JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

Website: http://www.jmrr.org **Vol.2, Iss.2 Pages:94-104**

SENTIMENT ANALYSIS ON COVID TWEETS

S Dhivya

Master of Computer Applications,
Coimbatore Institution of Technology, Coimbatore, Tamilnadu, India

Article Received: April 2021 Published: July 2021

Abstract

Sentiment analysis, also referred to as opinion mining, is a sub-machine learning task to determine the general sentiment of a given document. Using natural language processing, the subjective information of the document can be extracted and classify it according to its polarity such as positive, negative, and neutral. It is a really useful analysis to determine the overall opinion about a selling product, diagnose and treat a particular disease, movie review, and so on.

In this paper, tweets from Twitter are classified into "positive", "negative" or "neutral" sentiment by building a model based on probability. Twitter is a microblogging website where people can share their feelings quickly. Because of the usage of Twitter a perfect source of data to determine the current overall opinion about anything can be analyzed. This paper aims to develop a functional classifier for accurate and automatic sentiment classification of an unknown tweet stream. The tweets are classified by using four machine learning techniques like Logistic Regression, linear SVM, Random Forest classifier, and XGBoost classifier. It is concluded that the eXtreme Gradient Boosting (xgboost classifier) algorithm outperforms than other three algorithms.

Keywords: Machine learning, Natural Language Processing, Sentiment Analysis, Logistic Regression, Linear SVM, Random Forest classifier, XGBoost classifier

I. INTRODUCTION

At the end of 2019, the covid-19, ongoing coronavirus disease originated in Wuhan, China. The virus has spread and communicated locally in Wuhan and other places in china, despite strict intervention measures and efforts implemented within the region. It's affecting 203 countries and territories around the world as of 2 April 2020. Coronavirus affected 936,725 people, claimed quite 47260 lives as of two April 2020.

All countries are taking various steps to regulate the pandemic like Jana ta curfew, nation lockdown, canceling transport facilities, impose social distancing restrictions, etc. Twitter is one of the fastest information-sharing platforms among all online social networking media. Messages on Twitter range from personal information to worldwide news or events conducted throughout the world.

Sentiment analysis is well studied using Twitter data in recent days to predict and/or monitor health-related issues. Twitter contains a huge number of meaningless messages and unwanted or polluted content, which negatively affects the perception analysis performance. The normal techniques don, 't seem to be like-minded due to the short length of tweets, spelling and grammatical errors, and also the frequent use of informal languages.

In this research paper, the most aim is to get a far better understanding of the social opinions and perspectives on covid-19 and the way it's changed people's thinking over the past few months. Social media like Twitter is especially beneficial to extract information associated with the user's sentiments, opinions, and insights on numerous topics. Hence, tweet collected from twitter data for sentiment analysis of individuals on coronavirus using machine learning algorithms will help to check user's sentiments into three categories as positive, neutral, and negative during the disease outbreak.

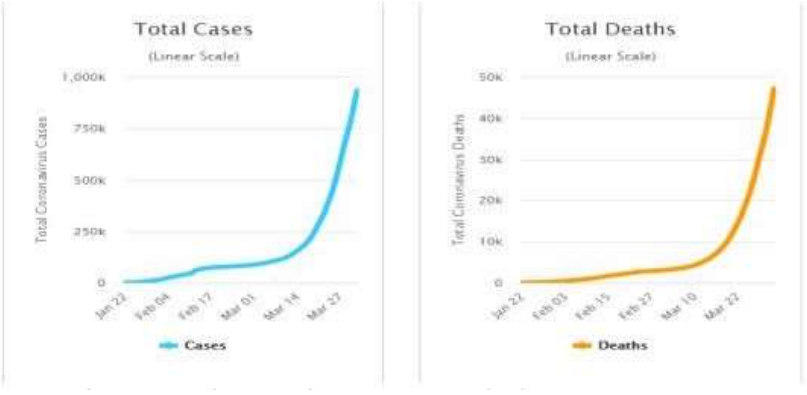

Fig 1Total coronavirus cases VS deaths data as of 2 April 2020

II. MACHINE LEARNING APPROACH

Machine learning methods are used to trained on datasets and created an evaluation model. Based on the accuracy of the model, the performance of the machine learning algorithm is acceptable. The three methods in machine learning algorithms are supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the model is trained using labeled data which contains both input and results. The two preprocessing phases are the training phase and the testing phase. Unsupervised learning methods do not use training data or labeled data. It finds the hidden structures or patterns from unlabeled data.

Supervised learning

Supervised learning requires a well-labeled dataset to train. Supervised learning is of two types namely regression and classification. Classification techniques help to find the appropriate class labels which can predict the positive, negative, and neutral sentiments. A machine learning model is developed which uses the labeled data to train, classify the tweets and predict the sentiments of the tweets. Logistic regression, random forest classifier, svm, xgbooster are the algorithms that are used in this method.

Proposed model

The dataset is collected from twitter data using Kaggle or UCI repository. Data that contain ids, tweets text, created at (date and time), and likes of the tweets related to the covid-19 pandemic and is processed through a set of five phases. The phases are tweet collection and pre-processing, tweets cleaning, feature selection, modeling, and evaluation.

Logistic regression:

Logistic regression is used to predict the outcome of a dependent variable based on previous observations. For example, an algorithm could determine the winner of an election based on previous election results. Logistic regression algorithms are popular in machine learning.

SVM:

SVM is a kind of non-binary linear classifier. It is trained with a series of data. Data is already classified into two categories, building the model, and is initially trained. The task of a SVM algorithm is to determine which category a new data entry will belong to.

Some applications of SVM include:

1. Text and hypertext classification
2. Image classification
3. Recognizing handwritten characters
4. Biological sciences, including protein classification

Random Forest Classifier:

Random forests or random decision forests are a learning method for classification, regression, and other tasks it operates by constructing multiple decision trees at training time, and output is the mode/mean/median prediction of the individual class.

XGBooster:

XGBoost is an algorithm that has recently more used machine learning algorithm for structured or tabular data. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

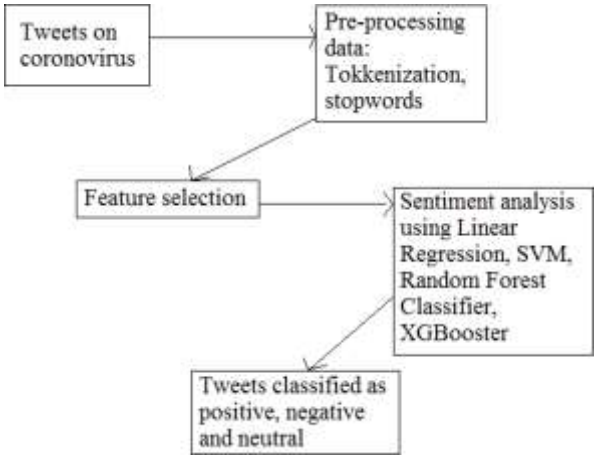

Fig 2 Corona sentiment analysis model

Phase i: Tweets collection and pre-processing

Tweets collection:

Download twitter dataset ('covidke tweets.csv') import the dataset.

Table 1 covidketweets.csv dataset

```
In [5]: #to read csv file(dataset)
        tweets = pd.read_csv('covidKE tweets.csv')
        #To view full dataset
        print(tweets)
```

```
          tweet_id                                               text  \
0     1.240000e+18  Everyday might not be good but there's always ...
1     1.240000e+18          The next one week 😁\n\n #coronaviruskenya
2     1.240000e+18  #coronaviruskenya #LockdownNow #UhuruKenyatta ...
3     1.240000e+18  Internet never forget we will remind you that ...
4     1.240000e+18  @ntsa_kenya @DCI_Kenya #coronaviruskenya If on...
...            ...                                                ...
1185  1.240000e+18  SIAYA EMERGENCY response team sent to trace co...
1186  1.240000e+18  @OleItumbi Time to put it mandatory😡😡I hate wh...
1187  1.240000e+18  When people lose trust in institutions, govern...
1188  1.240000e+18  @_mwendwa_felix @CisNyakundi @RobertAlai Level...
1189  1.240000e+18  The next Corona virus strain will be even more...

           created_at  likes
0     3/23/2020 16:17      1
1     3/23/2020 16:17      0
2     3/23/2020 16:17      0
3     3/23/2020 13:50      0
4     3/23/2020 16:17      0
...               ...    ...
1185  3/23/2020 12:22      4
1186  3/23/2020 12:21      1
1187  3/23/2020 12:21      1
1188  3/23/2020 12:21      2
1189  3/23/2020 12:20      1

[1190 rows x 4 columns]
```

The process of transforming unstructured data into structured data is the pre-processing stage. The data is downloaded from twitter directly through Twitter API. Hashtags, whitespaces, hyperlinks, URLs, usernames, stop words, etc are removed from tweets.

	tweet_id	text	created_at	likes	text lengthb
0	1.240000e+18	Everyday might not be good but there's always ...	3/23/2020 16:17	1	126
1	1.240000e+18	The next one week 😁\n\n #coronaviruskenya	3/23/2020 16:17	0	39
2	1.240000e+18	#coronaviruskenya #LockdownNow #UhuruKenyatta ...	3/23/2020 16:17	0	140
3	1.240000e+18	Internet never forget we will remind you that ...	3/23/2020 13:50	0	104
4	1.240000e+18	@ntsa_kenya @DCI_Kenya #coronaviruskenya If on...	3/23/2020 16:17	0	131

Fig 3 Before pre-processing stage

Table 2 Pre-processing stage

Remove hashtags	Remove whitespaces	Remove hyperlinks
Remove URL address	Remove HTML special entities	Remove usernames
Removal of stop words	Remove tickers	Remove Unicode strings from the tweets

	tweet_id	text	created_at	likes	text lengthb	text length
0	1.240000e+18	everyday might good always something good day ...	3/23/2020 16:17	1	126	86
1	1.240000e+18	next one week coronaviruskenya	3/23/2020 16:17	0	39	30
2	1.240000e+18	coronaviruskenya lockdownnow uhurukenyatta ima...	3/23/2020 16:17	0	140	108
3	1.240000e+18	internet never forget remind resorted internet...	3/23/2020 13:50	0	104	76
4	1.240000e+18	ntsa kenya dci kenya coronaviruskenya provide ...	3/23/2020 16:17	0	131	94

Fig 4 After pre-processing stage

Fig 4 is the output of data cleaning after removing hashtags, whitespaces, smileys, hyperlinks, stopwords.

Phase ii: Tweets cleaning

Tweets in their original form cannot be processed for analysis. Tokenization refers to splitting strings into smaller words called tokens. Tokenization consists of identifying nouns, verbs, adverbs, and adjectives, etc. Grouping of words with the same meaning is one of the processes under NLP. The cleaned tweets are now subjected to text processing where tokenization and stemming are done to the tweets.

	tweet_id	text	created_at	likes
0	1.240000e+18	[everyday, might, good, always, something, goo...	3/23/2020 16:17	1
1	1.240000e+18	[next, one, week, coronaviruskenya]	3/23/2020 16:17	0
2	1.240000e+18	[coronaviruskenya, lockdownnow, uhurukenyatta,...	3/23/2020 16:17	0
3	1.240000e+18	[internet, never, forget, remind, resorted, in...	3/23/2020 13:50	0
4	1.240000e+18	[ntsa, kenya, dci, kenya, coronaviruskenya, pr...	3/23/2020 16:17	0

Fig 5 Tokenization stage

Phase iii: building tweet dictionary & determining word density

After the text processing, the word dictionary is built. All words are categorized into positive, negative, and neutral across all datasets. Later the density of each word is calculated as the count of occurrences of every unique word across all the training datasets. The sentiments are

usually not affected by stop words. Hence stop words are filtered and words like not and do not say about the polarity of the tweet. This dictionary helps in evaluating the testing set.

After collecting a bag of words (bow). Collected words are shown in figure 5.

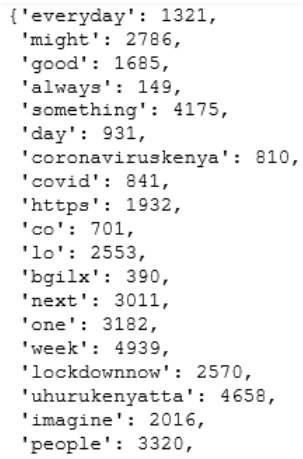

Fig 6 Bag of Words collected from Tweets

Weight for each in bow is calculated and is shown in figure 7.

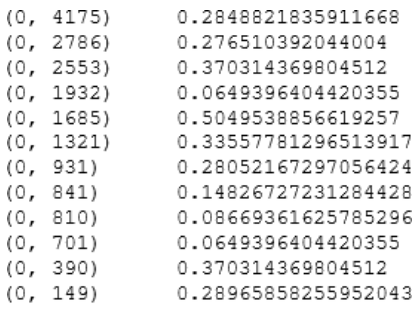

Fig 7 Calculating weight

Phase iv: Feature selection

Feature selection is based on word density. Word density is determined by Term Frequency Inverse Document Frequency (TF-IDF) method. TF-IDF identifies the frequently occurring words in the given tweet and the words that are not appearing frequently in the remaining training data set. Word density helps to know about the polarity of the tweet. The polarity of

the tweet is calculated through a term weight using TF-IDF. The positivity and negativity of the term are calculated based on the number of times the term occurs in a given tweet dataset. The TF-IDF is executed on all terms in the dataset to find the rank of each word. A high rank in TF-IDF shows the word is relevant in the given tweet and can contribute much to the polarity of the tweet. The overall approach works as follows:

Given a data set d, a term t, and an individual tweet (dt), dt ε d, we calculate:

Adsjusted tf-idf = ft, dt* log (|d| /ft+s, d)

Where ft, ds equals the no. Of times term appear in dataset.

|d| is the size of the data set, and f t+s, d equals the number of tweets in which the term t and its corresponding synonym word appears in d.

For each term in the term data, the corresponding synonyms are fetched from the word dictionary. Synonyms are considered to be the words equivalent to the original term and hence taken for the calculation of f t, d. Set of terms that are extracted using adjusted TF-IDF are used to judge the polarity of the tweet.

The polarity for each tweets is given in figure 7,

	tweet_id	text	created_at	likes	text lengthb	text length	polarity
882	1.240000e+18	happening usa apparently homeless important sa...	3/23/2020 13:50	3	140	87	-0.016667
647	1.240000e+18	melee tuskysofficial supermarket reason sneeze...	3/23/2020 14:29	2	118	88	0.000000
568	1.240000e+18	situation busia county general hospital medics...	3/23/2020 14:38	5	140	113	0.050000
1008	1.240000e+18	mbagathi full bed whole area wil messy contani...	3/23/2020 13:16	0	140	105	0.116667
656	1.240000e+18	burn covid coronaviruskenya mulikacorona coron...	3/23/2020 14:28	0	138	112	0.000000
1060	1.240000e+18	sorry say africa could soon worst hit covid pa...	3/23/2020 13:02	0	140	98	-0.750000
246	1.240000e+18	chat tom hickson director world dodgeball asso...	3/23/2020 15:33	1	140	111	-0.050000
16	1.240000e+18	together let safe stop spread covid coronaviru...	3/23/2020 16:16	1	126	84	0.500000
219	1.240000e+18	search ongoing individuals came contact rome b...	3/23/2020 12:34	0	140	103	0.000000
370	1.240000e+18	coronaviruskenya uhurukenyatta staysafestayhom...	3/23/2020 15:09	1	139	116	0.000000

Fig 8 polarity calculation

When polarity=0, it considers as “neutral” tweets, polarity >0 is “positive” tweets, polarity <0 is “negative” tweets it shown in the below figure 8,

	tweet_id	text	created_at	likes	text lengthb	text length	polarity	sentiment
0	1.240000e+18	everyday might good always something good day ...	3/23/2020 16:17	1	126	86	0.4000	positive
1	1.240000e+18	next one week coronaviruskenya	3/23/2020 16:17	0	39	30	0.0000	neutral
2	1.240000e+18	coronaviruskenya lockdownnow uhurukenyatta ima...	3/23/2020 16:17	0	140	108	0.0000	neutral
3	1.240000e+18	internet never forget remind resorted internet...	3/23/2020 13:50	0	104	76	0.0000	neutral
4	1.240000e+18	ntsa kenya dci kenya coronaviruskenya provide ...	3/23/2020 16:17	0	131	94	0.0000	neutral
5	1.240000e+18	like hassanalijoho said coronaviruskenya total...	3/23/2020 16:17	0	74	55	0.0000	neutral
6	1.240000e+18	us bite finger nails since childhood days happ...	3/23/2020 15:50	0	136	88	0.0000	neutral
7	1.240000e+18	need testing testing testing testing testing c...	3/23/2020 16:17	0	140	104	0.0000	neutral
8	1.240000e+18	wish would much anticipated presidential addre...	3/23/2020 12:44	2	140	90	0.1000	positive
9	1.240000e+18	mass testing please please fins way coronaviru...	3/23/2020 14:44	0	63	52	0.0000	neutral
10	1.240000e+18	need help mean ukipata corona nkama uko na aid...	3/23/2020 14:44	1	98	69	-0.3125	negative

Fig 9 Polarity classification as positive, negative, and neutral

Phase v: Model building & evaluation

The dataset has two parts where first the part is used to build the model and the second part is used to test the accuracy of the model. The four classification algorithm (Logistic Regression, SVM, Random Forest Classifier, XGBooster) is used to build the model. The experimental analysis reflects the sentiments of people towards coronavirus.

Initializing each model:

Logistic regression is calculated, with the features c, dual, max_iter. C is a positive floating-point number (1.0 by default) that defines the relative strength of regularization. Smaller values indicate stronger regularization. Max_iter is an integer (100 by default) that defines the maximum number of iterations by the solver during model fitting. Dual is a boolean (false by default) that decides whether to use primal (when false) or dual formulation (when true).

Logisticregression(c=4.0,class_weight=none, dual=false, max_iter=600)

SVM is also calculated with the features c, dual, max_iter.

Linearsvm(c=4.0, class_weight=none, dual=false, max_iter=600)

Random forest classifier is calculated with the features bootstrap, criterion, min_samples_split, n_estimators. N_estimators: the number of trees in the forest. Criterion: the function to measure the quality of a split. Supported criteria are “gini” for the gini impurity and “entropy” for the information gain. Note: this parameter is tree-specific. Min_samples_split: the minimum number of samples required to split an internal node.

Randomforestclassifier(bootstrap=false,

class_weight=none, criterion='entropy', min_samples_split=6, n_estimators=100, warm_start=false)

XGBooster is calculated with the objective and num-class of the features.

Xgbclassifier(objective='multi:softmax', num_class=3)

Train the models:

Fit the training dataset to each model to calculate the accuracy by using the cross_val_score function. Cross- validation is a statistical method used to estimate the skill of machine learning models. That k-fold cross- validation is a procedure used to estimate the skill of the model on new data.

```
cross_val_score(lr, features, tweets['target'], cv=95).mean())
cross_val_score(svc, features, tweets['target'], cv=95).mean())
cross_val_score(rfc, features, tweets['target'], cv=95).mean())
cross_val_score(xgb, features, tweets['target'], cv=95).mean())
```

Fig10 cross_val_score for each model

The model accuracy is showed below:

Fig 11 Accuracy of each model

From figure 11, it is concluded that the extreme gradient boosting (XGBoost classifier) algorithm outperforms than other three algorithms.

III. CONCLUSION

This work was carried on the coronavirus outbreak using twitter data from 1st April 2020 to 5th April 2020 where the virus spread across several countries and the outbreak became pandemic. This work helps to understand the people's perception about coronavirus and its impact on the public. The sentiments during the period were downloaded and the public's reaction towards the outbreak was analyzed. The machine learning algorithm is applied for data analysis and the accuracy of the model is nearly 70%. The people well understood the

government policies, safety measures, symptoms and precautionary measures to be taken during this period. They well followed and maintained the social distancing and sanitizing methods. This study helps the organizations to understand the opinion of people during the corona virus outbreak. As the virus is spreading vigorously, the study needs to be carried out every week to have a better understanding of the sentiments of the people.

IV. ACKNOWLEDGMENT

I would like to express my special thanks of gratitude to my professor (DR.J.B.Jona) as who gave me the golden opportunity to do this wonderful paper on the topic (Sentiment Analysis on Covid Tweets), which also helped me in doing a lot of research and I came to know about so many new things I am thankful to them. Secondly, i would also like to thank my parents and friends who helped me a lot in finalizing this paper within the limited time frame.

REFERENCES

1. Zhang, l., Ghosh, R., Dekhil, M., Hsu, M., & Liu, B. (2015). Combining lexicon-based and learning-based methods for twitter sentiment analysis. International journal of electronics, communication and soft computing science & engineering (ijecscse), 89, 1–8.
2. U. A. Siddiqua, T. Ahsan, and A. N. Chy, "combining a rule-based classifier with the ensemble of feature sets and machine learning techniques for sentiment analysis on the microblog," in 2016 19th international conference on computer and information technology (iccit), 2016, pp. 304– 309.
3. Johnnatan messias, Joao P. Diniz, Elias Soares, Miller Ferreira, Matheusaraujo, Jucas Bastos, Manoel Miranda, Fabricio benevenuto, towards sentiment analysis for mobile devices . 2016
4. Nels Oscar, Pamela A. Fox, Racheal Croucher, Riana Wernick, Jessica Keune, and Karen Hooker. Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on twitter. J Gerontol B Psycholscisocsci, 2017, vol. 72, no. 5, 742–751
5. Minchae Song, Hyunjung Park, Kyung-Shik Shin" attention-based long short-term memory network using sentiment lexicon embedding for aspect-level sentiment analysis in Korean." Information processing & management, 56 (3) (2019), pp. 637- 653
6. Ajay Aandi and Aziz fellah. Socio-analyzer: a sentiment analysis using social media data. Volume 64, 2019, pages 61–67 proceedings of 28th international conference on software engineering and data engineering
7. Ali Aasan, Sana Moin, Ahmad Aarim and Shahaboddinshamshirband "machine learning-based sentiment analysis for twitter accounts", Mdpi, 2018.

Cite this article:

S Dhivya, "Sentiment Analysis on Covid Tweets", Journal of Multidimensional Research and Review (JMRR), Vol.2, Iss.2, pp.94-104, 2021.