Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Machine Learning Based Twitter Data Mining to Analyse Sentiments of Tweets Allied to COVID-19 Epidemic & Its Patterns

[1]Rameshwer Singh and [2]Dr. Rajeshwar Singh

[1]Research Scholar, I.K. Gujral Punjab Technical University, Kapurthala, Punjab, INDIA

[2] Group Director, Doaba Khalsa Trust Group of Institutions, SBS Nagar, Punjab, INDIA

E-mail: [1]rameshwer.banga@gmail.com, [2]directordgc3@gmail.com

ABSTRACT

Sentiment analysis is the study of sentiments shared by users with the help of social networking sites like, Twitter, Facebook etc. Twitter is a microblogging site that contains abundant amount of data in the form of tweets on different topics. It is fruitful to perform sentiment analysis on datasets of twitter to extract fruitful information. In this research paper we present the sentiment analysis of tweets related to ongoing epidemic COVID-19, Corona Virus. It is declared as pandemic by World Health Organization (WHO) in the mid of March 2020. Our analysis is based on more than 40000 tweets of Corona Virus related hashtags. These tweets are collected from twitter between 20/03/2020 to 12/04/2020 using Twitter RESTFul API. We have used python based libraries for data collection, data cleaning, data analysis and graphical representation of results. Further, paper also highlights the spread pattern of COVID-19 worldwide with the help geographical information available in the tweets. In this paper sentiment analysis is performed using NLTK and TextBlob python libraries. Outcome shows that there are 39.3 % positive, 16.9 % negative and 43.8 % neutral tweets found in our data sets. We have done experimentation on datasets using different machine learning algorithms and our classifiers achieved accuracy up to 98%. As per our results Linear SVC performed better than other algorithms. Average classification accuracy of our algorithms used is more than 93%.   We also present the word clouds of tweets, hashtags and geo information of the tweets in the paper. Further we present how to identify appropriate hashtags for tweet search and how twitter data can be helpful to study the patterns of epidemic like COVID-19.

Keywords: *Sentiment Analysis, COVID-19, twitter data processing, machine learning*

1. INTRODUCTION

Opinion mining and sentiment analysis is the analysis of sentiments of people shared on microblogging sites such as Twitter. These sentiments are very much important and can play major role in predictions [1,2]. Sentiment analysis is useful in extracting the useful information from tweets [3]. COVID-19 is declared as pandemic by WHO on 13 March, 2020. First case of COVID-19 was reported in December 2019 from China. Wuhan is assumed as origin of this epidemic which become pandemic in the month of March 2020. Symptoms of COVID-19 are very common like dry cough, fever, throat infection, difficulty in berating etc. As per the studies there on no standard treatment is identified about the disease till date that’s why this disease is causing deaths. As per WHO (https://who.sprinklr.com/) total 1,211,214 cases are confirmed worldwide and 67,666 deaths reported till 06 April, 2020 [4]. As per paper [5], 87% of the COVID-19 cases reported in China were between the age of 30 to 79. Second meeting of emergency committee convened by WHO stated that COVID-19 epidemic outbreak is dangerous for the entire world [6]. Hence this epidemic is becoming the biggest topic of concern for the entire world. That’s why we have chosen COVID-19 epidemic for sentiments analysis of tweets.

There is huge amount of data available on the Twitter on each and every possible topic of the world. This is why Twitter is referred as gold mine for data scientists. Further Twitter also provides its own RESTful API. This API provides the access to Twitter data.

Sentiments of the people on microblogging sites are useful for the prediction of outbreaks and also its relation with actual data related to particular disease. Research papers [7,8,9] represents the significance of Twitter in predicting the outbreak. Accuracy of prediction depends upon the quality of data that we are using for our analysis. Article [10,11,12,13,14] presents the significance of data collection and cleaning to generate accurate output. In this paper, we have used Twitter API and we accessed the data available on Twitter with the help of Python programming language. While developing the program to fetch the tweets from Twitter in python, we have used some Python libraries. Tweepy API, it provides access to all the RESTful API methods of Twitter. Tweepy contains variety of methods that can be used to access tweets through twitter application. In our program we have used this API to authenticate our API keys so that we can access our application on Twitter to download the tweets [15]. Natural Language Processing Toolkit (NLTK), It is the python based platform that is used to deal with Natural Language Processing(NLP). It has many text processing functions which can be used for tokenization, classification, parsing etc. Further it also has properties to calculate polarity and subjectivity of the text [16]. TextBlob, it is another famous python library for textual data processing. It is useful to process the NLP. We can perform sentiment analysis, Part-Of-Speech Tagging, Translation etc. using TextBlob library [17]. Pandas, using Pandas library in python we can create python object from the input files in CSV, TSV, SQL format. It provides easy way to organize the data for processing and deriving the results from the data [18]. Similarly, wordcloud, matplotlib etc are used in the implementation of our proposed work in the paper.

Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

2. BACKGROUND

Research paper [19], presents the different techniques to perform sentiment analysis. Different phases to perform sentiment analysis are discussed in the paper. Paper presents both lexicon based and machine learning techniques based approach to perform sentiment analysis. In paper [20], authors describe the features of language that can be used to find the sentiments in the tweets. In this paper hashtags based tweets are collected and performed sentiment analysis on the same.

Paper [21], presents the study based on movie reviews. In the paper both approaches are presented i.e. machine learning and lexicon based approach. Outcome of the study shows that machine learning based approach outperformed the lexicon based approach. Similarly, paper [22], also presents the sentiment analysis on movie reviews by using Naïve Bayes. In paper [23], authors present the use of machine learning techniques, lexicon based approach and hybrid techniques to perform sentiment analysis. Paper shows that hybrid approach is suited where we have limited training data sets.

Study of the sentiments on twitter are useful to generate important health care information. This kind of study can be helpful in predicting the outbreak and its early detection. In paper [24], proposed model uses the twitter data sets and its analysis to find its relation with actual data sets. The study in the paper is based on swine flu epidemic. In similar way paper [25], conducted the study on Zika Virus based on the data collected from twitter and google search data for the predictions. This is the reason that we have chosen COVID-19 pandemic for our sentiment analysis.

3. DATA COLLECTION & CLEANING

We have used Twitter RESTful API to fetch the data from Twitter. We have developed a python based crawler that fetches the tweets from twitter having specified hashtags. Data collected from twitter tends to be noisy due to the use of special symbols, urls, emoji and other special symbols. Preprocessing of twitter data is very much required as it is difficult to produce accurate results with the noisy data. We have opted following procedure to fetch and clean the data related to COVID-19.

Algorithm for twitter data collection and cleaning:

- **Step 1: Create Twitter Developer Account and Twitter App:**
  In order to gather data from twitter, we need to follow a step by step process. In the very first step we need to create/ register Twitter Developer account on developer.twitter.com. After that we need to create an App on Twitter which will provide the access keys to access the twitter resources via programming [26,27].
- **Step 2: Generate Access Keys of Twitter Application**
  Create app on Twitter using Twitter Developer Account. After that Generate Access Keys of Twitter Application. These application keys will be used to access twitter resources and shall be used for authentication.
- **Step 3: Authentication using access tokens**
  This object will provide the access to Twitter RESTful API we need to pass the consumer key and consumer secret into it.
- **Step 4: Accessing and Saving Twitter Data**
  In the next step we have collected the twitted with hashtags #covid19, #covid-19, #covid, #coronavirus, #ncob2019, #novelcoronavirus, #covid—19, #2019NovelCoronavirus, #SARS-CoV-2 OR #covid-19 who, #covid-19india, #covid19india, #cornaupdateindia, #covid19 india, #coronavirus italy, #corona, #coronaindia, #COVID19, #COVID19INDIA, #nCOV, #nCOVID, #n-Covid' and saved them into CSV file.
- **Step 5: Cleaning of Tweets**
  Twitter data tends to be very noisy as it is basically natural language posed by different users. So there is a need of cleaning the tweets. In order to clean the tweets, we have removed the stopwords, extra spaces, additional symbols, Emoticons from the tweets.
- **Step 6 Create Data File**
  In the last step we have saved the save the data into the file. We have saved the cleaned tweets and other parameters in the form CSV file.

We have collected and cleaned more than 40000 tweets based on the aforementioned algorithm which are used for its sentiment analysis.

4. DATA ANALYSIS

4.1 SENTIMENT ANALYSIS OF TWEETS

After the collection and cleaning of tweets, we have applied the sentiment analysis in python using TextBlob & NLTK libraries. We have calculated the subjectivity, polarity of each tweet in CSV file. Value of subjectivity varies between 0.0 to 1.0 where 0.0 depicts that tweet is highly objective and 1.0 means tweet is highly subjective. Similarly, the score of polarity lies between -1 to 1, where -1 means highly negative and +1 means highly positive. So as per values of subjectivity and polarity sentiment is decided by using TextBlob. After that based of the sentiment score each tweet is marked as Positive, Negative or Neutral.

In our study, we have collected more than 40000 between 20 March 2020 to 12 April 2020 tweets allied to COVID-19 for sentiment analysis. As the result if the sentiment analysis we got 39.3% positive tweets, 16.9 % negative tweets and 43.8 % are neutral tweets.

Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Figure 1 Sentiments of each tweet calculated and added in the data using polarity of the tweet

	created_at	original_text	clean_text	sentiment	analysis
0	Fri Mar 20 00:43:49 +0000 2020	RT @PhilstarShowbiz: Kim Chiu @prinsesachinita...	Kim Chiu reacts co-star Christopher de Leon 's...	Sentiment(polarity=0.0, subjectivity=0.0)	Neutral
1	Fri Mar 20 00:57:02 +0000 2020	@CDCgov Social distancing implies #nCOV #covid...	Social distancing implies airborne .. aerosol ...	Sentiment(polarity=0.03333333333333333, subjec...	Positive
2	Fri Mar 20 01:01:12 +0000 2020	Is #nCOV #covid19 #coronavirus aerosol or drop...	Is aerosol droplet Aerosol implies fine partic...	Sentiment(polarity=0.41666666666666667, subject...	Positive
3	Fri Mar 20 01:11:37 +0000 2020	RT @bladerunner3049: Is #nCOV #covid19 #corona...	Is aerosol droplet Aerosol implies fine partic...	Sentiment(polarity=0.41666666666666667, subject...	Positive
4	Fri Mar 20 01:29:22 +0000 2020	#homeworkout remedy\r\n\r\n#Covid19 #Ncov #Cor...	remedy ...	Sentiment(polarity=0.0, subjectivity=0.0)	Neutral
...	...	...	...	...	...
19232	Fri Apr 03 21:12:05 +0000 2020	RT @anabevc: In all the #COVID19 madness, this...	In madness discovery made day soundsof cities ...	Sentiment(polarity=0.0, subjectivity=0.0)	Neutral
19233	Fri Apr 03 21:12:04 +0000 2020	RT @NGrossman81: @jeff82874662 @atrupar Seriou...	Serious cases COVID-19 new category France 's ...	Sentiment(polarity=-0.09848484848484848, subje...	Negative
19234	Fri Apr 03 21:12:03 +0000 2020	RT @plntbasedcutie: CALIFORNIA Our peak is pre...	CALIFORNIAOur peak predicted hit April 16th Th...	Sentiment(polarity=0.3111111111111111, subject...	Positive
19235	Fri Apr 03 21:12:02 +0000 2020	RT @SpiezLab: The #COVID19 crisis stretches ev...	The crisis stretches every part system mandate...	Sentiment(polarity=0.0, subjectivity=0.0)	Neutral
19236	Fri Apr 03 21:12:01 +0000 2020	RT @c9o9c9t9l: Thanks Mr. Ambassador for your ...	Thanks Mr Ambassador kind fair remarks We SA h...	Sentiment(polarity=0.575, subjectivity=0.75)	Positive

19237 rows × 5 columns

Table 1 Sample of tweets with different types of sentiment and their sentiment score

Sr. No	Tweet (Original Text)	Tweet (Cleaned Text)	Sentiment Score	Sentiment Type
1.	@spain Tell everyone STAY AT HOME. You have the worst death rate in Europe from #coronavirus - it's doubling everyâ€¦ https://t.co/kXbclcMizd	Tell everyone STAY AT HOME You worst death rate Europe 's doubling everyâ€¦	-1	Negative
2.	#COVID19 for #veterans in #NewOrleans continues to be especially bad. #VA caring for 1,600 veterans infected. 300 oâ€¦ https://t.co/mACSgUyTdo	continues especially bad caring 1,600 veterans infected 300 oâ€¦	-1	Negative
3.	@RT_com @TheOliverStone Please sign the petition to lift these inhumane US sanctions on #Iran and help it to fightâ€¦ https://t.co/8nRmI5zbgQ	Please sign petition lift inhumane US sanctions help fightâ€¦	-0.9	Negative
4.	@ChinaDaily I HATE YOU CHINA #coronavirus #CoronavirusPandemic you evil midgets	I HATE YOU CHINA evil midgets	-0.9	Negative
5.	RT @XHNews: A Chinese plane landed in Jakarta loaded with medical supplies Indonesia purchased from China to fight #coronavirus. #COVID19 hâ€¦	A Chinese plane landed Jakarta loaded medical supplies Indonesia purchased China fight hâ€¦	0	Neutral
6.	RT @JamesTodaroMD: The National Task Force for COVID-19 in India recommends HYDROXYCHLOROQUINE for PROPHYLAXIS against #COVID19 in "Asymptoâ€¦	The National Task Force COVID-19 India recommends HYDROXYCHLOROQUINE PROPHYLAXIS `` Asymptoâ€¦	0	Neutral
7.	Coronavirus in Lebanon: Delightful picture of a family riding a scooter wearing facemasks #coronavirusâ€¦ https://t.co/elzj5VAp8O	Coronavirus Lebanon Delightful picture family riding scooter wearing facemasks â€¦	1	Positive
8.	RT @TheOfficialSBI: Prevention is the best way to fight Coronavirus. Our staff from Bengaluru Circle ensures the safety of customers by encâ€¦	Prevention best way fight Coronavirus Our staff Bengaluru Circle ensures safety customers encâ€¦	1	Positive
9.	RT @TheOfficialSBI: Prevention is the best way to fight Coronavirus. Our staff from Bengaluru Circle ensures the safety of customers by encâ€¦	Prevention best way fight Coronavirus Our staff Bengaluru Circle ensures safety customers encâ€¦	1	Positive
10.	RT @PemaKhanduBJP: We are preparing best to meet any challenges on #COVID19. Inspected the preparedness level in Tomo Riba Institute of Heaâ€¦	We preparing best meet challenges Inspected preparedness level Tomo Riba Institute Heaâ€¦	1	Positive

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

The sentiment analysis of tweets demonstrates that the negative tweets are showing high use of word like death, infected etc. Which indicates that some serious epidemic outbreak is spreading with whom world is dealing. In case of neutral tweets data seems to be like some sort of information is being given. Further positive tweets are presenting most of the government initiatives and how public is responding to it. Total sentiment based count of tweets is as below:

Sentiment	Tweet Count	Percentage
Positive	15965	39.3
Negative	6854	16.9
Natural	17775	43.8

In the figure 2 graphs below we have plotted the sentiment of each tweet based on polarity and subjectivity of the tweets and figure 3 graph presents the count of positive, negative and neutral tweets.

Figure 2 Sentiment Analysis Graph Based on Polarity & Subjectivity

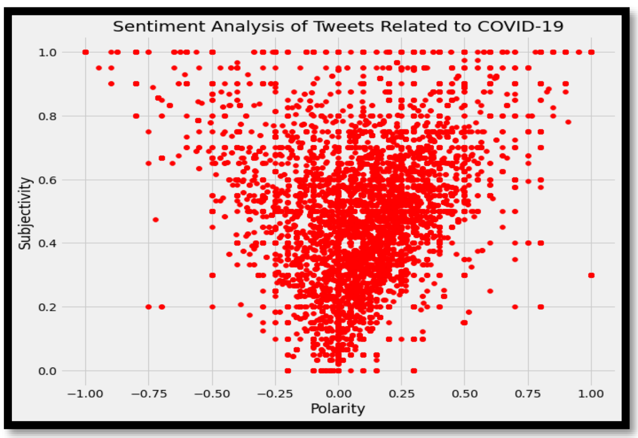
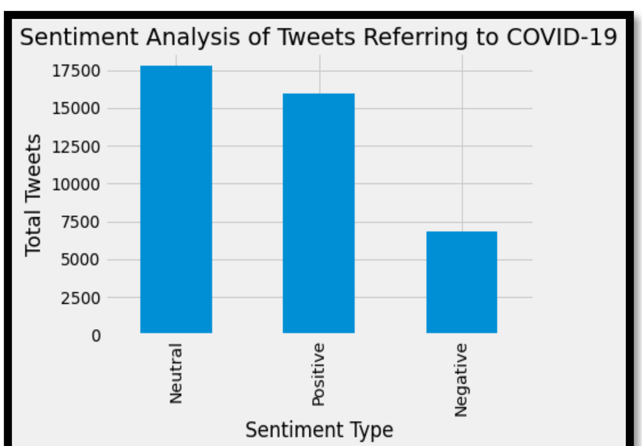


Figure 3 Count of Positive, Negative & Natural Tweet



4.2 MACHINE LEARNING TECHNIQUES BASED CLASSIFICATION OF SENTIMENTS

In the research paper, we have tried different machine learning based classification techniques to select the best classifier. We have performed experimentations using python Sci-kit Learn library. In this paper we have used Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, Linear SVC, Ada Boost Classifier, Ridge Classifier, Passive Aggressive Classifier, Perceptron machine learning classifiers.

We performed all the techniques with unigram, bigram, trigram and n-gram. We also presented the classification report for each algorithm. Which presents precision, recall, f1-suppor and support value of each and very classifier that we have used. We have used notation in equations as True Positive as TP, False Positive as FP, True Negative as TN and False Negative as FN. Precision is defined as the value of true positive divided by the sum of true positive and false positive. High precision means that we have less false positive ratio. Recall is the ratio between true positive divided by the sum of true positive and false negative. High recall means correctly classified. Accuracy is the ratio between sum of true positive and true negative divided by all observation i.e. classified correctly and classified incorrectly. F1 Score, is the harmonic mean calculated by using precision and recall value.

$$\textbf{Precision} = \frac{TP}{TP + FP} \qquad \text{(I)}$$

$$\textbf{Recall} = \frac{TP}{TP + FN} \qquad \text{(II)}$$

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{(III)}$$

$$\textbf{F1 Score} = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \text{(IV)}$$

In our experimentation we have calculated the sentiment score using TextBlob Sentiment analysis after that we have labelled the tweets into three categories -1 to represent negative sentiment, 0 to represent neutral sentiment and 1 to represent positive sentiment. These labels are used by different classifiers to perform classification using test and training sets.

Table 2 Performance of Different Machine Learning Classifiers with Unigram, Bigram and Trigram

Approach →	**Unigram**					**Bigram**				**Trigram**			
Machine Learning Technique	**Class**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Multinomial Naive Bayes	-1	0.90	0.83	0.87	0.89	0.94	0.88	0.91	0.93	0.96	0.88	0.92	0.94
	0	0.93	0.87	0.90		0.95	0.93	0.94		0.95	0.94	0.94	
	1	0.85	0.95	0.89		0.91	0.96	0.93		0.91	0.96	0.94	
Bernoulli Naive Bayes	-1	0.94	0.80	0.87	0.91	1.00	0.80	0.88	0.92	1.00	0.78	0.88	0.92
	0	0.93	0.91	0.92		0.90	0.97	0.93		0.87	0.98	0.92	
	1	0.87	0.94	0.91		0.93	0.93	0.93		0.96	0.90	0.93	
Logistic Regression	-1	0.98	0.93	0.96	0.97	0.99	0.92	0.95	0.97	0.99	0.90	0.94	0.97
	0	0.96	0.99	0.98		0.95	1.00	0.97		0.94	1.00	0.97	
	1	0.98	0.98	0.98		0.99	0.96	0.98		0.99	0.96	0.97	
Linear SVC	-1	0.98	0.96	0.97	0.98	0.99	0.93	0.96	0.97	0.98	0.92	0.95	0.97
	0	0.97	0.99	0.98		0.96	0.99	0.98		0.95	0.99	0.97	
	1	0.99	0.98	0.98		0.99	0.97	0.98		0.99	0.97	0.98	
Ada Boost Classifier	-1	0.89	0.58	0.70	0.78	0.89	0.58	0.70	0.78	0.89	0.58	0.70	0.78
	0	0.70	0.99	0.82		0.70	0.99	0.82		0.70	0.99	0.82	
	1	0.93	0.62	0.74		0.93	0.62	0.74		0.93	0.62	0.74	
Ridge Classifier	-1	0.97	0.93	0.95	0.96	0.99	0.94	0.96	0.97	0.99	0.91	0.95	0.97
	0	0.94	0.98	0.96		0.95	0.99	0.97		0.94	1.00	0.97	
	1	0.97	0.94	0.96		0.99	0.96	0.97		0.99	0.96	0.97	
Passive Aggressive	-1	0.98	0.95	0.97	0.98	0.98	0.92	0.95	0.97	0.98	0.91	0.95	0.97
	0	0.97	0.99	0.98		0.95	0.99	0.97		0.94	0.99	0.97	
	1	0.99	0.98	0.98		0.99	0.97	0.98		0.99	0.96	0.97	
Perceptron	-1	0.98	0.94	0.96	0.97	0.98	0.94	0.96	0.97	0.98	0.92	0.95	0.97
	0	0.97	0.98	0.97		0.96	0.99	0.98		0.95	0.99	0.97	
	1	0.98	0.98	0.98		0.98	0.97	0.98		0.99	0.97	0.98	

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Figure 4 Performance of Different Machine Learning Classifiers with Unigram

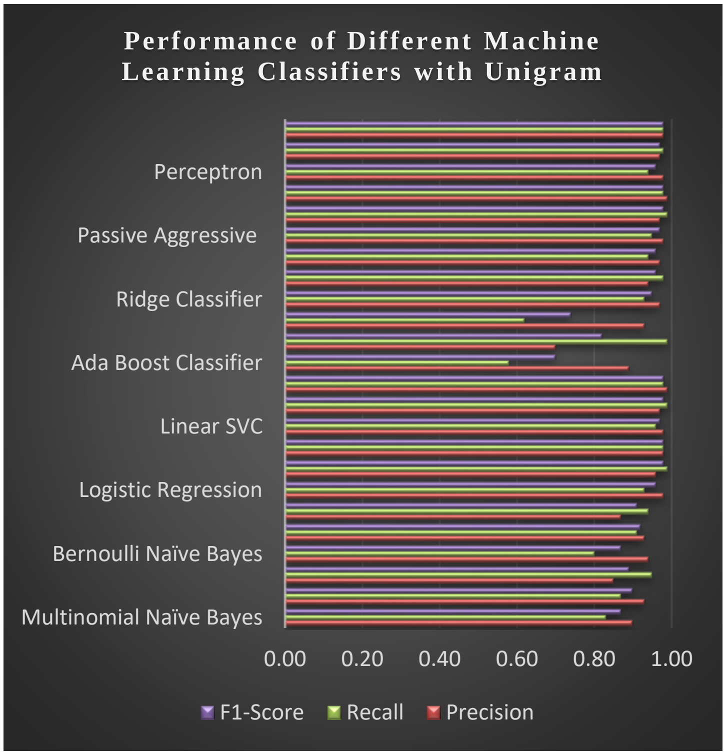


Figure 5 Performance of Different Machine Learning Classifiers with Bigram

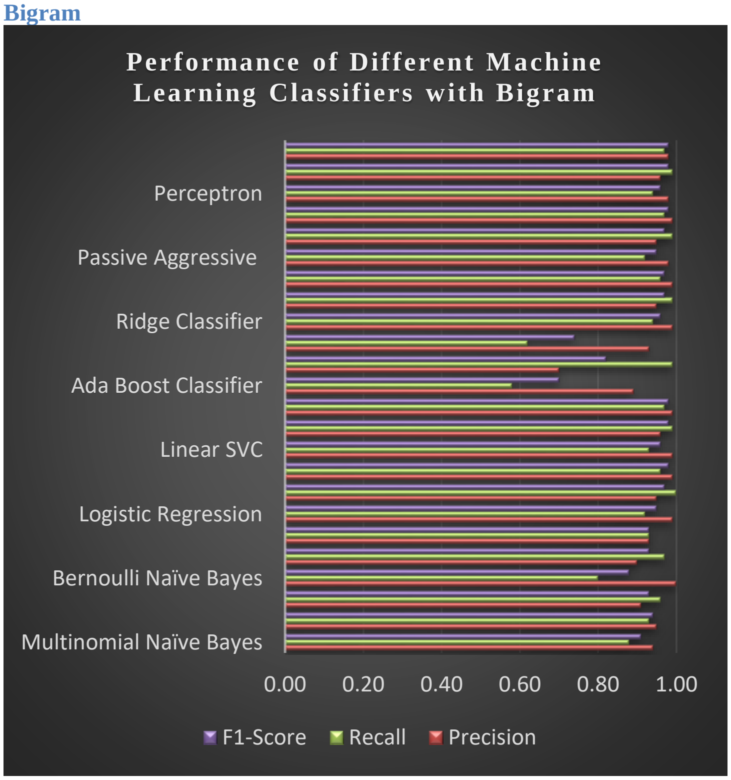
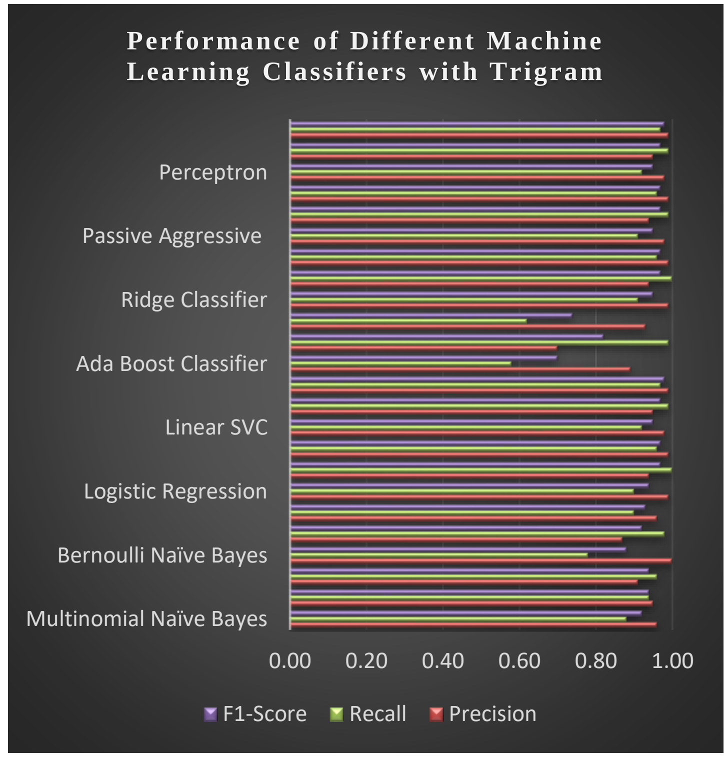


Figure 6 Performance of Different Machine Learning Classifiers with Trigram



Further we have applied k-fold cross validation on all the classifier and to test the accuracy and performance of all the algorithms that we have used in out experimentation. We have set the number or folds to 10. Results shows that Linear SVC performs very well in all the kinds i.e unigram, bigram and trigram. Detailed results are show as below in table 3.

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Table 3 Results of K-Cross Validation of Different Classifiers

Type	Machine Learning Classifier	Accuracy	Mean CrollVal Train Score	Mean Crossval Test Score	Standard Deviation CrossVal Train Score	Standard Deviation CrossllVal Test Score	Time	Accuracy Ranking
Bigram	Linear SVC	0.974180735	0.99997242	0.967152131	9.19337E-06	0.004096596	171.3898802	1
Trigram	Linear SVC	0.971698113	0.99997242	0.961001754	9.19E-06	0.003988936	378.150995	1
Unigram	Linear SVC	0.979145978	0.999849841	0.975150336	9.92534E-05	0.004627108	73.00034666	1
Unigram	Passive Aggressive Classifier	0.978152929	0.999779358	0.972419963	0.000154323	0.004920666	27.47272086	2
Bigram	Perceptron	0.973187686	0.999954033	0.964642326	2.47067E-05	0.003999621	34.83742785	2
Trigram	Perceptron	0.968718967	0.999957098	0.963290785	2.03E-05	0.004073521	57.9482882	2
Unigram	Logistic Regression	0.974180735	0.996613775	0.966848721	0.000169093	0.004232665	39.70787144	3
Trigram	Ridge Classifier	0.96673287	0.99997242	0.957636952	9.19E-06	0.004421426	156.9782124	3
Bigram	Passive Aggressive Classifier	0.971698113	0.99997242	0.96588344	9.19337E-06	0.003697691	52.84428239	3
Unigram	Perceptron	0.973187686	0.999687426	0.967344999	0.000243157	0.004607928	28.50086379	4
Trigram	Logistic Regression	0.965243297	0.999929517	0.955402933	1.40E-05	0.004360804	289.0346026	4.5
Bigram	Ridge Classifier	0.969712016	0.99997242	0.961718835	9.19337E-06	0.004067433	113.9167092	4.5
Bigram	Logistic Regression	0.969712016	0.999911131	0.960422565	1.6503E-05	0.003773465	99.44127059	4.5
Trigram	Passive Aggressive Classifier	0.965243297	0.99997242	0.959733039	9.19E-06	0.003977444	49.01539636	4.5
Unigram	Ridge Classifier	0.957795432	0.998415676	0.956092489	0.000114696	0.003390243	42.69442129	5
Trigram	Multinomial Naïve Bayes	0.935451837	0.995222509	0.923272176	0.00027801	0.005922788	66.69413209	6
Unigram	Bernoulli Naïve Bayes	0.905660377	0.935796551	0.892106769	0.000804414	0.006586738	11.56183171	6
Bigram	Multinomial Naïve Bayes	0.932472691	0.990175347	0.920045275	0.000348945	0.006485418	22.73764229	6
Trigram	Bernoulli Naïve Bayes	0.917080437	0.980942138	0.90575867	0.000120472	0.004112417	69.60406661	7
Bigram	Bernoulli Naïve Bayes	0.924031778	0.979970703	0.911605736	0.000222399	0.004632872	22.75503802	7
Unigram	Multinomial Naïve Bayes	0.892750745	0.931104861	0.885487434	0.000705953	0.006137698	11.30515122	7
Unigram	Ada Boost Classifier	0.7775571	0.775641847	0.774173932	0.001510247	0.007758435	76.17947006	8
Trigram	Ada Boost Classifier	0.7775571	0.775641847	0.774173932	0.001510247	0.007758435	244.4566879	8
Bigram	Ada Boost Classifier	0.7775571	0.775641847	0.774173932	0.001510247	0.007758435	78.21291232	8

Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

4.3 RECURRENT KEYWORDS AND HASHTAGS IN TWEETS

After the processing we have calculated the most used word in the tweets. We have used the Word Cloud python library to show the words cloud. We have created two clouds one contains the words from original tweets and other contains the words from cleaned tweets as shown below:

Figure 7 Word Cloud based on original text of tweets

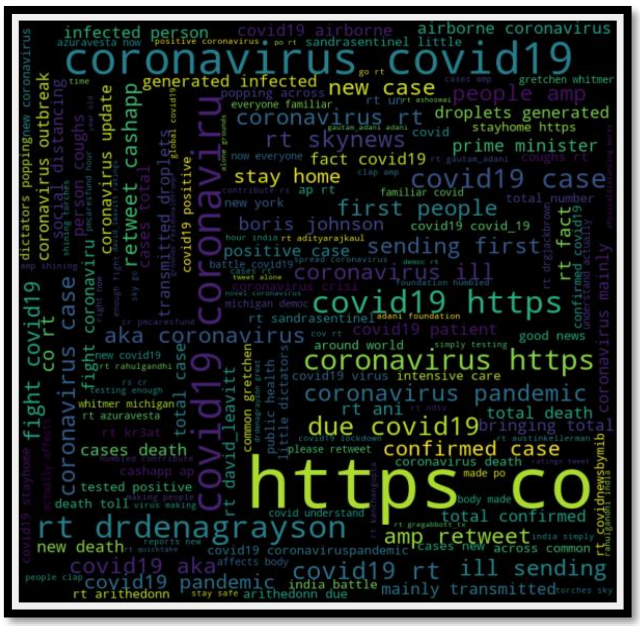
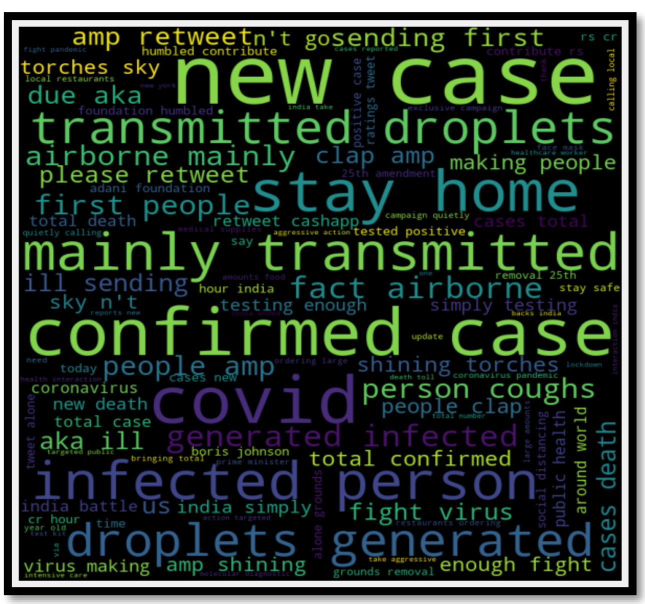


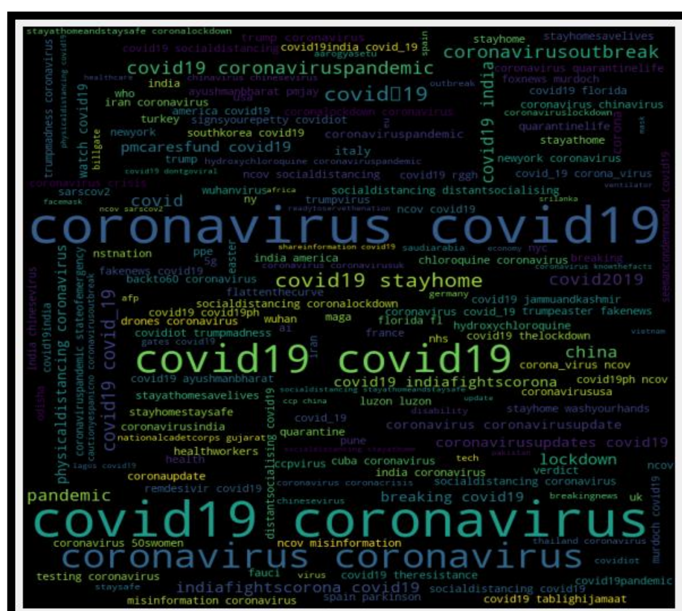
Figure 8 Word Cloud based on original cleaned text of tweets



As shown in figure 7, word cloud based of uncleaned tweets. It is clear that there is very high use of keywords like covid19, coronavirus, confirmed cases, fight coronavirus etc. http is also highlighted as many tweets contain reference to some videos, images, websites etc. Similarly figure 8, is showing the commonly used words in cleaned tweet data sets. We can see the dominated keywords like confirmed cases, covid, mainly transmitted, infected person etc. Hence word cloud of cleaned tweets seems more significant as not referring useless terms. So benefit of the cleaning the tweets can be clearly identified here. Such high used of these kinds of word is useful in generating hint about the COVID-19 epidemic. Further the deep analysis of the same also gives us the hint reading modes which may be the cause of spreading the disease. For instance, use of word infected person, droplets generated, mainly transmitted etc. are providing information about how disease may spread. Hence high use of keyword like mainly transmitted is indicating that its infectious in nature. Thus it may be become an epidemic.

Similarly, in figure 9, word cloud hashtags used in the data presents the most frequently used has hashtags for tweets about COVID-19. This is usable to make more accurate search or tweets related to COVID-19 as the word cloud shows most dominant hashtags that are in our data sets like coronavirus, covid-19 etc. We have noticed some other hashtags which are being in frequent use like socialdistancing, quarantine etc. These kind of missing hashtags can be added in our searches to get more related data. It will help in better selection for hashtags for current and future researches.

Figure 9 Popular hashtags related to COVID-19



Volume 9, Issue 2
April 2020

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

4.4 ANALYSIS OF TWEET GEO LOCATION INFORMATION TO ANALYZE EPIDEMIC PATTERNS

Twitter data mining is also helpful in studying the epidemic patterns with the help geotagging information available in the tweets. As shown in figure 10, there is huge variety of courtiers, places from where the tweets are being posted. It's clear from the geo location information of the tweets that which country seems to be most affected place with the COVID-19 epidemic may face effects of COVID-19.

Figure 10 Top courtiers, places from where the tweets are being posted

If the number of different places is increased from where tweet is being posted. It will increase the chances of the disease to convert into an epidemic or pandemic. It is clear from the word cloud related to location information in tweets that many countries are posting tweets related to COVID-19 like India, UK, USA and many other. Hence, location analysis is very important to understand spread patterns of the COVID-19.

The duration in which we have collected the tweets (20 March, 2020 to 12 April, 2020) was very sensitive time for India as on 20 March, 2020 India was very near to lockdown which is prevention mechanism followed by many countries. Similarly, cases of COVID-19 increased quickly in USA during the phase we have collected the datasets. Hence its effect is also visible through geo location information in tweets. This kind of hint can be useful to make prediction or early detection of epidemic like COVID-19.

5. CONCLUSION

In this research paper, we have performed the sentiment analysis on recent COVID-19 outbreak using twitter. Our analysis shows that such kind of studies are very helpful to generate the hint about the epidemics, early detection of epidemics and spread patterns of the epidemic. In our research paper, we have found that the higher amount of tweets is categorized under the category of neutral as the nature of most of the tweets were advisory or the government planning about the COVID-19 outbreak. We have also performed experimentation on classification of tweets using popular machine learning algorithms our classifiers able to achieve accuracy up to 98%. Paper also shows that Linear SVC is the most optimal algorithm as it produces the high accuracy. Similarly, Passive Aggressive Classifier, Perceptron, Logistic Regression also able to achieve 95% to 97% accuracy in classification. Ridge Classifier, Further paper also shows that how researchers can identify relevant hashtags that should be used while fetching the data from twitter. Paper also presents the idea how geo location information in the tweets can be helpful to analyze the patterns of epidemics like COVID-19. These kinds of analysis are useful to build better models that can predict the epidemic outbreak and its patterns.

REFERENCES

[1] Annett, Michelle, and Grzegorz Kondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." Advances in artificial intelligence. Springer Berlin Heidelberg, 2008. 25-35.

[2] Paul, Michael J., and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." ICWSM. 2011.

[3] Singh, Rameshwer, Rajeshwar Singh, and Ajay Bhatia. "Sentiment analysis using machine learning techniques to predict outbreaks and epidemics".

[4] Coronavirus (COVID-19), WHO Online: Avaible on [https://who.sprinklr.com/ as accessed on 06 April, 2020].

[5] Wu, Zunyou, and Jennifer M. McGoogan. "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention." Jama (2020).

[6] World Health Organization (WHO), "Statement on the second meeting of the International Health Regulations

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

(2005) Emergency Committee regarding the outbreak of novel coronavirus (2019-nCoV) on 30 January 2020 Online. Available [https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov) as assed on 22/02/2020].

[7] Albances, L. Z., et al. "Application of C5. 0 Algorithm to Flu Prediction Using Twitter Data." 2018 International Conference on Platform Technology and Service (PlatCon). IEEE, 2018.

[8] Xue, Hongxin, et al. "Regional level influenza study based on Twitter and machine learning method." PloS one 14.4 (2019): e0215600.

[9] Yanqing Shen, Covid-19 Outbreak: Tweet Analysis on Face Masks Online. Avaible:[https://towardsdatascience.com/covid-19-outbreak-tweet-analysis-on-face-masks-27ef5db199dd as accessed on 28/03/2020].

[10] Dilan K Jayasekara, Extracting Twitter Data, Pre Processing and Sentiment Analysis using Python 3.0 Online. Avaiable:[https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf as accessed on 25/02/2020].

[11] Anas Al-Masri, Creating The Twitter Sentiment Analysis Program in Python with Naive Bayes Classification Online. Avaiable [https://towardsdatascience.com/creating-the-twitter-sentiment-analysis-program-in-python-with-naive-bayes-classification-672e5589a7ed as accessed on 25/02/2020].

[12] Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," PLoS ONE, vol. 5, no. 11, p. e14118, 2010.

[13] Tweepy, An easy-to-use Python library for accessing the Twitter API. [Online]. Available [http://docs.tweepy.org/en/latest/getting_started.html#introduction as accessed 01/03/2020].

[14] Ranjan, Sandeep, and Sumesh Sood. "Social network investor sentiments for predicting stock price trends." International Journal of Scientific Research and Review 7.02 (2019): 90-97.

[15] Tweepy, An easy-to-use Python library for accessing the Twitter API. [Online]. Available [http://docs.tweepy.org/en/latest/getting_started.html#introduction as accessed 01/03/2020].

[16] NLTK 3.5b1 documentation [Online]. Available [https://www.nltk.org/ as accessed on 01/03/2020.

[17] TextBlob: Simplified Text Processing [Online]. Available [https://textblob.readthedocs.io/en/dev/ as accessed on 01/03/2020.

[18] A Quick Introduction to the "Pandas" Python Library [Online]. Available [https://towardsdatascience.com/a-quick-introduction-to-the-pandas-python-library-f1b678f34673 as accessed on 01/03/2020].

[19] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." Ain Shams Engineering Journal 5.4, Elsevier (2014): 1093-1113.

[20] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." ICWSM 11 (2011): 538-541.

[21] Annett, Michelle, and Grzegorz Kondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." Advances in artificial intelligence. Springer Berlin Heidelberg, 2008. 25-35.

[22] Kumari, Pooja, et al. "Sentiment Analysis of Tweets." IJSTE (2015).

[23] Kharche, Ms Swapna R., and Lokesh Bijole. "Review on Sentiment Analysis of Twitter Data." International Journal Of Computer Science And Applications 8.2 (2015).

[24] Grover, Sangeeta, and Gagangeet Singh Aujla. "Twitter data based prediction model for influenza epidemic." 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom). IEEE, 2015.

[25] McGough, Sarah F., et al. "Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data." PLoS neglected tropical diseases 11.1 (2017): e0005295.

[26] Yanqing Shen, Covid-19 Outbreak: Tweet Analysis on Face Masks Online. Avaible:[https://towardsdatascience.com/covid-19-outbreak-tweet-analysis-on-face-masks-27ef5db199dd as accessed on 28/03/2020].

[27] Dilan K Jayasekara, Extracting Twitter Data, Pre Processing and Sentiment Analysis using Python 3.0 Online. Available: [https://towardsdatascience.com/extracting-twitter-data-pre-processing-and-sentiment-analysis-using-python-3-0-7192bd8b47cf as accessed on 25/02/2020].

AUTHORS PROFILE

RAMESHWER SINGH, received his degree of MCA from IKG-PTU. He is pursuing Ph.D. in App-Sci-Computer Applications from IKG-PTU, INDIA. His areas of interest are Machine Learning, Data Mining, NLP. He is specialized in JAVA, Python, Android and PHP programming. He is having more than 11 years of experience as faculty of Computer Science.

DR. RAJESHWAR SINGH, is presently working as Group Director at Doaba Khalsa Trust Group of Institutions, SBS Nagar, Punjab, India. His current research interests include Swarm Intelligence based Optimization (Energy, Security, Routing), Wireless Sensor Network, intellectual information technology, Mobile Ad Hoc Networks.

ITEE Journal
Information Technology & Electrical Engineering
ISSN: - 2306-708X