**Twitter data handling and sentiment analysis using big data tools and frameworks**

**Big Geodata**

**Submitted by,**

**Aranya Jha**

**Meghaa S**

**Utkarsh**

**Vishwas R**

**M.Sc. Geoinformation Science and Earth Observation**

**April 2023**

**Dehradun**

| **FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION UNIVERSITY OF TWENTE, ENSCHEDE, THE NETHERLANDS** |  |
| --- | --- |
| **INDIAN INSTITUTE OF REMOTE SENSING,  INDIAN SPACE RESEARCH ORGANISATION, DEHRADUN, INDIA** | ***iirs*** |

**Table of Content**

[**Introduction 3**](#_30j0zll)

[Data Collection 4](#_1fob9te)

[Study Area 4](#_3znysh7)

[**Literature Survey 5**](#_2et92p0)

[Objective 6](#_tyjcwt)

[**Methodology 7**](#_yviay4xcs9xt)

[Flow chart 7](#_w74cvou2li6k)

[Data Hydration 7](#_4d34og8)

[Pre-processing 7](#_2s8eyo1)

[Training & Analysis 8](#_17dp8vu)

[**Result & Discussion 9**](#_3rdcrjn)

[Python Library - TextBlob 10](#)

[Lexicon Based Approach 12](#)

[**Conclusions 14**](#_44sinio)

[Contribution of Members 14](#)

[**References 15**](#_3j2qqm3)

# 

# **Introduction**

Social media platforms in today’s world have become important tools for connecting people globally. In our day-to-day lives, we come across various social media platforms like Facebook, WhatsApp, Instagram, etc. However, the most influential social media platform today is Twitter. Twitter not only provides common folks with a platform to showcase their views to the world but also provides a platform to share any happenings and seek help if required.

The COVID-19 pandemic shattered the world's normalcy, leading to grief, deaths, and mourning, and India was no exception. India, along with its neighbouring countries, was one of the most affected regions in the world. Twitter activity increased manifold during the pandemic, as more and more people joined Twitter to share their emotions on various happenings and government policies. These tweets (in Twitter, the messages used to communicate are called Tweets) were a mix of both positive and negative sentiments towards various issues. These Twitter data (Geo-tagged data) can be a valuable asset in analysing the public sentiments throughout the pandemic in different regions of the world if analysed properly.

Opinion mining, popularly known as sentiment analysis, is done to analyse the sentiment behind the body of text. The process of sentiment analysis involves cleaning texts by removing punctuation, emoticons, URLs etc such that the sentiment of the refined text can be assessed using relevant machine learning model, taking into consideration the way a human thinks (Gautam & Yadav, 2014). Natural Language Processing (NLP) is a distinguished method to analyse the sentiments. It involves concepts of data mining as well as machine learning simply and vectorize texts (natural language) and classify them as positive, negative, or neutral.

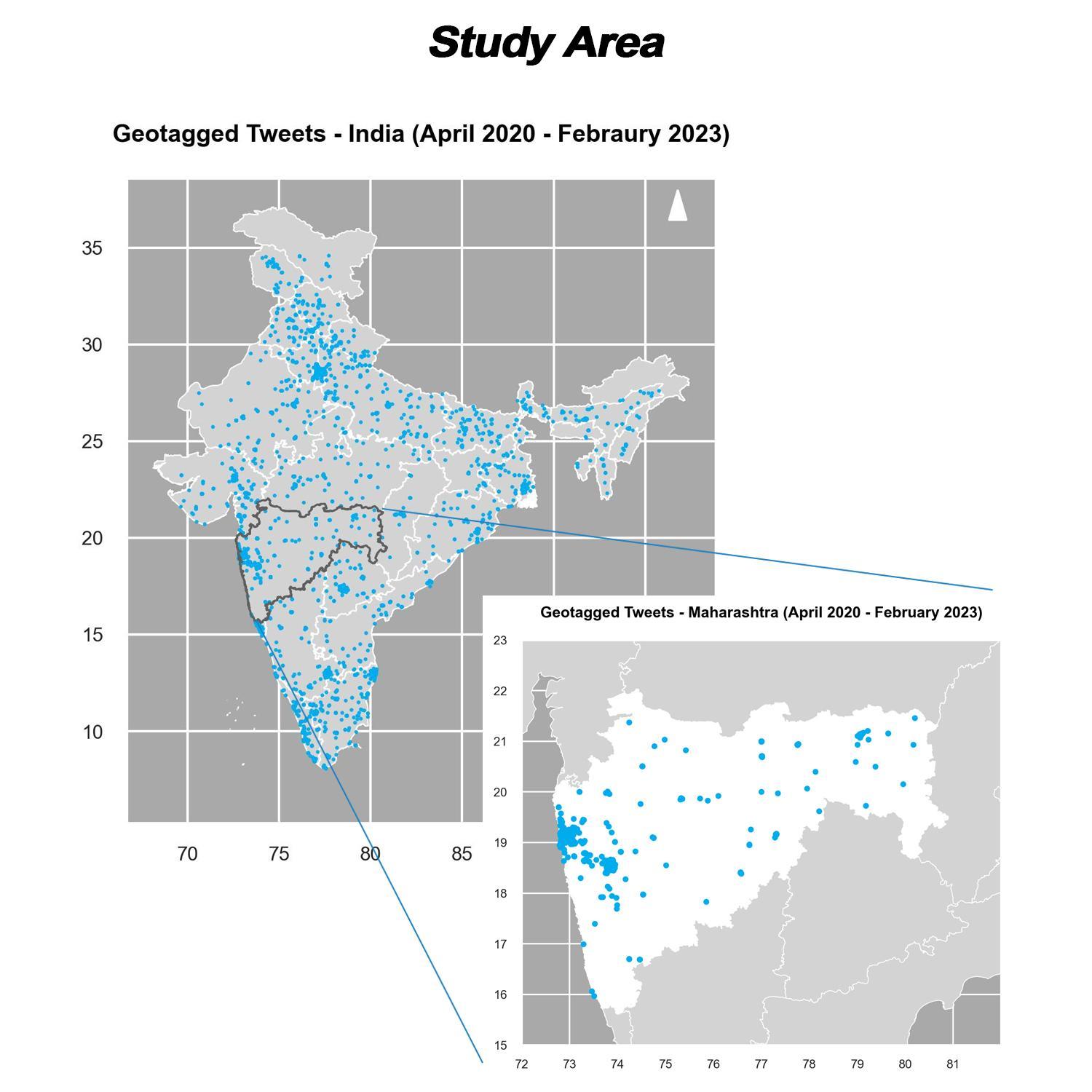
In the present study, geotagged tweets of the Maharashtra state in India were analysed. For this analysis, the "Coronavirus (COVID-19) Geo-tagged Tweets dataset" was downloaded from IEEE data port for the years 2020, 2021, and 2022. This data was made to undergo data hydration, filtering, and pre-processing before it was actually used for the analysis. Broadly, three labelling techniques were considered beginning with manually labelling the sentiment. Later on, Lexicon based approach and TextBlob were also used to generate polarity for textual data. Thereafter, the entire data was divided into training and testing data sets (70:30), and provided as input to two popular ML models namely SVM and Random Forest. to compute accuracy.

## Data Collection

The data set used “Coronavirus (COVID-19) Geo-tagged Tweets dataset”, was downloaded from IEEE dataport ([IEEE dataset](https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset)). Tweets about COVID-19 for the years 2020, 2021, and 2022 were included in this dataset. and constituted approximately 500,000 entries organized in a CSV file format over a period of 36 months. The datasets contained two fields: tweet ID and sentiment score. In this study, only the tweet IDs were utilized for data hydration.

## Study Area

The study area for the project was chosen as the state of **Maharashtra** in India**,** which has a population of more than 13 crore and covers an area of approximately 3 lakh square kilometres. Maharashtra was selected as it was one of the most affected states during the COVID-19 pandemic and Mumbai, the capital of Maharashtra, is known as the business capital of the country, with a significant number of active Twitter users. Another important thing to consider was that Maharashtra has several important cities including Mumbai, Pune, Nagpur and Aurangabad. During the pandemic, there has been a series of news regarding COVID-19 spread in the slums of Mumbai (Hollingsworth & Mitra, 2020) and people migrating (Desk, 2020). Twitter was flooded with emotions regarding these issues during the same period.



*Figure 1: Study Area - Maharashtra*

# Literature Survey

(BalakrishnanGokulakrishnan, Pavalanathan Priyanthan, Prasath, & AShehan Perera, 2012) describes the sentiment analysis processes involved including pre-processing, classification of tweets into Positive, Neutral and Negative based on the emotional content of the tweets. Performance of different classification algorithms were analysed in this paper based on metrics like precision and recall. It has pre-processed the tweet to eventually obtain a cleaned format of tweets by omitting irrelevant contents. This paper talks about machine learning algorithms commonly used like Naive Bayes, Random Forest and Support Vector Machines (SVMs) for the twitter sentiment analysis. This study states that SVM and Random Forest classifiers give acceptable results.

(Tao, Hauff, Houben, Abel, & Wachsmuth, 2015) discuss the challenges and difficulties faced while processing and analysing huge volumes of unstructured data generated by various social media platforms like Facebook, twitter etc. They have identified a set of common traits and generalised a variety of typical Twitter data use cases and developed the Twitter Analytical Platform (TAP) in order to use data from Twitter for analytical purposes. According to authors, the platform would allegedly offer a Twitter Analysis Language (TAL) that is customised to a certain topic as the interface to its feature set.

(Lamsal, 2021) describes in detail about the reason behind preparation of geotagged COVID- 19 twitter dataset, how the dataset was prepared and how to handle large scale data. Keywords related to pandemic were used to identify tweets related to COVID-19 thus forming the dataset that contains relevant tweet IDs. Twitter does not give permission to openly display entire tweet information, so this paper gives two ways to hydrate tweet ids - Using the hydrator app or using the twarc library in python. The paper also talks about the python library TextBlob that computes sentiment scores in a continuous range rather than discrete values.

(Naseem, Razzak, Khushi, Eklund, & Kim, 2021) have used a python library - tweepy for hydration and then used TextBlob to generate sentiment score polarity. By computing the score as a polarity [-1 to 1], TextBlob can reveal a sentence's mood. This paper uses both machine learning and deep learning classifiers to classify the sentiment of text.

(El Rahman, Alotaibi, & Alshehri, 2019) have used both supervised and unsupervised Machine Learning algorithms in their study to classify the tweets as Positive, Neutral and Negative based on their sentiment. The lexicon-based approach is used labelling the tweets where two documents with positive and negative labels are used to compare the tweet and provide a sentiment score. This was followed by training and classification of tweets using several Machine Learning algorithms including SVM, Random Forest, Naive Bayes, Maximum Entropy, Bagging. This study proves that the Maximum Entropy method was best among all algorithms for their dataset.

(Aminuddin et al., 2021) developed a web-based system to visualise the pandemic status in Malaysia using Twitter data. The data was collected using the web scraping tool Twint and filtered by keywords, hashtags, and location tags on tweets. The SVM classification model was used to conduct sentiment analysis on the data, and the model was evaluated using accuracy, recall, precision, and F1-measure. The study aimed to contribute to effective monitoring and management of the COVID-19 pandemic by utilising social media data analysis.

## Objective

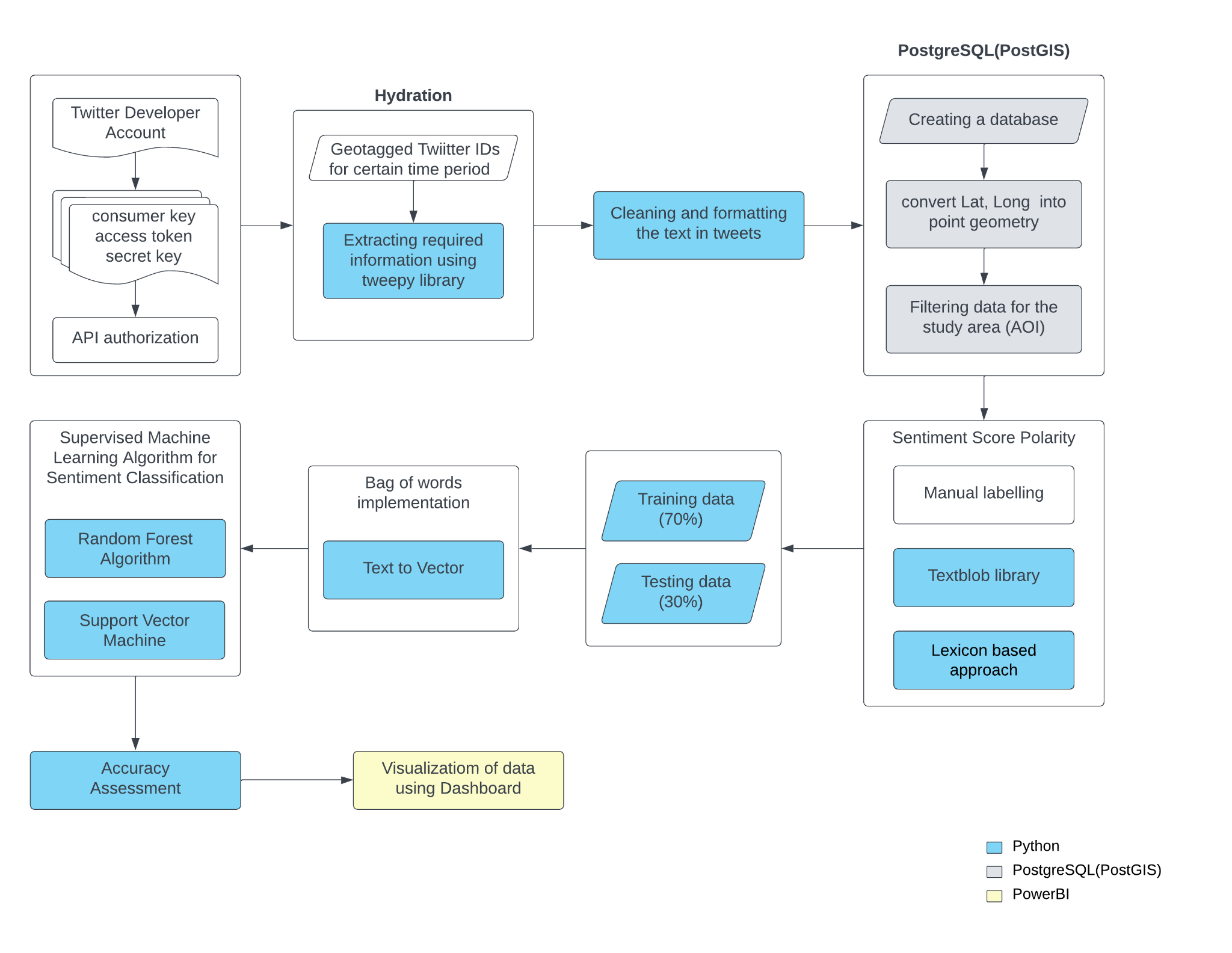
The prime focus of this study is to utilise the potential of machine learning algorithms namely Random Forest (RF) and Support vector machine (SVM), to analyse sentiment of geotagged Tweets related to COVID-19 times.

The sub-objectives that were identified for the study are:

* To acquire the necessary information from geotagged tweets within the study area through a data hydration process using big geo-data tools.
* To perform sentiment analysis using Machine learning algorithm with different labelling methods and visualise the findings using a dashboard.

# Methodology

## Flow chart



*Figure 2: Work flow of Methodology*

## Data **Hydration**

A dataset containing IDs of geo-tagged tweets related to the COVID-19 pandemic was downloaded as multiple csv files from [IEEE DataPort](https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset). A process known as Hydration was carried out on these IDs to extract information such as name, place, co-ordinates, date and time, Tweet text etc. Hydration can be done in python programming using several libraries like twarc, tweepy (Lamsal, 2021) etc or with a dedicated [Hydrator](https://github.com/DocNow/hydrator) app. Both processes require an API authorization from a Twitter Developer account to retrieve the required information. In this project, Tweepy library was used.

## Pre-processing

Total number of tweets IDs was around 5 lakhs for the entire world. Tweepy library (Naseem et al., 2021)was used in the current study as it provided an option to filter down the data by region. For India, 11858 tweets were gathered from March 2020 to February 2023.

A Geodatabase was created using PostgreSQL (PostGIS) for further analysis of the data. The Latitudes and Longitudes were converted into point geometry. The visualisation of data on map showed distribution of tweets throughout the country and thus helped in the selection of study area which consisted of 2628 tweets.

Regex library was used to clean and format the text portion of the tweets (Aminuddin et al., 2021). The resulting new text was free from any non-alphanumeric symbols, indentations, weblinks and stop-words.

## Training & Analysis

**Training:** After the pre-processing, the cleaned tweets were labelled using three different methods: Manual Labelling, Python Library: Text Blob and Lexicon based Approach. The tweets were labelled as Positive-1 (score>0.25), neutral-0 (-0.25<score<0.25) and Negative-1 (score<-0.25).

In Manual Labelling, the labels were provided without the intervention of a machine.

In the second method, the python library TextBlob (Naseem et al., 2021)was used to generate the Sentiment Score Polarity.

Finally, in a Lexicon based approach (El Rahman et al., 2019), two text files containing positive and negative words respectively were used to compare the tweets and assign a suitable sentiment score.

Once the training was completed as per requirements, the final dataset (containing cleaned tweets with appropriate sentiment score) had to be fed to the appropriate ML model for analysis.

*Table 1: Polarity of Scores for different Labelling methods*

|  | Manual Labelling | Text Blob | Lexicon based Approach |
| --- | --- | --- | --- |
| Positive (+1) | 350 | 232 | 355 |
| Neutral (0) | 341 | 731 | 506 |
| Negative (-1) | 309 | 38 | 104 |

**Analysis:** On the basis of literature survey, two well-known algorithms were considered for classification - Random Forest and Support Vector Machine (BalakrishnanGokulakrishnan, Pavalanathan Priyanthan et al., 2012). Our labelled dataset was split into 70% training and 30% testing. Both the train and test dataset had the cleaned tweets which had to be converted to a structured format i.e., vectors. This was achieved using the Bag of words model. Training data containing vectorized text and reclassified labels were used to fit the model, while the testing data containing only vectorized text was used for prediction. Later the labels predicted by model and the actual label in the testing set were compared for accuracy assessment.

# **Result & Discussion**

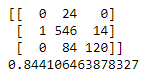
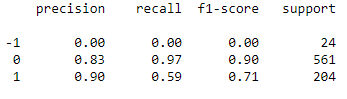
After successfully hydrating & pre-processing as well as generating sentiment score polarity using different methods, sentiment classification is done using supervised machine learning (ML) algorithms. The final data is split into a training and testing set which are fed to ML models. The metrics used for validation are: confusion matrix, overall accuracy score and the classification report.

## 

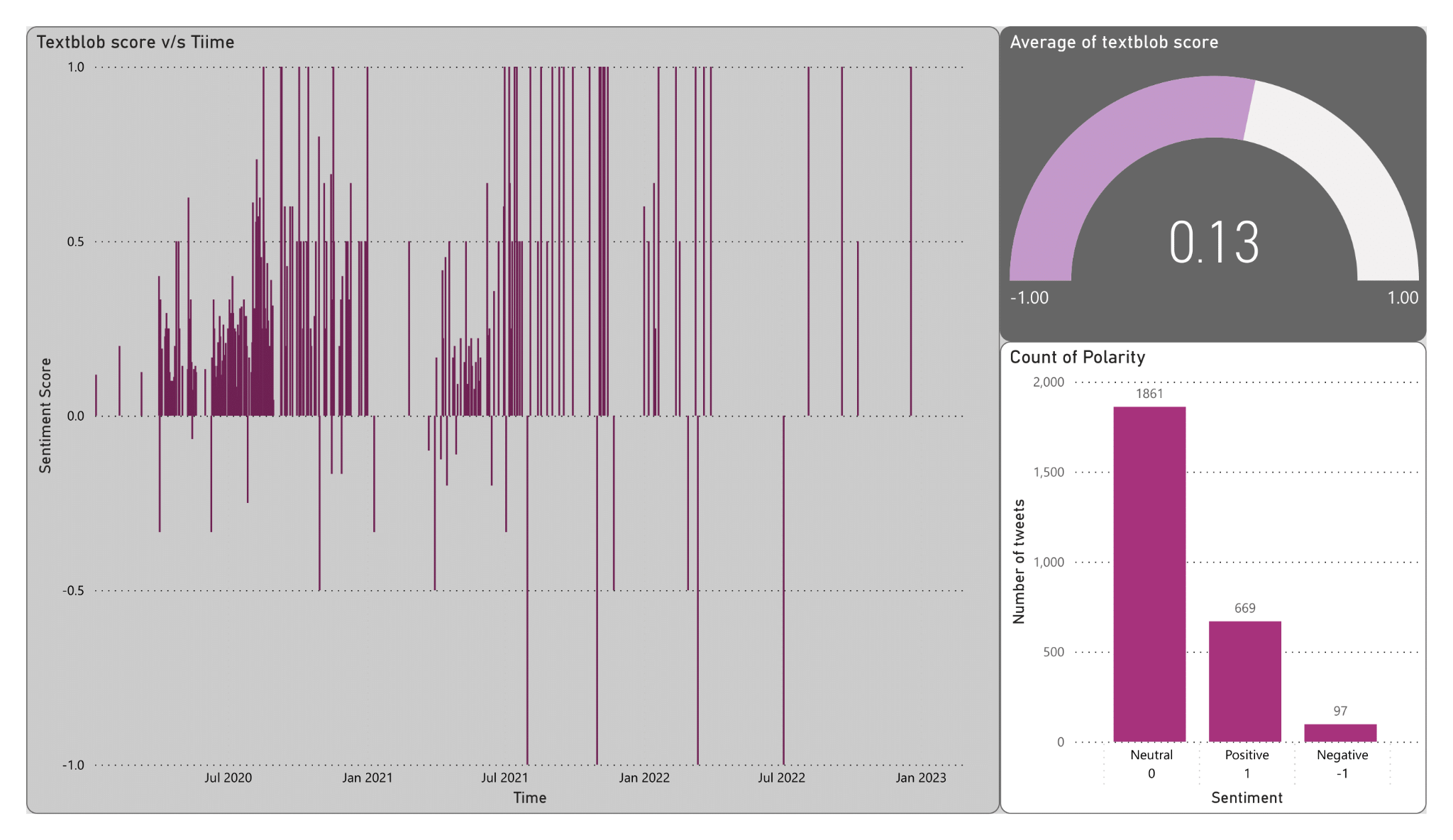
*Figure 3: Dashboard showing distribution of acquitted data*

## Python Library - TextBlob

1. Random Forest: The major hyperparameter is the number of trees whose optimal value came out to be 50 after trial and error. For our dataset we obtained an accuracy of 84.41%.

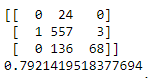
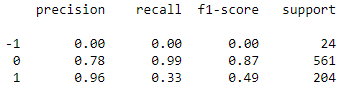
*Figure 4: Confusion Matrix, Accuracy Score & Classification Report - RF (Textblob)*

****

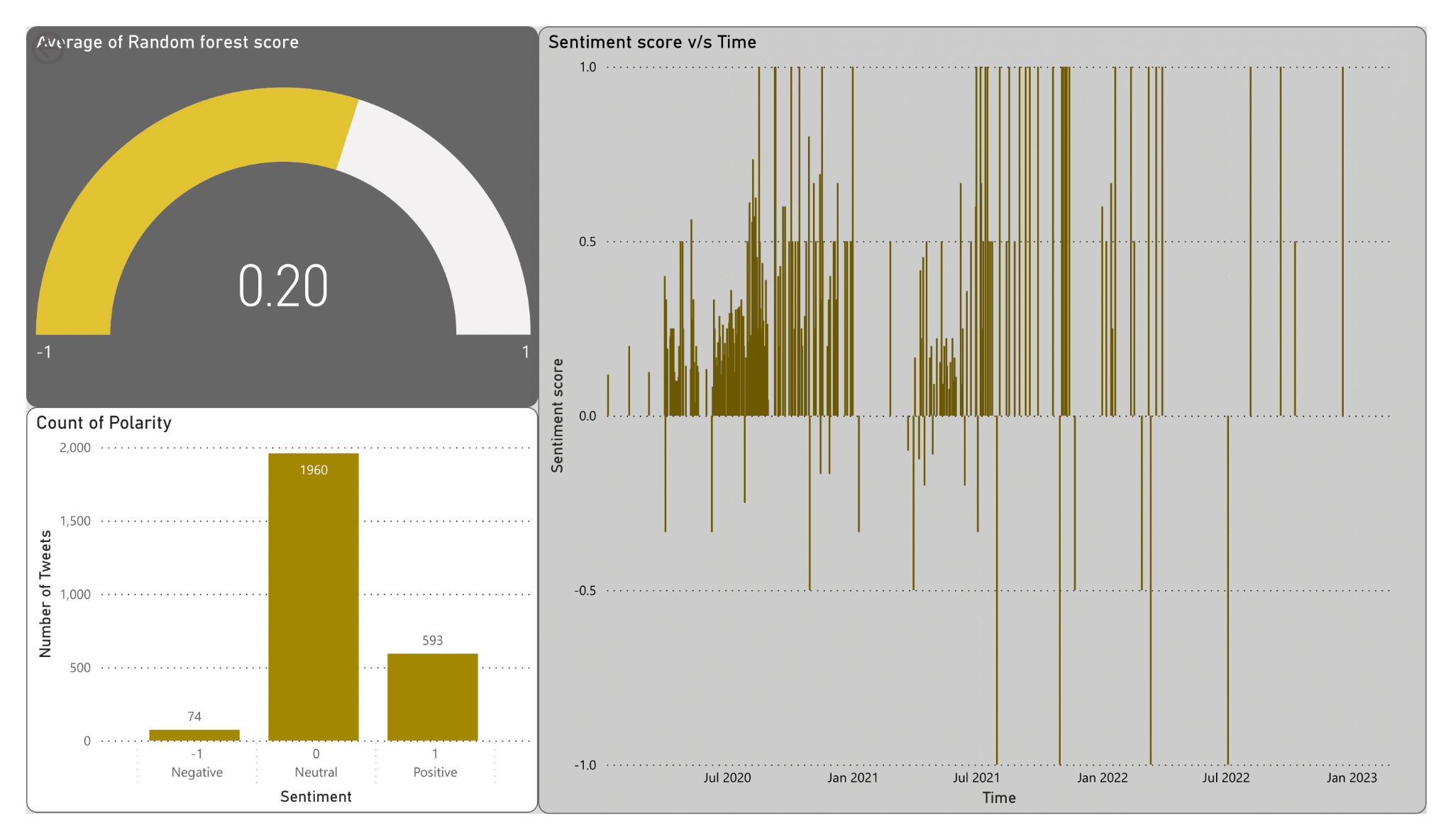
*Figure 5: Dashboard for TextBlob data*

1. Support Vector Classifier: Classification was done considering two kernels (used to convert linearly inseparable data to linearly separable), linear and radial basis function (rbf). For our dataset, an accuracy of 79.21% was obtained for the rbf kernel and accuracy of 84.28% for the linear kernel

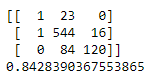
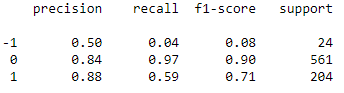
.

*Figure 6: Confusion Matrix, Accuracy Score & Classification Report - SVC RBF (Textblob)*



*Figure 7: Dashboard for TextBlob - RF*

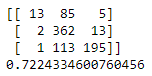
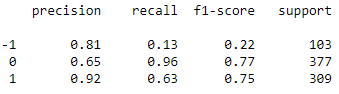
*Figure 8: Confusion Matrix, Accuracy Score & Classification Report - SVC Linear (Textblob)*

## 

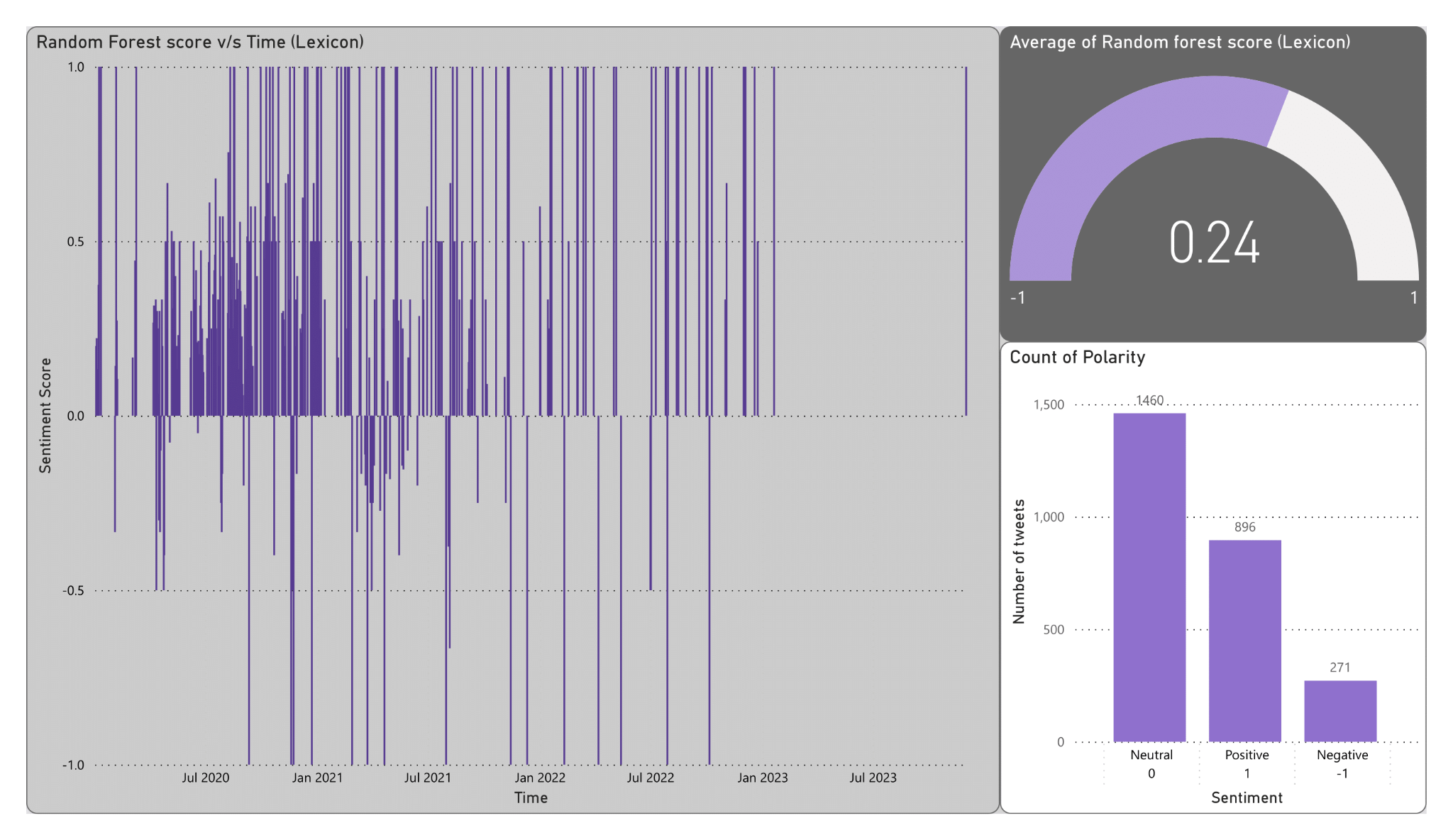
*Figure 9: Dashboard for TextBlob - SVC (Linear)*

## Lexicon Based Approach

1. Random Forest: Classification accuracy obtained for random forest is 72.24%

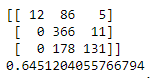
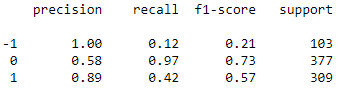
 

*Figure 10: Confusion Matrix, Accuracy Score & Classification Report - RF (Lexicon)*

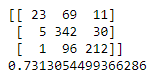
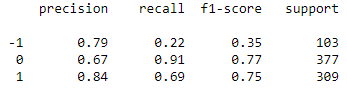


*Figure 11: Dashboard for Lexicon - RF*

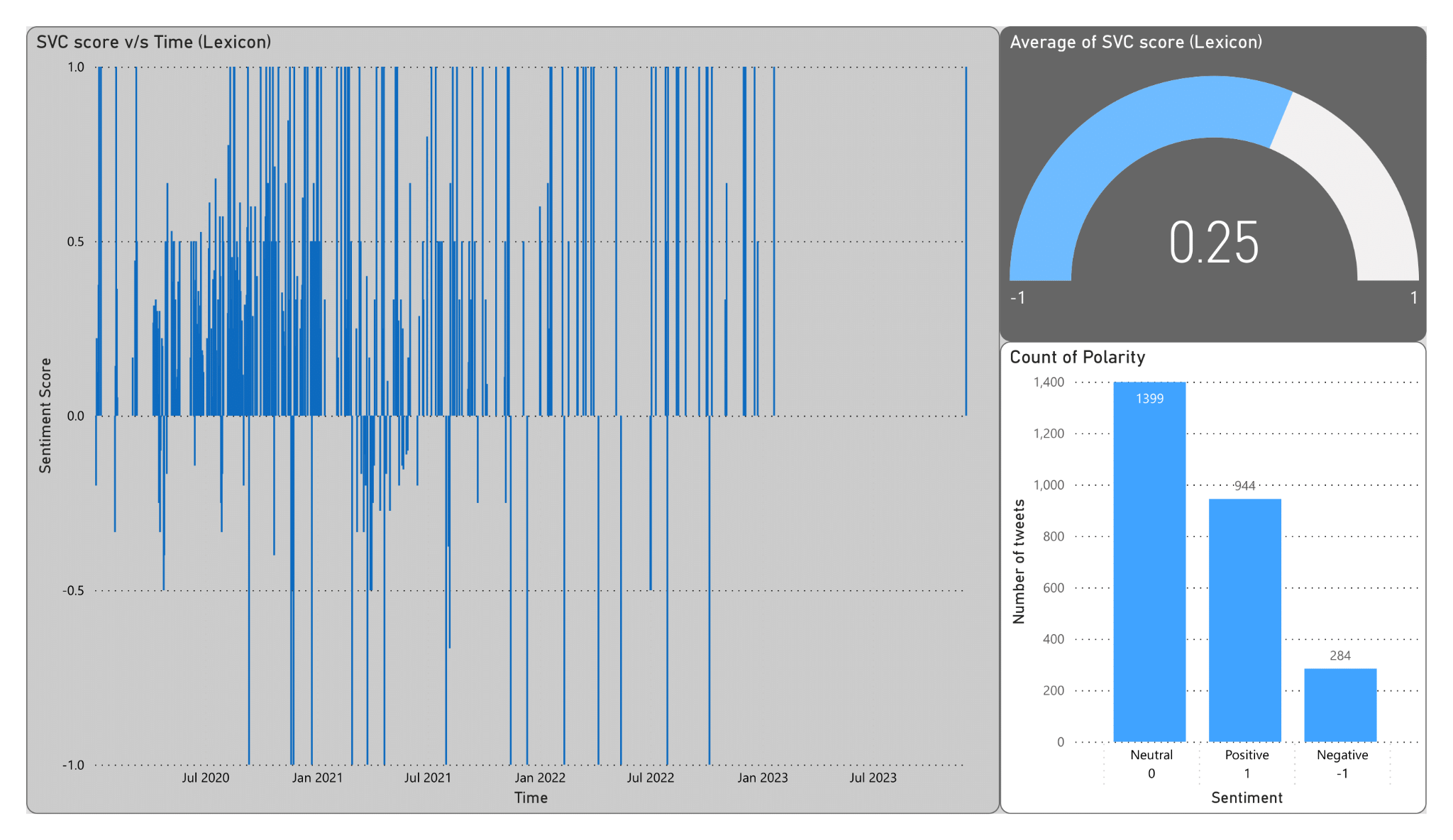
1. Support Vector Classifier: When the kernel is chosen as rbf, accuracy comes out to be 64.51% and when kernel is linear, accuracy is 73.13.

*Figure 12:: Confusion Matrix, Accuracy Score & Classification Report - SVC RBF (Lexicon)*

*Figure 13: Confusion Matrix, Accuracy Score & Classification Report - SVC Linear (Lexicon)*



*Figure 14: Dashboard for Lexicon based approach - SVC (Linear)*

For labelling the tweets, multiple approaches were adopted. Manual labelling of tweets did not provide good results as it's a rigorous method. It's not manually feasible to provide labels for a huge dataset, at the same time if the task of manual labelling is divided then there may be mismatch between the labels due to difference in perception. For this dataset when training and testing ratio is taken as 90:10, accuracy came around 62% but for ration 70:30, accuracy was around 50%.

Thus, we decided to label the tweets using TextBlob and Lexicon based Approach and we got satisfactorily good results. Both approaches were able to predict the sentiment score when provided with the vectorized form of tweets.

Apart from this, in case of support vector classifier, for all three labelling approaches, it was observed that SVC with linear kernel was outperforming SVC with rbf kernel in terms of accuracy. This may have happened because our dataset has mainly 3 classes (positive, neutral, negative) which may already be linearly separable.

Visualization of data and results for the current study was accomplished by creating an interactive dashboard using Power BI. Figure 3 represents general information of the geotagged twitter dataset for COVID-19, along with a map of Maharashtra showing the location of tweets in the form of points. For each labelling method i.e., TextBlob & Lexicon based approach, a total of 3 dashboards are created.

* The first one represents the relation between the sentiment score by TextBlob and time as well as the count of polarity for each sentiment.
* The second dashboard for each labelling method shows relation between random forest score and time as well as count of polarity.
* The third dashboard for each labelling method shows relation between support vector classifier (kernel - linear) and time as well as count of polarity.

A very interesting pattern was observed via the dashboard regarding the two modelling approaches. Even though accuracy for ML models using TextBlob scores was better than that of the Lexicon based approach, the number of negative sentiments were less than 100 in case of TextBlob method whereas count of negative tweets was above 250 in case of lexicon-based approach.

# **Conclusions**

To summarise, in this project we have attempted to assess the performance of machine (ML) learning algorithms like Random Forest (RF) and Support Vector Machine (SVM) to perform sentiment analysis of the Geotagged tweets related to the COVID-19 in Maharashtra. In addition, we have tried to shed light on the effect of different labelling methods and how they affect the overall accuracy.

The overall result shows that both SVC (linear kernel) and Random Forest have performed better for this dataset with the TextBlob labelling method with an accuracy of 84.41 % and 84.28% respectively. The Lexicon based labelling method as well had good accuracy of 72.24% and 73.13%, while the manual labelling method was found to give a poor performance with 50 % and 48% accuracy. It was also observed that in case of support vector classifiers, the accuracy when the kernel is linear is better than that of the radial basis function kernel.

This project can be further extended to analysing the performance of unsupervised Machine Learning methods, since only supervised Machine Learning methods were followed in this project. Apart from this, deep learning models like neural networks can be implemented as they might give even better results. Also, the tweet sentiments in the covid context are significantly different from the general context. Therefore, a unique labelling method to meet the covid context, could be developed in order to accomplish a more reliable sentiment analysis.

## Contribution of Members

*Table 2: Contribution of Individual Members*

| **Contributions** | **Contributors** |
| --- | --- |
| Training the data for ML | Coding part : Meghaa & hydration - All 4 |
| ML model Implementation | Aranya Jha, Utkarsh |
| Git Repository | Utkarsh |
| Dashboard | Vishwas R |
| Twitter Data handling   * Download * Cleaning & Formatting * Hydration Code * Hydrating data | Meghaa  Vishwas R  Meghaa  All 4 members |
| Report | All 4 members |
| PowerPoint PPT | All 4 members |

Git hub Repository: <https://github.com/Utkarsh-iirs/Team-IIRS>

# **References**

Aminuddin, R., Bistamam, M. A., Ibrahim, S., Mangshor, N. N. A., Fesol, S. F. A., & Wahab, N. I. F. A. (2021). A Sentiment Analysis Framework on COVID-19 in Major Cities of Malaysia based on Tweets using Machine Learning Classification Model. *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, 25–30. https://doi.org/10.1109/ICSET53708.2021.9612527

BalakrishnanGokulakrishnan, Pavalanathan Priyanthan, T., Prasath, N., & AShehan Perera. (2012). Opinion Mining and Sentiment Analysis on a Twitter Data Stream. *The International Conference on Advances in ICT for Emerging Regions -*, 229.

Desk, E. W. (2020). The long walk of India’s migrant workers in Covid-hit 2020. *Indian Express*, p. 1.

El Rahman, S. A., Alotaibi, F. A., & Alshehri, W. A. (2019). Sentiment Analysis of Twitter Data. *2019 International Conference on Computer and Information Sciences, ICCIS 2019*. https://doi.org/10.1109/ICCISci.2019.8716464

Gautam, G., & Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 7th International Conference on Contemporary Computing, IC3 2014*, (September), 437–442. https://doi.org/10.1109/IC3.2014.6897213

Hollingsworth, J., & Mitra, E. (2020). More than half of India’s Mumbai slum residents may have been infected with Covid-19, study suggests. *CNN*, p. 1.

Lamsal, R. (2021). Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, *51*(5), 2790–2804. https://doi.org/10.1007/s10489-020-02029-z

Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., & Kim, J. (2021). COVIDSenti : A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis. *IEEE TRANSACTIONS ON COMPUTATIONAL SOCIAL SYSTEMS*, 1–13.

Tao, K., Hauff, C., Houben, G. J., Abel, F., & Wachsmuth, G. (2015). Facilitating Twitter data analytics: Platform, language and functionality. *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, 421–430. https://doi.org/10.1109/BigData.2014.7004259