IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

COVID PREDICTION USING REAL-TIME ANALYSIS

[1] PONNU JALARTH, [2] JASMINE JOSE

[1] Student, [2]Assistant Professor

[1,2]Department of Computer Science

[1,2]St. Joseph's College (Autonomous), Irinjalakuda, Thrissur, Kerala

Abstract:

Social networking sites have gained much popularity in last decades. People are more connected to such platforms like twitter and give updates about their opinions and also their current situations. A major amount of data generated now a days is through such sources. By classifying these data through sentiment analysis ,we can acquire many fruitful informations. Sentiment analysis is a method in text analysis that is very suitable for such social media evaluation. So, by proposing a real time system we can find enough information regarding Covid situation for predicting the future of this pandemic. Also we use machine learning models here for real time analysis.

Keywords: Twitter, Sentiment analysis, Machine learning

I. Introduction

Corona Virus or COVID-19 is a novel infection illness that began somewhat recently of 2019. Around the world, as per 25 June 2021, there have been 179M affirmed cases of COVID-19, including 3M deaths, according to WHO. As begun from the most populated country on the planet, China, COVID-19 has spread and killed a large number of individuals from different nations, including Italy, Spain, the USA, Iran, India and numerous different nations. As of late, particularly during the last half of May, while this overall pandemic has relentlessly kept on influencing the existences of millions of every few nations, a few nations have no other arrangement except for to turn to add up to lockdown. Web-based media is exceptionally intelligent, which permits pandemic infection specialists to follow a sickness' spread inside a populace to a profoundly precise degree. The inbuilt highlights of online media stages including Twitter work with ongoing connection among clients and during the pandemic, they utilize this platform to share data about the spread of the pandemic, which can be utilized to educate the arrangement regarding measures to control the spread. Sentiment analysis is a tool based on machine learning methodologies and quite possibly the main fields in natural language processing. Numerous individuals have utilized informal communities, for example, Twitter to state their viewpoints and

perspectives towards COVID-19 and offer their encounters in confronting this infection. In this manner, the significance of Twitter has expanded like never before, and sentiment analysis for Twitter data has become an interesting issue in information science research.

The paper contributions can be summarized as follows:

(1) Fostering a Real time system to foresee the Covid pandemic utilizing Twitter streaming data.
(2) Contrasting machine learning classifications to track down the ideal model used to predict coronavirus sentiment in real-time.

II. Methodology

Data Collection: Twitter is quite possibly the most generally utilized social stages for individuals' cooperation, content posting, sharing, and remarking on different subjects individuals talk about, including medical problems. These days, the world experiences the COVID-19 pandemic, and individuals begin posting a storm of tweets about it. This huge data could be a decent source, however it needs investigation in the age of the COVID-19 pandemic in its inspiration and antagonism. Hence, the supposition examination of Twitter streaming data assumes a part in dispersing clinical data for the Covid pandemic. To realize the system's data collection stage, Twitter APIs have been considered for ingesting streaming information. Twitter's Streaming APIs are arranged into the Search API and Stream API . Twitter's Search API is utilized to assemble chronicled Twitter information offline, while Twitter's Stream API is utilized to stream real-time data through the online stage. For the offline stage, we have gathered recorded Twitter information about Covid in length between 23/01/2020 and 01/06/2020 filtered by #COVID-19 and #Corona infection hashtags. In doing as such, we have made an approval association with Twitter utilizing Twitter Streaming API. Specifically, the OAuth verification convention is utilized to approve applications to get to Twitter administrations. Twitter's Stream API is utilized to stream real-time data from Twitter, which utilizes in the online stage.

Data Preprocessing: Data preprocessing is basic in social network-based analysis system as it straightforwardly impacts the viability of the sentiment analysis because of the data’s complexity. As per our work, in spite of the fact that Twitter is viewed as a gold mine of data, it is viewed as one of the noisiest information since it comprises of numerous connections, hashtags, unique images, emoticons, etc. Consequently, the gathered Twitter information has been preprocessed utilizing the accompanying methods:
noise removal, tokenization, normalization, and stemming

Pretrained Tweets- Sentiment analysis : It recognizes the feelings or attitudes the essayist holds (i.e., Twitter handle/user), regardless of whether these feelings/perspectives can be positive, negative, or neutral. Here, we utilizes TextBlob , a Python library, to perform sentiment analysis on information gathered from Twitter. As controlled by the TextBlob, TextBlob utilizes the Na¨ıve Bayes (NB) model for grouping, and it returns two properties as yields, specifically, polarity and subjectivity. The TextBlob contradiction implies recognizing sentiment direction (positive, neutral, and negative), though subjectivity implies communicating some personal feelings, sentiments, assessments, or convictions. We have utilized the yield polarities of tweets to name the gathered dataset to be fitted in the machine learning models in the assessment step in regards to this work.

Data Splitting: In this progression, the subsequent pretrained dataset is parted into 90% of the training dataset and 10% of the testing dataset utilizing a defined strategy. The training set is utilized to enhance and prepare the machine learning models, while the inconspicuous test set is utilized to assess the machine learning models.

Feature Extraction: One of the text based data analysis challenges is feature extraction because of gaining from the high-dimensional information. It is ideal to utilize some feature extraction strategies to change text into a matrix (or vector) of features. In this way, we have applied two of the most mainstream feature extraction strategies on the truly gathered tweet information, namely, n-gram and TF-IDF. N-gram modeling is a well known feature selection and investigation strategy widely utilized in text mining and normal language processing. As per textual data analysis, n-gram is utilized to register a contiguous sequence of words with length n inside a given window. The term frequency-inverse document frequency (TF IDF) is an acclaimed technique utilized in assessing the significance level of a word in a document utilized in recovering data and natural language processing. The objective of TF-IDF is to compute the word recurrence inside the content in an enormous document corpus. The TF IDF technique utilizes the relative frequency level through the reference document corpus, which can be viewed as good advantage.

Optimization and Training Models: The machine learning models utilized are as the following:
Decision tree (DT), support vector machine (SVM), random forest classifier (RF), logistic regression classifier (LR), and K-nearest neighbor (KNN). For each test, the training set is utilized to advance the hyper parameters of the model by utilizing the grid search procedure. The grid search technique with delineated 10-fold cross-validation (CV) has been utilized to find the ideal hyper parameters of ML algorithms. At last, the subsequent models were assessed by utilizing the concealed test set. We have utilized a grid search with delineated 10-fold cross-validation to track down the ideal hyper parameters of all models.

Assessing the Models: Four standard measurements were used to assess the models' accuracy, precision, recall, and F1-score, where TP is true positive, TN is true negative, FP is false positive, and FN is a false negative given in the accompanying conditions:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} - \text{score} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

III. Conclusions

This paper has presented a real time system for pandemic prediction by analysing Twitter data. Here, Twitter data is collected using Twitter API, then processed and subjected to Sentiment Analysis. The resultant data is split into training test and test data. Using the training data, machine learning models are prepared and evaluated using test data. In this we apply, two most prominent methods of feature extraction, n-gram and TF-IDF. Four machine learning models are used here. Then, these models are evaluated using four standard measurements.

IV. References

[1]. Predicting Coronavirus Pandemic in Real-Time Using Machine Learning and Big Data Streaming System Xiongwei Zhang,1 Hager Saleh , 2,3 Eman M. G. Younis,2 Radhya Sahal , 4 and Abdelmgeid A. Ali2

[2]. Real-time Prediction of the Daily Incidence of COVID-19 in 215 Countries and Territories Using Machine Learning: Model Development and Validation
Yuanyuan Peng1* , BSc; Cuilian Li2* , MBBS; Yibiao Rong3 , PhD; Chi Pui Pang2,4 , DPhil; Xinjian Chen1 , PhD; Haoyu Chen2 , MD

[3].What social media told us in the time of COVID-19: a scoping review
Shu-Feng Tsao, Helen Chen, Therese Tisseverasinghe, Yang Yang, Lianghua Li, Zahid A Butt

[4]. How to Predict the Return of COVID-19? A Warning System Using Social Media Data Shan Shan Yulei Li

[5]. Predicting the pandemic: sentiment evaluation and predictive analysis from large-scale tweets on Covid-19 by deep convolutional neural network Sourav Das & Anup Kumar Kolya

[6]. Public Sentiment Analysis on Twitter Data during COVID-19 Outbreak Mohammad Abu Kausar1 , Arockiasamy Soosaimanickam2 , Mohammad Nasar3

[7]. A Proposed Sentiment Analysis Deep Learning Algorithm for Analyzing COVID-19 Tweets Harleen Kaur, Shafqat Ul Ahsaan, Bhavya Alankar & Victor Chang

[8]. Prediction of COVID-19 Waves Using Social Media and Google Search: A Case Study of the US and Canada Samira Yousefinaghani[1], Rozita Dara[1*], Samira Mubareka[2] and Shayan Sharif[3]

[9]. COVID-19 Outbreak Prediction with Machine Learning Sina F. Ardabili 1 , Amir Mosavi 2,3,*, Pedram Ghamisi 4 , Filip Ferdinand 2 , Annamaria R. Varkonyi-Koczy 2 , Uwe Reuter 3 , Timon Rabczuk 5 , Peter M. Atkinson 6

[10]. COVID-19 Outbreak Prediction with Machine Learning Sina F. Ardabili 1 , Amir Mosavi 2,3,4,* , Pedram Ghamisi 5,6 , Filip Ferdinand 7 , Annamaria R. Varkonyi-Koczy 3,4 , Uwe Reuter 8 , Timon Rabczuk 9 and Peter M. Atkinson 10

[11]. WHO Coronavirus (COVID-19) Dashboard
https://covid19.who.int