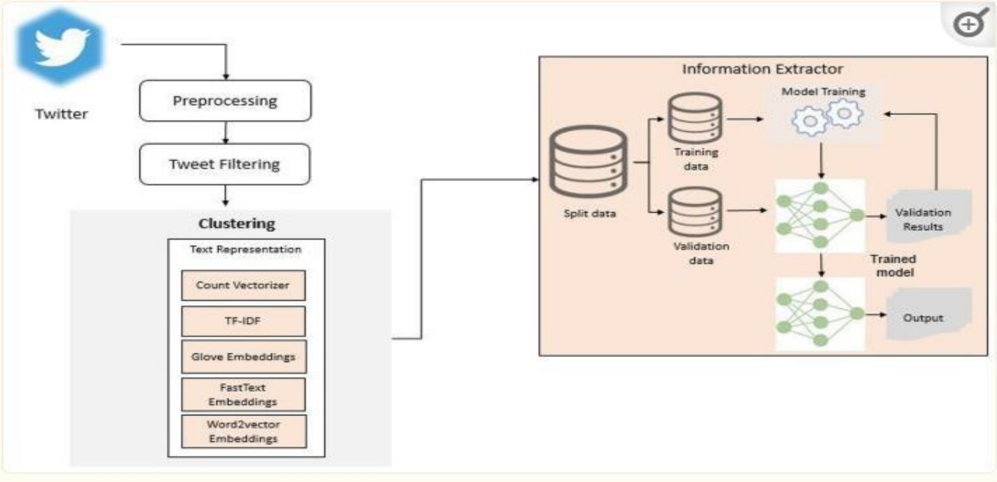


Figure 1. Overview of the Proposed Approach

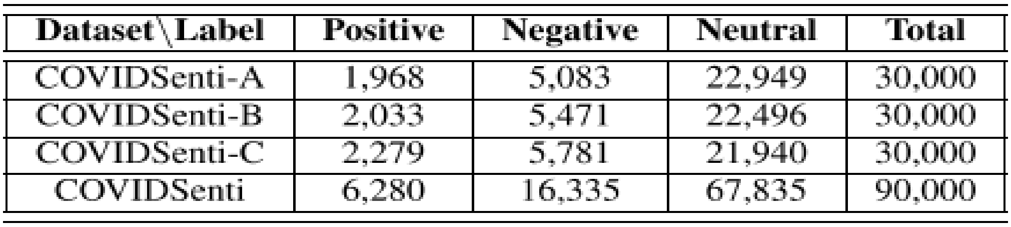
Data Collection and Labelling

The data set collected for predicting given data is split intoTraining set and Test set. Generally, 7:3 ratios are applied to split the Training set and Test set. The Data Model which was created using Deep Learning algorithms are applied on the Training set and based on the test Result accuracy, Test set prediction is done. The predicting the Covid tweets sentiment problem DL Neural network predication model is effective because of the following reasons. It provides better results in classification problem

Labelling

As tweets which are get from the twitterby using twitter API are many times short, unstructured, casual, and uproarious, the initial step of sentimental analysis is to preprocess the information. For pre-processing the tweets, we have to do ,the series of following procedures in order to improve the text. One renowned method for examining COVIDSenti data is to work out word frequencies to comprehend how frequently words are utilized in tweets. In this way, the initial step is lemmatization that cycles with the utilization of a jargon and morphological analysis of expressions and returns root words.

DATA SET DISTRIBUTION



Dataset\Label	Positive	Negative	Neutral	Total
COVIDSenti-A	1,968	5,083	22,949	30,000
COVIDSenti-B	2,033	5,471	22,496	30,000
COVIDSenti-C	2,279	5,781	21,940	30,000
COVIDSenti	6,280	16,335	67,835	90,000

- The subsequent step is to eliminate the stop words. It is the most reasonable method to beat the clamor from the tweets, (for example, "the," "a," "an," "in"). Stop words can be filtered from the message to be handled, and it really does never again influence comprehension of a tweet sentence's valence. We eliminated the renowned stop-words that are available in a text, for instance, "a," "an," "in," and "the."
- Contraction process is a course of shortening the by through supplanting or dropping letters with the guide of a punctuation. Many individuals speak with one another; individuals for the most part utilize abbreviated structures and shortenings of words in their text. We utilized the compression planning technique that drops the vowels from the words. Evacuation of compression mapping is connected with text normalization, and it is useful while working with tweets in sentiment analysis.
- The primary benefit of using textblob is its numerous capacities like phrase extraction, pos-tagging, and sentiment analysis.
- Twitter information is loud, which influences the exhibition of a classifier, so the pre-processor removes URLs, @user_mentions. We eliminate alphanumeric or extraordinary characters and eliminate non-ASCII characters

and numbers from our dataset due to they don't assist us with identifying emotions. We can replaceemojis with their corresponding reaction in text.

For hashtags, we take out the "#" imagefrom the beginning of the expression. We utilized tokenizer to part hashtags into proper words, for instance, "#stayhomestaysafe," tokenizerchanged over it into "remain," "home," "remain," "safe." Many words are connected all in all, and we likewise performed word-division to accomplishthis.

Exploratory Analysis

In this part, we direct exploratoryinvestigation to get a more extensiveperspective on our informational collection.

- Keyword Trend Analysis: We originally performed Keyword trend analysis on our preprocessed corpus to find out the most often referenced words. We observed that individuals arediscussing Covid cases, the Covid flare-up, social removing, the Covid pandemic, the emergencies due to Covid, and remaining at home.
- Topic Modeling: To quantitativelybreak down the points in our informational index, we analyse the point circulations with LDA. Top hashtags in the COVIDSenti informational index. LDA is a calculation for subject displaying, and that really intends that a text is made from a combination of subjects. After the LDA learning, themes depicted by the dissemination of words and the theme conveyance of the archives are learned. In LDA, we set the quantity of themes as six. Themes addressed by a distribution of words and the subject disseminations of the archives are learned after LDA preparing separately.

Feature Extraction

In this research, We used count vectorizer, TF-IDF, and word techniques for feature extraction. The most commonly and frequently used words in the given tweets is discovered by converting the tweets into vector space and technique called count vectorizer feature. The count vectorizermakes a word matrix where each distinct word denotes the of the network, and the selected text from the documentdenotes the column of the grid. Along with the count vectorizer we also used the term frequency-inverse document frequency(TF-IDF) feature extraction technique. The TF-IDF takes the TF andits corresponding IDF as an product to get the weight of the feature in a document . The length decides the TF ofelements in a single document. It is characterized in Equation (1).

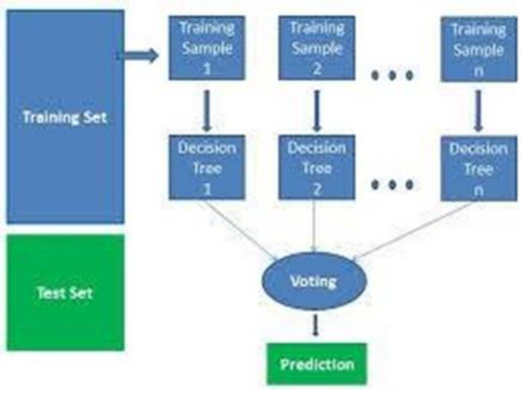
$$TF = \frac{count_{t,d}}{totalcount_{d}} \qquad (1)$$

Classification

To give a far reaching examination, we utilized ML-and DL- based classifiers to measure execution in the sentiment characterization task. ML- based classifiers, like support vectormachine (SVM), Naive Bayes (NB), Decision Tree (DT), and Random Forest (RF), are utilized in our investigation. Furthermore to customary ML, we have likewise applied two DL- basedclassifiers, specifically convolutional neural network(CNN) and bidirectional long-short memory (Bi- LSTM).Our CNN comprised of three convolutional layers, and to reshape the input size, a maxipool of filter sizer threehas been applied followed by a flatterninglayer after each layer and drop out layer with a pace of 0.5.Atlast ,a Thick layer followed by an yield layer, softmax utilizer as the enactment work. In ourtests, four Bi-LSTM cells with various quantities of secret hubs are utilized. After the primary Bi-LSTM a dropout of 0.5 is utilized for regularization, and a dropout of 0.25 is utilized after three Bi- LSTM layers. The Bi-LSTM layers. The Bi-LSTM cell's result is associated with dense layers with Relu as an activation work and abide by a softmax activatioin function.

Random Forest

Random forests or randomdecision forests are an ensemble learningmethod for classification, regression and other tasks, that operate by constructing amultitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set. Random forest is a type of supervised Machine Learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multipletimes to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the sametype i.e. multiple decision *trees,* resultingin *a forest of trees*, hence the name "Random Forest". The random forest algorithm can be used for both regressionand classification tasks. The following are the basic steps involved in performing the random forestalgorithm:



Diagrammatic representation of Random Forest

- Pick N random records from the dataset.
- Build a decision tree based on these Nrecords.
- Choose the number of trees you wantin your algorithm and repeat steps 1 and 2

- In case of a regression problem, for a new record, each tree in the forest predicts a value for output. The final value can be calculated by taking the average of all the values predicted by all the trees in the forest predicts the category to which the new record is assigned to the category that wins the majority vote.

Logistic Regression

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (inwhich there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.). In other words, the logistic regression model predicts P(Y=1) as a function of X. Logistic regression Assumptions:

- Binary logistic regression requiresthe dependent variable to be binary.
- For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
- Only the meaningful variablesshould be included.
- The independent variables should be independent of each other. That is, the model should have little.
- The independent variables are linearly related to the log odds.

Deep Learning RNN with LSTM get best accuracy result

Recurrent Neural Network is a generalization of feed forward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input. Unlike feed forward neural networks, RNNs canuse their internal state (memory) toprocess sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other. First, it takes the X(0) from the sequence of input and then it outputs h(0) which together with X(1) is the input for the next step. Similarly, h(1) from the next is the input with X(2) for the next step and so on

Conclusion

We presented a study with novel Deep Learning models for sentimental analysis during the rise of COVID infections. The number of tweets significantly lowered towards the peak of new cases. Furthermore, the optimistic, annoyed and joking tweets mostly dominated the monthly tweets with much lower number of negative with much lower number of negative sentiments expressed. We found that most tweets that have been associated with "joking" were either "optimistic" or "annoyed" and minority of them were also "thankful" in terms of the annoyed" sentiments in tweets, mostly were either "surprised" or "joking". These predictions generally indicate that although the majority have been optimistic, a significant group of population has been annoyed towards the way the pandemic was sentiment analysis which can provide more details for emerging topics during the rise of COVID-19 from specific regions. We took advantage of COVID-19 dataset of 41,000 hand-labelled tweets for training the respective Deep Learning models handled by the authorities. The major contribution of the paper is the framework which provides sentiment analysis in a population given the rise of the COVID-19 cases. Future work can use the framework for different regions, countries, ethnic and social groups to understand their behavior given multiple peaks of novel cases. The framework can be extended to understand reactions towards vaccinations with the rise of anti-vaccine sentiments given fear, insecurity and unpredictability of COVID-19.

References

1. Naseem U, Razzak I, Khushi M, Eklund PW, Kim J. Covidsenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis. *IEEE Trans Comput Soc Syst.* (2021) 8:1003–15. 10.1109/TCSS.2021.3051189
2. C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2013.
3. N. Ahmad and J. Siddique, "Personality assessment using Twitter tweets," *Procedia Comput. Sci.*, vol. 112, pp. 1964–1973, Sep. 2017.
4. T. Ahmad, A. Ramsay, and H. Ahmed, "Detecting emotions in English and Arabic tweets," *Information*, vol. 10, no.3, p. 98, Mar. 2019.
5. A.Bandi and A. Fellah, "Socio analyzer:A sentiment analysis using social media data," in *Proc. 28th Int. Conf. Softw. Eng. Data Eng.*, in EPiC Series in Computing, vol. 64, F. Harris,
6. S. Dascalu, S. Sharma, and R. Wu, Eds.Amsterdam,The Netherlands: EasyChair, 2019, pp. 61–67.
7. F. Barbieri and H. Saggion, "Automatic detection of irony and humour in Twitter,"in *Proc. ICCC*, 2014, pp. 155–162.
8. R. Bhat, V. K. Singh, N. Naik, C. R. Kamath, P. Mulimani, and N. Kulkarni, "COVID 2019 outbreak:The disappointment in Indian teachers," *Asian J. Psychiatry*, vol. 50, Apr. 2020, Art. no. 102047.
9. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
10. P. Boldog, T. Tekeli, Z. Vizi, A. Dénes,
11. F. A. Bartha, and G. Röst, "Risk assessment of novel coronavirus COVID-19 outbreaks outside China," *J. Clin. Med.*, vol. 9, no. 2, p. 571, Feb. 2020.
12. G. Carducci, G. Rizzo, D. Monti, E. Palumbo, and M. Morisio, "TwitPersonality: Computing personality traits from tweets using word embeddings and supervised learning," *Information*, vol. 9, no. 5, p. 127, May 2018.
13. J. P. Carvalho, H. Rosa, G. Brogueira, and F. Batista, "MISNIS: An intelligent platf orm for Twitter topic mining," *Expert Syst. App* ., vol. 89, pp. 374–388, Dec. 2017.

14. B. K. Chae, “Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data forsupply chain practice and research,” *Int.J. Prod. Econ.*, vol. 165, pp. 247–259, Jul. 2015.
15. M. De Choudhury, S. Counts, and E. Horvitz, “Predicting postpartum changes in emotion and behavior viasocial media,” in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, Apr. 2013,pp. 3267–3276.
16. A. Depoux, S. Martin, E. Karafifillakis,
17. R. Preet, A. Wilder-Smith, and H. Larson, “The pandemic of social media panic travels faster than the COVID-19 outbreak,” *J. Travel Med.*, vol. 27, no. 3,Apr. 2020, Art. no. taaa031.
18. J. Devlin, M.-W. Chang, K. Lee, and K.Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, vol.
19. Minneapolis, MN, USA: Associationfor Computational Linguistics, Jun. 2019, pp. 4171–4186.