Analysing Tweets of Covid 19 Vaccination Drive in India

Annie Ann Abraham[1a], Diya Susan Eapen[2a], Jo Cheriyan[3a], Lekshmi S Nair[4b]

[a]*Department of Computer Science and Engineering, SAINTGITS College of Engineering, Kerala, India*

[b]*Department of Computer Science and Engineering, AMRITA Vishwa Vidyapeetham, Amritapuri, Kerala, India*

[1]*annieannabraham@gmail.com,* [2]*diyaseapen@gmail.com,* [3]*jo.cheriyan@saintgits.org*

[4]*lekshmisn@am.amrita.edu*

Abstract

Social media marks a vital role in social influence nowadays. India ranked second highest usage of social media, Twitter, and tweets influence the social community to a great extent. The analysis of tweets has been an exciting topic for data analysts to consider. Analysing the tweets relating to COVID-19 vaccination is such a topic. In this article, we focus on analysing public opinion and perceptions toward vaccination drive engaged in India. By applying machine learning algorithms, we analyse and conclude the public opinion of the COVID 19 vaccination drive.

Keywords. Covid-19, Machine learning, Sentiment analysis, Twitter

1. INTRODUCTION

The COVID-19 pandemic evolved in last 2019 and spread around the world. The first case of such pandemic was reported in INDIA in March 2020. The World Health Organisation(WHO) invoked the global pharmaceutical industries to develop vaccines. The Covid-19 vaccines focus on preventing symptomatic and severe illness. Many technology platforms are involved in the research and development to create an effective vaccine against COVID-19. India administered the first vaccination drive on January 16, 2021. Currently, India approved the vaccines developed under several pharmaceuticals with trade names like Covishield, Covaxin, Sputnik V, Moderna, and ZyCoV-D.

The present COVID-19 pandemic has resulted in an upsurge in the use of social media as a platform for debating numerous pandemic themes, including vaccinations [1]. Indeed, negative feelings and disinformation may be spread through social media, impacting individual perspectives and resulting to vaccination rejection [2]. Hesitation towards vaccines causes a threat to health. Misinformation paves the way towards lower vaccination uptake. This article evaluates the public opinion among the Indians towards the vaccination

drive over the social network platform. We took tweets related to the vaccination drive in INDIA and analyzed the approaches, like and dislike towards vaccination.

The remaining part of the article is structured as follows: Section II gives details about the literature for the proposed work. The proposed work is detailed in section III. The experimentation and their validation are described in section IV followed by results and their discussions in section V. Finally, the article is concluded in section VI.

2. BACKGROUND STUDY

The Sentiment analysis is a powerful way for expressing and labelling the sentiments showed by crowd or community from the text source [3]. Usually, people take up social media like Facebook or Twitter to express their sentiments regarding a topic. These emotions vary from solid likes or dislike towards a product or a policy taken by a community. Expressing emotions against any such event on a social platform would generate considerable information and is impractical to process manually. Machine learning algorithms are adapted to process and analyse such information. For analysing such extensive data, the method of opinion mining helps in this scenario [4]. In [5], [6], the authors use supervised KNN for the analysis of tweets. The Sentiment analysis (or opinion mining) evaluates whether data is positive, negative, or neutral. Sentiment analyses helped to improve in domains like business strategies, evaluating customer feedback, customer needs [7], financial time series forecasting [8].

Capturing tweets for sentiment analysis directly from social media platforms helps access the insights of social mentions at par with time [9]. Opinion mining combines computational linguistics and NLP to extract sentiments(positive, negative, or neutral). It helps to understand customers' likes and dislikes and redesign the product or services. Opinion mining can be performed on structured or unstructured texts using appropriate natural language processing (NLP) [10].

Several works are done as part of social network analysis. Twitter dataset analyzed as part of the COVID-19 pandemic in various aspects. Analysing the sentiment towards the vaccination is one such example. Twitter has become an essential platform for gathering public opinion widely [11]. Tweets help understand people's feelings about a situation from a social network platform. Analysis of these data helps understand the people's opinion and perseverance about various topics. In [12] author performed text mining to identify addiction concerns during the COVID-19 pandemic. In [13], the authors performed an analysis towards the topic of usage of "masks" through Tweets. The volume of tweets related to masks increased by March 2020.

In the proposed work, we collected people's sentiments as tweets against the COVID-19 vaccination drive in India. Using Machine Learning algorithm we perform sentiment analysis on tweets. We use Naive Baye's and Logistic Regression to perform sentiment analysis and have derived the exact public opinion prevailing in India towards the vaccination process taking place.

3. PROPOSED WORK

3.1 Sentiment Analysis

3.1.1 Pre-processing

The tweets specifically about the vaccination drive in India are downloaded. The tweets may contain emojis, acronyms, or even the rating of experience. We clean up the entire dataset by considering the following:

- Remove hashtags, mentions, and links
- Punctuations removal (including alphanumeric characters if necessary)
- Tokenization
- Stop words removal

Natural language processing (NLP) is gaining popularity in developing applications like chatbots, language translations, data analysis. Advancement in Machine learning and Deep learning helped consider more similar data to linguistic forms. Several NLP libraries like Textblob, Spacy, Gensim, CoreNLP are used in text processing. TextBlob generates part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, language translation more easily. Given the sentiments, Textblob returns polarity and subjectivity. Polarity is a float between -1 and 1, where 1 suggests that it is a positive statement and -1 means a negative statement. Subjective sentences express personal feelings, emotions, or judgments, whereas objective sentences express facts. Subjectivity also lies in the range of [0,1]. Due to these advantages, we are using Textblob for classifying our tweets.

3.1.2 Naïve Bayes

Naive Bayes is a supervised classification algorithm that is suitable for both binary and multi-class classification [14]. Naive Bayes is heavily used in text classification. The Naive Bayes assumes that each characteristic contributes equally towards the outcome. For using the Naive Bayes classifier, we look into the following:

- Feature matrix contains all the rows in the dataset. In our dataset, the feature is 'Tweet.'
- The response vector is the prediction value for each row of the feature matrix.

Naive Bayes Classifier combines the features to detect weights using probability. We classified the tweets obtained from class, namely sentiment class. This classification of sentiment obtained positive and negative. Unigram Naive Bayes approach is used for tweet classification, and it is observed that Naive Bayes delivers better results than Support Vector Machine when the small dataset size. The Naive Bayes algorithm exhibited data consistency and measurement classification.

3.1.3 Logistic Regression

Logistic regression is suitable to connect one or more independent variables to the dependent variable of the type of category. The relationship between the dependent variable and one or more independent variables is established by estimating probabilities using a logistic regression equation. The equation used in the algorithm is,

$$\log(p /1 - p) = \beta_0 + \beta_{num}$$

Here, If the log(p/(1-p)) is greater than zero, then the success ratio appears to be greater than half of 100 percent every time.

$$F_1 \text{ - score} = 2 * [A*B] / [A+B]$$

3.1.4 Support Vector Machine

Support Vector Machine works towards structural risk minimization (SRM) to find the best hyperplane that separates two input spaces. SVM could classify positive, negative, and neutral sentiments.

4. EXPERIMENT AND VALIDATION

A python based machine learning system is developed for the experiments. A single system with switch mode to Naïve Bayes, Logistic Regression, and SVM for comparison. The system analyses the tweets about covid vaccine drive in India for the methodologies, each other. The packages such as Pandas, Numpy, Sklearn were used to build the system. The result of each analysis is found to be promising. The accuracy of each methodology is compared.

4.1 Dataset

We retrieved 209930 publicly available tweets. Following extraction, we identified vaccine sentiments and opinions of tweets. A tweet can contain many things, from plain text, mentions, hashtags, links, punctuation to many other things. When working on a data science or machine learning project, it is necessary to remove these things before processing the tweets further. It involves the following steps:

- Lowercasing all letters
- Removing hashtags, mentions, and links
- Punctuations removal (including filtering non- alphanumeric characters if necessary)
- Tokenization – Here, the text is split into smaller components, for example, a paragraph into a list of sentences or a sentence into a list of words.
- Stop words removal - Stop words are considered unimportant to the meaning of a text. These words may seem important to us humans, but to machines, these words may be considered a nuisance to the processing steps.

	tweet
0	isn t best poll promise ever free covid vaccin...
1	now states shall wait thier vidhan sabha elect...
2	
3	they said vaccine when free covid vaccine new ...
4	bjp presenting free covid vaccine state manife...

```
print(df.head(10))
```

```
                                               tweet  polarity  subjectivity
0  isn t best poll promise ever free covid vaccin...  0.362500      0.595833
1  they said vaccine when free covid vaccine new ...  0.378788      0.718182
2  bjp presenting free covid vaccine state manife...  0.400000      0.608333
3  the shame facedness bjp crossed boundaries get...  0.400000      0.500000
4  just days ago pm said roadmap ready provide fr...  0.300000      0.650000
5  what non bjp ruled states indians didn t vote ...  0.400000      0.800000
6  big pharma big money big egos year old univers...  0.025000      0.125000
7  latest astrazeneca oxford world beating vaccin...  0.500000      0.900000
8  free covid vaccine bihar with kind crowd ralli...  0.316667      0.866667
9  read s manifesto bihar elections covid vaccine...  0.100000      0.833333
```

Figure 1. A sample cleaned data and dataset with polarity and subjectivity

It is also important to keep in mind that stop words are largely language-dependent. The stop words such as for, to, and, or, in, out in English. The dataset is cleaned using the tweet-preprocessor library in python, and the polarity and subjectivity using TextBlob, as shown in Figure 1.

5. RESULT AND DISCUSSION

In the experiment, we found the difference in the pervasiveness of like and dislike towards the vaccination drive in INDIA. in the analysis of sentiments, with positive being the dominant polarity. The negative polarity is taken as 0, and the positive polarity is taken as 1. Polarity is a float value that lies in the range [0,1]. Values near 0 and 1 indicate negative sentiments and positive sentiments, respectively.

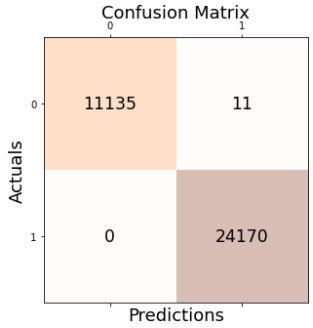

Figure. 2. The polarity of the tweets processed with SVD and shows that precision = 0.99, accuracy = 0.99, and recall = 1.0

The Figure 2 shows the confusion matrix for the polarity of tweets processed using SVM. There were 11135 actual false values which were predicted false. There were 11 actual false values which were predicted true. No true value was predicted false. There were 24170 true values which were predicted true.

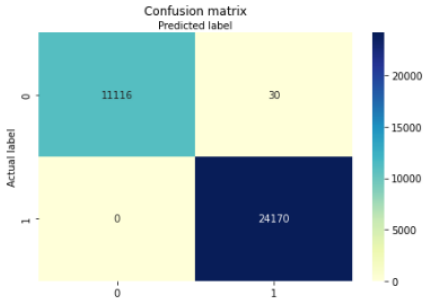

Figure 3. The prediction of labels with respect to positive polarity and negative polarity using logistic regression shows that precision = 0.999150, accuracy = 0.998760, and recall = 1.0

The Figure 3 shows the confusion matrix for the polarity of tweets processed using Logistic Regression. There were 11116 actual false values which were predicted false. There were 30 actual false values which were predicted true. No true value was predicted false. There were 24170 true values which were predicted true.

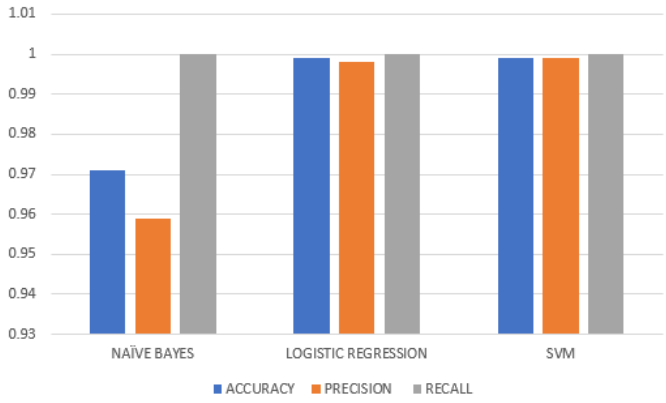

Figure 4. The statistical measures shows various level of precision, accuracy and reall for three different classification algorithm on sentiment based on tweets and find that Logistic Regression and SVM are the best classification for sentiment analysis

The barchart shows the accuracy, precision and recall of different sentimental analysis models like Naïve Bayes, Logistic regression and SVM. The X-axis plots the models, and the Y-axis plots accuracy, precision, and recall. The SVM and Logistic Regression is the best model for Sentimental Analysis compared to Naïve Bayes, as shown in Figure 4.

Subjectivity is also a float value within the range [0,1]. Subjectivity refers to personal opinion or factual information in the text. High subjectivity indicates that the text contains more personal opinions, emotions, or judgements, whereas low subjectivity indicates factual information.

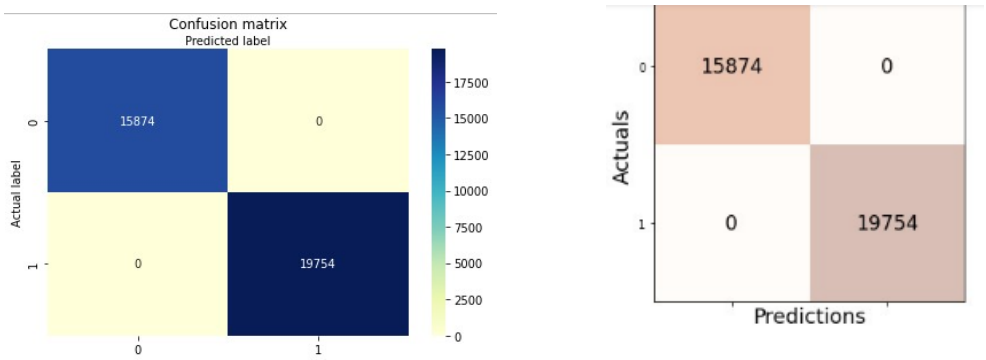

Figure 5. The sentiment analysis about vaccination against COVID-19 by using logistic regression and SVM

The confusion matrix for the subjectivity of tweets processed using SVM is shown in Figure 5. The 15874 actual false values were predicted false and the 19754 true values were predicted true. The second matrix for the subjectivity of tweets is processed using Logistic Regression. The 15874 actual false values were predicted false and the 19754 actual values were predicted true.

The result showed more positive tweets (almost 70%) than negative ones. Sentiment analysis towards COVID-19 vaccines can help the government make wise decisions regarding allocating funds and vaccination roll-out plans. The developed models using the Naïve Bayes and logistic regression algorithm can help classify tweets according to their polarity, especially in English.

6. CONCLUSION

The sentiment analysis of vaccination drive in INDIA was analyzed with different machine learning algorithms. The data is processed to get fine-tuned for the analysis. The sentiment towards the vaccination drive through logistic regression explicitly achieved well regarding all algorithms. The results are well compared with the same of different machine learning techniques. The logistic algorithms limit the assumptions further. The Naive Bayes algorithm is comparatively performed well. The analysis results give an understanding of the sentiments and opinions about vaccinations that can help public health agencies boost positive messages and remove negative ones to improve vaccinations. The role of social media plays a good role in the present contagious circumstances.

REFERENCES

[1] J. Cheriyan, V. S. Chandran, and L. S. Nair, “An awareness to multisystem inflammatory syndrome (mis-c) using social networks”, in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 01–07.

[2] R. Feldman, “Techniques and applications for sentiment analysis”, Communications of the ACM, vol. 56, no. 4, pp. 82–89, 2013.

[3] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams engineering journal, vol. 5, no. 4, pp. 1093–1113, 2014.

[4] L. Bing, "Sentiment analysis and opinion mining (synthesis lectures on human language technologies)," University of Illinois: Chicago, IL, USA, 2012.

[5] M. Shamrat, S. Chakraborty, M. Imran, J. N. Muna, M. M. Billah, P. Das, and O. Rahman, "Sentiment analysis on twitter tweets about covid-19 vaccines using nlp and supervised knn classification algorithm," Indonesian Journal of Electrical Engineering and Computer Science, vol. 23, no. 1, pp. 463–470, 2021.

[6] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3. IEEE, 2004, pp. 32–36.

[7] A. Jishag, V. Rakhesh, S. Mohan, N. Vinayak Varma, V. Shabu, L. S. Nair, and M. Menon, "Automated review analyzing system using sentiment analysis," in Ambient Communications and Computer Systems. Springer, 2019, pp. 329–338.

[8] L.-J. Cao and F. E. H. Tay, "Support vector machine with adaptive parameters in financial time series forecasting," IEEE Transactions on neural networks, vol. 14, no. 6, pp. 1506–1518, 2003.

[9] N. F. Da Silva, E. R. Hruschka, and E. R. Hruschka Jr, "Tweet sentiment analysis with classifier ensembles," Decision Support Systems, vol. 66, pp. 170–179, 2014.

[10] N. M. Lal, S. Krishnanunni, V. Vijayakumar, N. Vaishnavi, S. Siji Rani, and K. Deepa Raj, "A novel approach to text summarisation using topic modelling and noun phrase extraction," in Advances in Computing and Network Communications. Springer, 2021, pp. 285–298.

[11] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: a systematic review," American journal of public health, vol. 107, no. 1, pp. e1–e8, 2017.

[12] E. M. Glowacki, G. B. Wilcox, and J. B. Glowacki,"Identifying# addiction concerns on twitter during the covid-19 pandemic: A text mining analysis," Substance abuse, vol. 42, no. 1, pp. 39–46, 2021.

[13] A. C. Sanders, R. C. White, L. S. Severson, R. Ma, R. McQueen, H. C. A. Paulo, Y. Zhang, J. S. Erickson, and K. P. Bennett, "Unmasking the conversation on masks: Natural language processing for topical sentiment analysis of covid-19 twitter discourse," medRxiv, pp. 2020–08, 2021.

[14] K. P. Murphy et al., "Naive bayes classifiers," University of British Columbia, vol. 18, no. 60, pp. 1–8, 2006.