IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Sentiment Analysis Of covid – 19 Twitter Dataset Using ML

Apoorva Bhimshetty[#1], Dr Sridevi Hosmani [*2]

1.Apoorva Bhimshetty Dept Of Computer Science And Engineering Sharnbasva University Engineering and technology (Exclusively For Women),Kalburgi, India

2. Dr Sridevi Hosmani Dept Of Computer Science And Engineering Sharnbasva University Engineering and technology (Exclusively For Women),Kalburgi, India.

Abstract- The global community has been inundated with reports of the COVID-19 epidemic thanks to social media. Over time, there was a deluge of COVID-19-related messages, upgrades, video, and postings, much like the actual epidemic. In besides the health danger that COVID-19 posed, widespread panic ensued. Unsurprisingly, widespread fear spread owing to misunderstandings, a lack of knowledge, and even deliberate disinformation concerning the nature and effects of COVID-19. Ex post facto analysis of the first information flows on social media during the epidemic, in addition to a test case of growth of public sentiment on social networks, is therefore topical and essential. This research hopes to inform legislation that may be implemented on social media platforms, such as figuring out how much moderation is required to reduce disinformation. This investigation also examines the perspectives of Twitter users in regards to COVID-19. We provide a big new sentiment data set called COVIDSENTI, which comprises of 90,000 tweets gathered in the early phases of epidemic, during February to March of 2020, and provides a foundation for our research. A favorable, negative, and neutral category has been assigned to each tweet. Using many feature and classifier sets, we investigated the twitter data for sentiment classification. Public opinion was heavily conditioned by negative commentary; for example, we found that individuals approved lockdown measures at the outset of the epidemic but, as predicted, public opinion switched by mid-March. The results of our research lend credence to the idea that public health agencies, in the wake of a pandemic, need to have a more proactive and nimble online presence to counter spreading of defeatist effect.

***Keywords*—COVID-19, Negative Sentiment, sentiment,**

I. INRODUCTION

CORONAVIRUS illness (also known as COVID-19) is a newly identified virus that got its name from the year it first emerged. Many nations have been hit by the pandemic, and every nation is fighting to stop the sickness from spreading, even those with relatively few cases. On January 30, 2020, it was officially labeled a pandemic by WHO, which has been working tirelessly to contain the outbreak ever since. The research and development of vaccinations is highly anticipated and promising. Currently, there is a dearth of scholarly research on the issue to benefit researchers, with the exception of Bhat et al. & Boldog et al. Because of this, it has been difficult to investigate the effects of COVID-19 on psychological health or the economic impact it might have over the world. Twitter, Facebook, Reddit, & Instagram, among others, have been hard at work analyzing and fact-checking to counteract the spread of disinformation since crazy conspiracy theories began circulating around COVID-19. To spread inaccurate information with the intent to mislead the public is an example of misinformation. This highlights the necessity for the development of analytic approaches that might be swiftly implemented to comprehend data flows or to analyze how public mood evolves in pandemic situations. Most studies offer the analysis of preventative treatment and recovery, health, social media platform, and economic data, but there hasn't been much study on assessing conspiratorial communication patterns on social media including accumulated personal-level information. One common technique for capturing human emotion is to analyze text posted on social media sites like Twitter and Facebook.

II. RELATED WORK

In the words of C. C. Aggarwal & C. K. Reddy [1]. This paper delves into the process of developing and reporting on experiments to study the usefulness of trying to apply a combined systemic estate of a text's sentences & phrase growth utilizing WordNet as well as a local thesaurus in choosing the most suitable abstractive text succinct summation for a given document. We used LCS technique to find the most comparable group of texts after we categorized and standardised them. The similarity between the sentences in the text was determined using LCS. A normalized value is produced and utilized to rank sentences. A chosen top set among the most comparable phrases then is segmented to obtain a collection of essential keywords or concepts. The created words were subsequently enlarged into two

subgroups utilizing 1) WorldNet; & 2) a local tech dictionary/thesaurus. The three sets acquired also were re-cycled further to improve & expand the list of chosen phrases from the original text. The procedure was done number of times in determining most effective representative group of phrases. We narrowed down the list of possible phrases to summarize to the very best ones. In order to test the usefulness of the technique, a number of studies was done utilizing an email corpus. The findings were compared to those that were generated by authors as well as to those obtained utilizing some simple sentences similarity calculating approach. Positive findings were obtained, with high correlations to both human experts & Overall sentence similarity. N. Ahmad & J. Siddique, [2]Social media fosters continual communication, between its users and the world via sharing personal data and their perspectives in every part of life. The primary purpose of this research is to investigate how twitter (dataset) might be utilized to enhance the user experience using character evaluation. In its main context, the paper demonstrates how the dataset may be used to create psychological profiles of people. The results of such profiling may be put to good use in a number of ways, including improving work conditions and morale and enabling customization of user interfaces. Utilizing DISC analysis, we offer a method by which a user's personality may be predicted via information mapping accessible on their public Twitter profile. The conclusions of this research may be beneficial for data acquisition, content selecting technique & marketing goods & services.

A.Proposed System

Proposed feelings using the COVID-19 database by taking into account all possible classifications. We used ML models to compare the performance of several methods for categorizing users' attitudes about COVID-19 and found that random forest yields the best results. We extended our investigation of DL models to identify users' emotions regarding COVID-19, computed their prediction performance, compared their findings to those of ML models, and shown as Most times DL prototypes produce superior outcomes than ML.

III. METHOLOGY

Decision Tree (DT): In a DT, rectangles represent nodes closer to center of tree, while the ovals represent nodes farther out. When it comes to supervised learning, the Decision Tree approach is where it's at.

This algorithm will iteratively carry out the below operations.

1. Establish first node of tree 1.
2. You should return the 'positive' leaf node if all examples are positive.
3. If all the given instances are negative, however, 'negative' leaf node should be returned.
4. Find the entropy of state H as it exists right now (S)
5. Determine entropy of each attribute with regard to the attribute 'x,' where 'x' is an arbitrary identifier (S, x)
6. 6. Pick the most valuable quality in terms of IG (S, x)
7. Discard greatest IG-providing characteristic from consideration.
8. Continue until either all characteristics have been used up or decision tree has no more leaves to explore.

Random Forest (RF) is a well-known supervised learning technique that utilizes a supervised learning method to enhance the performance of a model by combining many types of classification features. It is an ensemble classifier consisting of numerous decision trees, each of which is generated from a portion of the training data and parameters.

These stages outline the procedure that must be followed to get the job done:

1. Pick K training data points at random.
2. Construct decision trees connected to the points in question (Subsets).
3. decide on N as the total number of decision trees you'll be constructing.
4. Repeat Stages 1 and 2

Extreme Gradient Boost (XGBoost): Recently, XGBoost has become the go-to method in the field of applied ML. XGBoost is a fast and efficient implementation of a gradient-boosted decision tree. To achieve optimal performance, XGBoost models need more data and model modification than methods like a random forest.

Specifications of XGBoost The library has been designed with computational efficiency and model performance in mind, therefore it doesn't have many

extra features. Three primary types of gradient boosting are handled by this model.**:**

1) Gradient Boost
2) Stochastic Gradient Boost
3) Regularized Gradient Boost
4) Characteristics of System
5) This library contains the following for usage in various computer settings:
6) Building trees in parallel
7) Computational Distributed Learning for Massive Model Training
8) Algorithm & data structure cache optimization

.

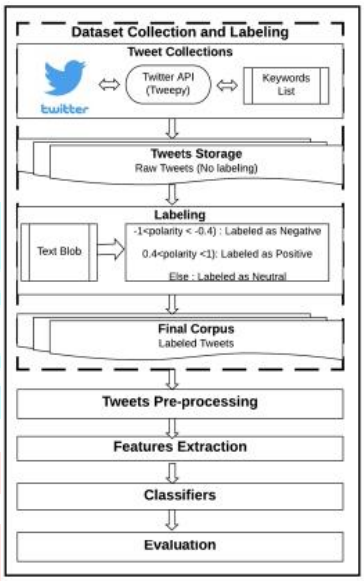

Fig2: System Architecture

V IMPLEMENTATION

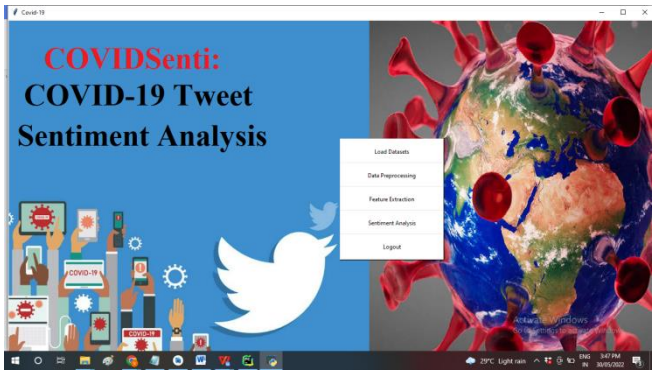

Fig3:Main
Fig page portrays every module as
1)loading datas
2) data preprocessing
3) extracting feature
4) sentiment study
5) logout

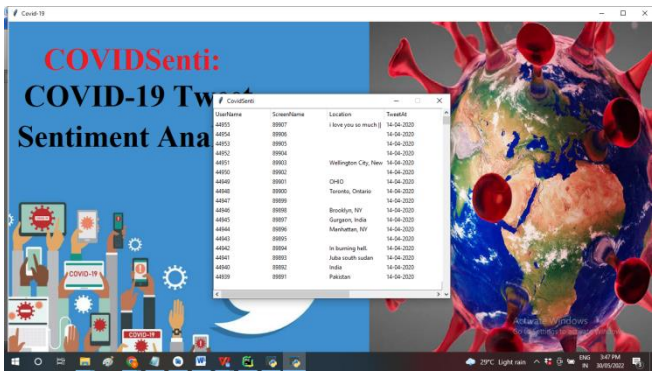

Fig5.7.2: Read Data
We crawled 2.1 million tweets during February and March 2020 and included 90,000 unique tweets from 70,000 individuals that were eligible for inclusion in fig 5.7.2. Based on the results of our research, we have discovered 12 distinct categories in which positive, negative, very negative, & neutral mood might be expressed, like lockdown, lockdown, & staying inside.

```
RangeIndex(start=0, stop=41157, step=1)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 41157 entries, 0 to 41156
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   UserName       41157 non-null  int64
 1   ScreenName     41157 non-null  int64
 2   Location       32567 non-null  object
 3   TweetAt        41157 non-null  object
 4   OriginalTweet  41157 non-null  object
 5   Sentiment      41157 non-null  object
dtypes: int64(2), object(4)
memory usage: 1.9+ MB
```

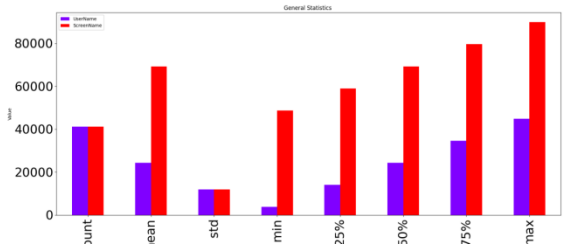

Fig5.7.3 Preprocessing
Pre-processing information is shown in Fig. 5.7.3. We can't get any useful information from columns labeled "UserName" and "ScreenName" in our data. So, we aren't using these characteristics in our model development. All of the information we have gathered from tweets sent in February and March of 2020. The next bar chart displays the total number of distinct values within every column.
A crucial first step in text mining is the preparation of raw text data. To better determine the tone of tweets, this process aims to eliminate irrelevant data, like punctuation (.,?,” etc.), **special characters**(@,%,&,$, etc.), **numbers**(1,2,3, etc.), **tweeter handle, links**

```
==========================================
Feature Extraction
==========================================
      UserName  ScreenName  ...          Sentiment complexityEvaluati
0         3799       48751  ...            Neutral           Sentime
1         3800       48752  ...           Positive           Sentime
2         3801       48753  ...           Positive           Sentime
3         3802       48754  ...           Positive           Sentime
4         3803       48755  ...  Extremely Negative           Sentime
...        ...         ...  ...                ...
41152     44951       89903  ...            Neutral           Sentime
41153     44952       89904  ...  Extremely Negative           Sentime
41154     44953       89905  ...           Positive           Sentime
41155     44954       89906  ...            Neutral           Sentime
41156     44955       89907  ...           Negative           Sentime

[41157 rows x 7 columns]
```

Fig5.7.4 Extracting Feature

Figure 4 displays a feature extraction, which involves the categorization of comments into positive, very positive, negative, and neutral categories.

```
   UserName  ...                                       CleanedTweet
0      3799  ...      Gahan   https   t co iFz9FAn2Pa and https ...
1      3800  ...  advice Talk to your neighbours family to excha...
2      3801  ...  Coronavirus Australia  Woolworths to give elde...
3      3802  ...  My food stock is not the only one which is emp...
4      3803  ...  Me  ready to go at supermarket during the   ou...

[5 rows x 7 columns]
   Actual  Predicted
0       0          4
1       4          4
2       1          4
3       2          2
4       3          2
AUC of the predictions: 0.6231713165790018
```

Fig 5.7.5(a) Sentiment analysis

For instance, the # /?? @ symbols are all removed and the clean tweet is shown in fig 5.7.5(a) below after sentiment analysis has been applied. We have used vectorization to standardize the test data and save it in the variables x test and y test. We also made forecasts about future values. Using actual and forecasted data, we calculated an area under the curve (AUC) of 0.6231713165790018.

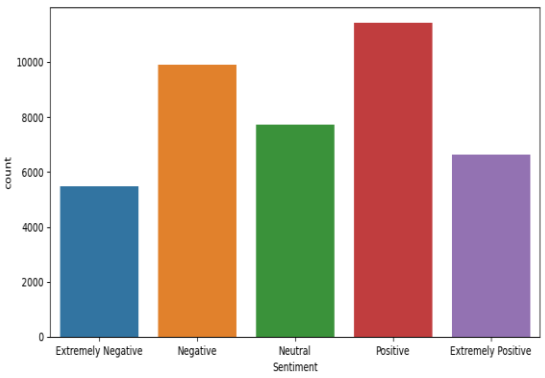

Fig5.7.5(b) Bar Graph

Fig 5.7.5(b) is a bar chart displaying the comments' overall tone, from very negative to positively neutral to positively favorable.

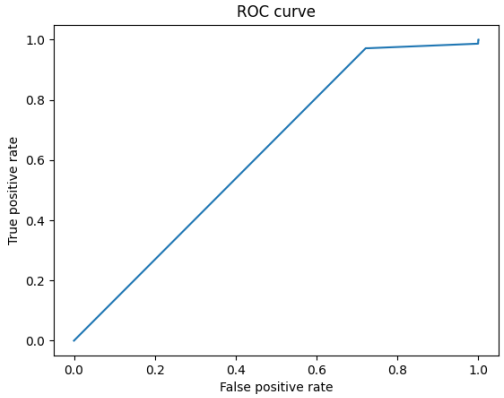

Fig5.7.5(c) ROC Curve

Given that the Naive Bayes classifier we used here received an AUC of -0.64, we can conclude that it is not very effective but is serviceable. As the area under the curve (AUC) gets closer to 1, the greater the ability to classify.

Any kind of emotional analysis of tweets is possible in the same manner.

VI. Conclusion

The purpose of developing Twitter sentiment analysis was to better understand how consumers feel about factors crucial to a company's commercial success. The software incorporates natural language processing methods into its analysis of sentiment in addition to machine learning, which provides more precise results.

References

1.C. C. Aggarwal and C. K. Reddy, Data Clustering: Algorithms and Applications, Boca Raton, FL, USA:CRC Press, 2013.

2.N. Ahmad and J. Siddique, "Personality assessment using Twitter tweets", Procedia Comput. Sci., vol. 112, pp. 1964-1973, Sep. 2017.

3.T. Ahmad, A. Ramsay and H. Ahmed, "Detecting emotions in English and Arabic tweets", Information, vol. 10, no. 3, pp. 98, Mar. 2019.

4.A. Bandi and A. Fellah, "Socio-analyzer: A sentiment analysis using social media data", Proc. 28th Int. Conf. Softw. Eng. Data Eng., vol. 64, pp. 61-67, 2019.

5.F. Barbieri and H. Saggion, "Automatic detection of irony and humour in Twitter", Proc. ICCC, pp. 155-162, 2014.

6.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.

7.D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet allocation", J. Mach. Learn. Res., vol. 3, pp. 993-1022, Jan. 2003.

8.P. Boldog, T. Tekeli, Z. Vizi, A. Dénes, F. A. Bartha and G. Röst, "Risk assessment of novel

coronavirus COVID-19 outbreaks outside China", J. Clin. Med., vol. 9, no. 2, pp. 571, Feb. 2020.

9.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, pp. 127, May 2018.

10.X. Carreras and L. Màrquez, "Boosting trees for anti-spam email filtering", arXiv:cs/0109015, 2001, [online] Available: https://arxiv.org/abs/cs/0109015.

11.J. P. Carvalho, H. Rosa, G. Brogueira and F. Batista, "MISNIS: An intelligent platform for Twitter topic mining", Expert Syst. Appl., vol. 89, pp. 374-388, Dec. 2017.

12.B. K. Chae, "Insights from hashtag #supplychain and Twitter analytics: Considering Twitter and Twitter data for supply chain practice and research", Int. J. Prod. Econ., vol. 165, pp. 247-259, Jul. 2015.

13.M. De Choudhury, S. Counts and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media", Proc. SIGCHI Conf. Hum. Factors Comput. Syst., pp. 3267-3276, Apr. 2013.

14.A. Depoux, S. Martin, E. Karafillakis, 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.

15.J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding", Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics Human Lang. Technol., vol. 1, pp. 4171-4186, Jun. 2019