ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 6| June 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

SENTIMENT ANALYSIS OF PEOPLE DURING COVID- 19 USING SVM AND LOGISTIC REGRESSION ANALYSIS

Sayan Majumder[1], Anuran Aich[2], Satrajit Das[3]
[1,2,3]*Department of Computer Science & Engineering, Gargi Memorial Institute of Technology, Baruipur, Kolkata*

Article DOI: https://doi.org/10.36713/epra10424
DOI No: 10.36713/epra10424

ABSTRACT

Starting from China, novel coronavirus is now the most dangerous threat to humans, all over the world. India is also not an exception. Besides using mask and sanitizers, Indian Government has also decided to maintain proper social distancing or lockdown. Our sentiment during this lockdown period also varied from man to man. In this paper, we have collected twitter data of people across India, during March to June and then using NLP, the polarity is measured, i.e. positive, negative or neutral. After that. We have used SVM classifier and Logistic Regression analysis to classify the sentiments of people. Python programing language is used in Anaconda distribution to simulate the result. After simulation, we have found that SVM gives 91.50% accuracy, whereas Logistic Regression gives 87.75% accuracy. At last, a comparative study between these two results, is also represented.

KEYWORDS: ***Sentiment Analysis, NLP, Machine Learning, SVM, Regression.***

I. INTRODUCTION

The Coronaviruses which eventually belongs to a large class of viruses that cause illness ranging from common cold to more severe diseases, among them novel coronavirus has apparently new characteristics that have not been seen in humans earlier. The COVID-19 [1] pandemic is the most affecting and dangerous one in the list of biggest known pandemics in the world history. COVID-19 causes severe respiratory illness and thus leads to deaths of several living beings. The COVID-19 pandemic started in India on 30th January 2020, as on that day first case was informed; this initiated from China and spread all across the world very quickly. In India most cases were reported mainly in six cities from Delhi, Mumbai, Ahmedabad, Chennai, and Pune to Kolkata. On 22nd March 2020, India observed a 14-hour public curfew. Further, on 24th March, the directive for a 21 days nationwide complete lockdown was given by our prime minister and as situations didn't improve again the lockdown was extended. Nowadays we can exchange our valuable information through robust social medium like, Facebook, Twitter etc. People express their sentiments through various posts. So, such information can be used in an organized way, for studying about human society and many other things.

We chose Twitter for its popularity among social networking sites which can be implemented for analysis and understanding the feeling of people across India regarding the lockdown. We have collected the data from tweets of various people across India which is collected using the twitter API. The ease of the tweets may vary from public statements to personal thoughts. Table 1 demonstrates tweets collected from twitter.

Sentiment Analysis is done as it is important and important to know about the feelings of so many people across India and helps to find, whether people see this national lockdown positively as a measure to fight against the pandemic or they have negative feelings or neutral thoughts on this.

Classification is the most helpful process to segregate tweets into positive, neutral and negative. Here positive means how much people are positively taking the lockdown as an effective measure to fight against COVID-19, negative defines concerns of various people and neutral means those people who have neutral thoughts on this matter.

Amidst various classification processes, we have selected Support Vector Machine Classifier and Logistic Regression [2] Classifier. There were many studies regarding the analysis of human sentiment. Bac Le and Huy Nguyen [3] performed his research on twitter data only. Recently emotional analysis was done by Amrita Mathur, Purnima Kubde and Sonali Vaidya [4] using Twitter Data during COVID-19 to understand the mental health of people. Jiang et. Al. [5] did a Verb Expression attitude analysis and proposed that Logistic Regression gave excellent results.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 6| June 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

Tweets:-

@PMOIndia Just a thought can't we use workforce of @swiggy_in @Zomato @amazonIN @bigbasket_com etc. to give ration to needy people in India. As they already have people on ground

@Republic_Bharat Arnabh Bhai, is India left with only one option i.e. lockdown. Continuing lockdown is not a solution to this Covid-19. People like me left with no rashion/money anymore. I will opt to die better than begging before government.

The Government of India has taken a tough decision to extend the nationwide #lockdown for two weeks. We must support the Government as fighting #Covid_19 is need of the hour. #lockdown3 #Lockdownextention #coronahaaregaindiajeetega @pmoindia @hmoindia @cmoguj @reliancejio

RT @hubermantamir: Work From Home (WFH) may become the new normal for startups as the world is witnessing an extended lockdown due to the CÃ¢â‚¬Â¦

II PROBLEM SOLVING METHODOLOGY

Data Preperation

We chose Twitter, being one of the most favored social platforms to fetch tweets and prepare a dataset. Fetching data required API access, so we submitted an application for request for API access and after five to six days of verification, twitter provided their access token and other keys to access tweets which were gathered by writing a python script using a python library called Tweepy, which uses API class, usually used to provide access to the entire RESTful API methods. Each method can accept various parameters and return responses; so we gathered set of tweets of people across India and thus prepared a dataset.

Data Preprocessing

Data cleaning is an essential task, for which few python libraries like string and regular expressions were used and again a python function was written and applied on the text data that converted the text into lower case, and then unwanted special characters, Twitter user names (e.g. @Satish), Twitter particular words (like "RT" which means an abbreviation of retweet, which means reposting of a message), punctuations, hyperlinks etc. were removed. After that we checked whether the dataset contained any null or missing values and treated them, we also checked whether the set of data contained any duplicate values using panda's library and treated them, thus obtained a cleaned dataset to work upon. We have used the Label Encoding technique (that converts the labels into numeric form as the machine learning algorithms can decide in a better way on how those labels must be operated) to convert the neutral, negative, positive tweets into 0,1,2 which is an important step to preprocess the data.

Feature Extraction

The dataset which is collected is therefore used to draw out attributes that will be very needful for training of our classifier, so we used Text Blob, which is an extremely powerful NLP [6] [7] library for python constructed on the base of NLTK (Natural Language Toolkit).TextBlob has several functionalities, firstly it performs Tokenization that indicates to slice large text into set of words, where token typically refers to a single term, then lemmatization is performed that is reducing an expression. Then it finds N-Grams (n no. of amalgamation of words) which is important for finding where a text is positive, neutral or negative. Then using Text Blob, we found out the polarity and expression of opinion of the tweets, so polarization can be varied between -1 to 1. Tweets with negative polarity have negative point of view about the lockdown while the positive one has positive point of view. Then we analyzed the tweets and write a method to classify the tweets to segregate them into positive, neutral and negative using the polarity value.

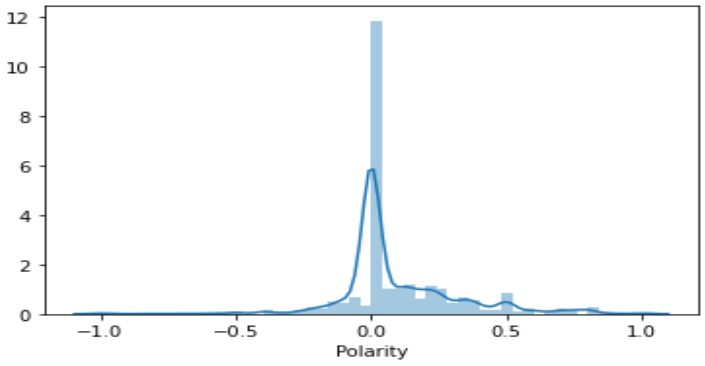

Fig. 1 polarity extracted from tweets

It is proved from the picture above that nearly all of the tweets are positively polarized within 0 and 0.5; this resembles how much positively people of India reacted to the lockdown to stop the spread of COVID19.

Classifier

We used SVM [8] and Logistic Regression [9] classifiers to separate the 3 different emotions from the tweets and made a comparative study to find whose accuracy is more or which algorithm yields the best results.

Support Vector Machine

SVMs belong to the class of widely applicable linear classification. An exclusive property of SVM is that it can concurrently minimize the classification error and also has the capability of maximizing the geometric margin. So SVM [10] is known as Maximum Margin Classifier [11]. SVM has the capability of mapping input vectors to space which belongs to greater dimension such that a supreme splitting hyperplane is

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 6| June 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

created. So, further on each side of the subspace another two parallel subspaces are built that helps in segregating the data. Here we assume that, the maximal is the gap between these parallel subspaces.

Step 1: - We have taken the data points in the form of-

$$\{(a_1, b_1), (a_2, b_2),(a_3,b_3),(a_4,b_4)\ldots\ldots\ldots,(a_m, b_m)\}.$$

Where b_m=1 / -1, it is a constant that signifies to the class, point a_m belongs to.
m = test number. Each a_m being real vector with dimension p.

Step 2: - From the data used to train we observe that, we divide (or separate) the hyperplane, which takes

$$z.\ x + c = o \text{ ----- (1)}$$

Here z is p-dimensional Vector and c is scalar. The vector z points erect towards the separating hyperplane. By adding the parameter c, we can make a greater in size of the margin. If c is absent then the subspace is forced to go through the centre, thus limiting the result. Parallel hyperplanes are narrated by implementing the equation

$$z.x + c = 1$$

$$z.x + c = -1$$

The data which is used to train can be linearly distinguishable, so we select those hyperplanes, thus no points are present within those and then we can make a try to inflate their farness. With the help of geometry, we found out the distance between the subspaces are $2 / |w|$. So $|z|$ should be minimized.

$$z.\ x_i - c \geq 1 \text{ or } z.\ x_i - c \leq -1$$

Step3: - Again this may be expressed as,

$$y_i\ (z.\ x_i - c) \geq 1,\ 1 \leq i \leq m \text{ ------(2)}$$

A hyperplane which differentiates the substantial margin is expressed as $M = 2 / |w|$, so it specifies that support vectors meet those data points used for training close to it.

The vectors xi which are responsible for training, paint into a greater proportional space by the relation Φ. Then a linear distinguishable subspace with the greatest difference is found by the SVM. Here C > 0 is interpreted as the penalty boundary of the faulty term.

Logistic Regression

To obtain the possibility of class existing, such as success/failure or win/lose etc. we can use logistic regression technique. In our sentiment analysis, we obtained the feelings of Indian people from their tweets regarding the lockdown ordered by Prime Minister of India.

Step 1: - A model is considered with two x_1, x_2 as predictors and one variable J as binary response, by which we indicate p=P(J=1).

Step 2: - A linear interrelation is presumed as-

$$l = \log_b (p/1-p) = \beta_0+\beta_1.x_1+\beta_2.x_2$$

Step 3: - $((p)/(1-p)) = b^{\beta 0+\beta 1.x1+\beta 2.x2}$

By using concept of algebra, the probability that J=1 is

$$P = (b^{\beta 0+\beta 1.x1+\beta 2.x2})/ (b^{\beta 0+\beta 1.x1+\beta 2.x2}+1) = 1/ (1+ b^{-(\beta 0+\beta 1.x1+\beta 2.x2)})$$

From the formula above we observe that when β_i is constant.

III SIMULATION AND RESULT

We have used Jupyter Notebook and Python for simulation and thus describe the results, here. Combination of both gives us the right environment to perform the Sentiment Analysis of People during Lockdown Period.

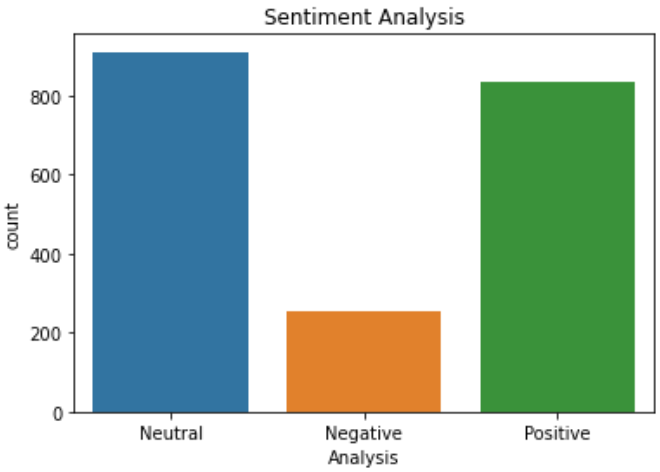

Fig. 2 Count of tweets containing neutral, positive and negative sentiments.

From this we found the percentage of positive tweets is 41.8%, negative tweets are 12.7% and rest is neutral.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 6| June 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

RESULTS OF SVM MODEL

```
#SEE THE PERFORMANCE OF THE SVC

from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	1.00	0.66	0.79	44
1	0.91	0.97	0.94	200
2	0.96	0.97	0.96	156
accuracy			0.94	400
macro avg	0.96	0.87	0.90	400
weighted avg	0.94	0.94	0.93	400

Fig. 3 Performance of Support Vector Machine Classifier

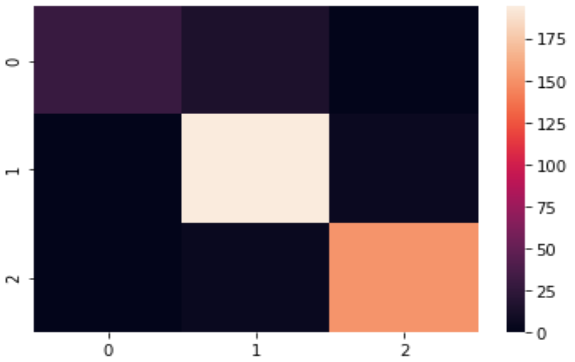

Fig. 4 Confusion Matrix of SVM classifier

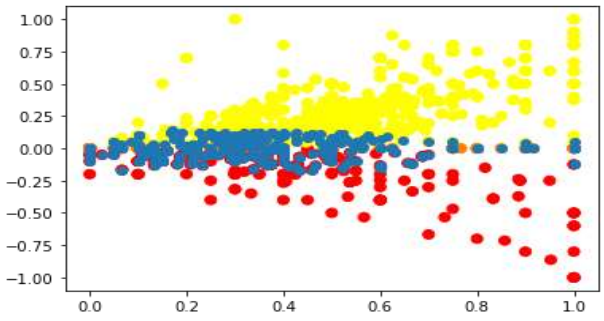

Fig. 5 Plotting the support vectors

RESULTS OF LOGISTIC REGRESSION MODEL

```
from sklearn.metrics import classification_report, confusion_matrix
print(classification_report(y_test,predictions))
```

	precision	recall	f1-score	support
0	1.00	0.52	0.69	44
1	0.87	0.92	0.89	200
2	0.90	0.95	0.92	156
accuracy			0.89	400
macro avg	0.92	0.80	0.83	400
weighted avg	0.89	0.89	0.88	400

Fig. 6 Shows the performance of Logistic Regression

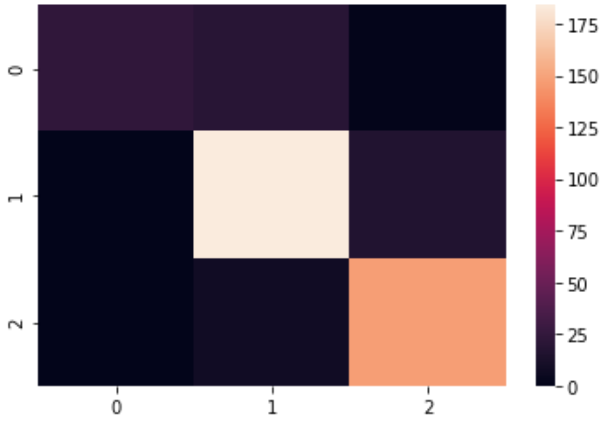

Fig. 7 Confusion Matrix of Logistic Regression

Accuracy Comparison

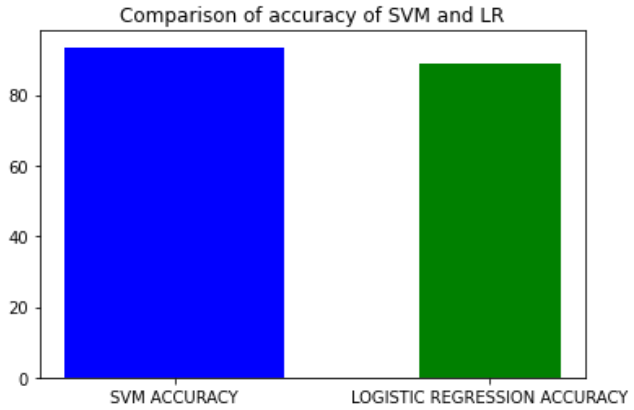

Fig. 8 Comparison between the accuracy of the two models.

ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 6| June 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

Table 2. Comparison between the two algorithms used

Parameters	SVM	Logistic Regression
Simulation Time	260 seconds	200 seconds
Simulation Area	255 × 255 pixels	255 × 255 pixels
Accuracy	91.5 %	87.75%

IV. CONCLUSION

We have done here a comparative study between SVM and Logistic Regression for performing sentiment analysis by classifying positive, negative and neutral sentiments.

Both the algorithms have given close accuracy but SVM performed slightly better.

But from the observations we can conclude people across India have positive sentiments regarding the lockdown ordered by Indian Prime Minister as a protective measure to stop the spread of COVID19 and very few people had negative sentiments regarding the lockdown while others had neutral sentiments which mean they had mixed feelings.

Three main advantages of SVM over Logistic Regression are:-

1. SVM tries to find the "best" margin that separates the classes and this reduces the risk of error on data, while logistic regression does not.
2. 2. The risk of overfitting is less in SVM, while Logistic regression is vulnerable to overfitting.
3. SVM gives 91.50% accuracy whereas Logistic Regression gives 87.75% accuracy.

REFERENCES

1. *V. Chamola, V. Hassija, V. Gupta and M. Guizani, "A Comprehensive Review of the COVID-19 Pandemic and the Role of IoT, Drones, AI, Blockchain, and 5G in Managing its Impact," in* **IEEE Access**, *vol. 8, pp. 90225-90265, 2020.*
2. *Majumder S., Bhattacharyya D. (2020) Relation Estimation of Packets Dropped by Wormhole Attack to Packets Sent Using Regression Analysis. In: Mandal J., Bhattacharya D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore*
3. *Le B., Nguyen H. (2015) Twitter Sentiment Analysis Using Machine Learning Techniques. In: Le Thi H., Nguyen N., Do T. (eds) Advanced Computational Methods for Knowledge Engineering. Advances in Intelligent Systems and Computing, vol 358. Springer, Cham*
4. *A. Mathur, P. Kubde and S. Vaidya, "Emotional Analysis using Twitter Data during Pandemic Situation: COVID-19,"* 2020 5th International Conference on Communication and Electronics Systems (ICCES), *COIMBATORE, India, 2020, pp. 845-848, doi: 10.1109/ICCES48766.2020.9138079.*
5. *Jiang D., Tao Q., Wang Z., Dong L. (2019) An Intelligent Logistic Regression Approach for Verb Expression's Sentiment Analysis. In: Patnaik S., Jain V. (eds) Recent Developments in Intelligent Computing, Communication and Devices. Advances in Intelligent Systems and Computing, vol 752. Springer, Singapore*
6. *H. Jelodar, Y. Wang, R. Orji and H. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," in* IEEE Journal of Biomedical and Health Informatics, *doi: 10.1109/JBHI.2020.3001216.*
7. *A. Gupta, A. Singh, I. Pandita and H. Parashar, "Sentiment Analysis of Twitter Posts using Machine Learning Algorithms,"* 2019 6th International Conference on Computing for Sustainable Global Development (INDIACom), *New Delhi, India, 2019, pp. 980-983.*
8. *A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis,"* 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), *Bangalore, 2016, pp. 628-632, doi: 10.1109/ICATCCT.2016.7912076.*
9. *H. Hasanli and S. Rustamov, "Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM,"* 2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT), *Baku, Azerbaijan, 2019, pp. 1-7, doi: 10.1109/AICT47866.2019.8981793.*
10. *Zainuddin, Nurulhuda & Selamat, Ali. (2014). Sentiment analysis using Support Vector Machine. I4CT 2014 - 1st International Conference on Computer, Communications, and Control Technology, Proceedings. 333-337. 10.1109/I4CT.2014.6914200. Author, F.: Article title. Journal 2(5), 99–110 (2016).*
11. *Sahu, S.K., Behera, P., Mohapatra, D.P. et al. Sentiment analysis for Odia language using supervised classifier: an information retrieval in Indian language initiative. CSIT* ***4,*** *111–115 (2016) https://doi.org/10.1007/s40012-016-0117-9*