**Twitter data handling and sentiment analysis using Big Data tools/frameworks**

**Group Members:**

KANKUKURI SANDEEP KUMAR

IIRS2021418035335

NANCHARAIAH MADDALA

IIRS2021418035468

GANGADHAR KANCHI

IIRS2021418035467

DHANASEKARAN K

IIRS2021031611883

AMAN TIWARI

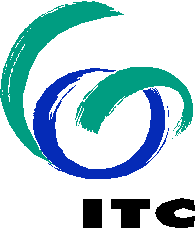
IIRS202103169978

**Submitted To:**

Mr. KAPIL OBERAI

Scientist/Engineer - SF

Geoinformatics Department



FACULTY OF GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

UNIVERSITY OF TWENTE, ENSCHEDE, THE NETHERLANDS

***iirs***

INDIAN INSTITUTE OF REMOTE SENSING, DEHRADUN, INDIA

**Table of Contents**

[**1.** **Introduction** 3](#_Toc102716717)

[Dataset collection 4](#_Toc102716718)

[Study Area 4](#_Toc102716719)

[**2.** **Literature Review** 5](#_Toc102716720)

[**3.** **Objectives** 6](#_Toc102716721)

[**4.** **Selection of database** 6](#_Toc102716722)

[**5.** **Selection of Dashboard** 6](#_Toc102716723)

[**6.** **Methodology** 7](#_Toc102716724)

[**7.** **Result and Discussion** 9](#_Toc102716725)

[Random Forest algorithm: 9](#_Toc102716726)

[Support Vector Classifier (SVC): 10](#_Toc102716727)

[SVC using Linear as kernel**:** 10](#_Toc102716728)

[SVC using Radial Basis Function (RBF) Kernel 11](#_Toc102716729)

[Naive Bayes: 12](#_Toc102716730)

[**8.** **Conclusions** 13](#_Toc102716731)

[**9.** **Contribution of Members** 13](#_Toc102716732)

[**10.** **References** 14](#_Toc102716733)

# **Introduction**

In the modern world, we are highly dependent on the social media. We use various platforms for sharing, expressing our thoughts, get news, entertainment, shopping, and many more. Social media becomes a powerful and an active source of information at the time of some crisis. Be it anthropogenic or natural like landslides, earthquake, or disease outbreak. Social media platforms like Facebook, Instagram, Twitter, LinkedIn, etc create an astronomical amount of data on day-to-day basis. People use these platforms for sharing the ground truth of the scenario and seeking help in need. This type of data becomes valuable because with proper analysis and planning, informed decisions can be taken. (Lamsal, 2021) It is a huge, complex, and always changing data. This type of data is also called as big data.

To better understand the users, a ton of analysis is done on these datasets. One such analysis is called as Sentiment Analysis/Opinion Mining/Emotion AI. It is a practice to extract the spectrum of subjective information using automatic tools in order to create structured and actionable knowledges. (Pozzi, Fersini, Messina, & Liu, 2016).

In this project, Sentiment Analysis of Covid-19 Twitter data is done. Twitter is a famous social media and microblogging platform. The users can interact using messages called as tweets. As of current information, 396.5 million twitter accounts are active. Tweets have a limit of 280 characters in non CJK (China, Japan, and Korean) languages. CJK language is restricted to 140 characters only. It has a well-documented Application Programming Interface (API) for accessing the data present on the platform (Lamsal, 2021).

From the available tweet ID’s, actual tweet information was extracted. This process of extraction is called as hydration. These hydrated data is managed using a database. Model fitting is done based upon the training samples using various algorithms in python (Jupyter Notebook). Using this model, sentiment prediction is done. An interactive dashboard is prepared for the visualization using Metabase.

## Dataset collection

Freely available *Geotagged Tweets Dataset* from [IEEE DataPort](https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset) is used in the project. This dataset contained the geo-tagged tweet IDs related to COVID-19 from March 20th, 2020 to March 31st, 2022. The complete data contains about 4.1 Lakhs tweets.

## Study Area

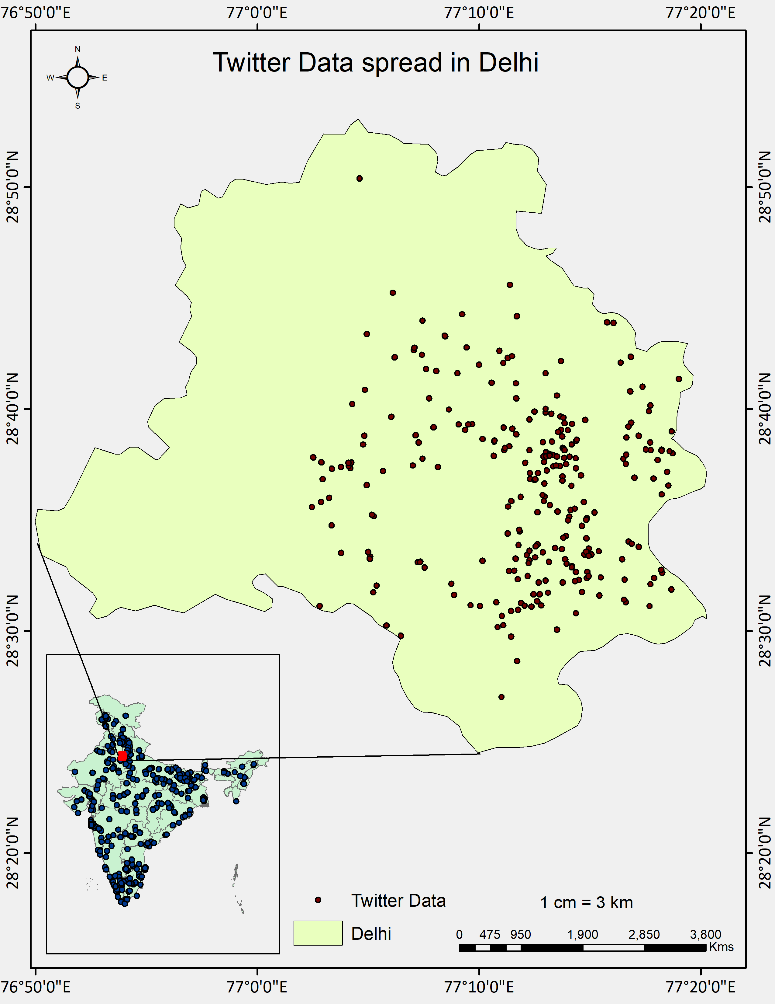
The selected study area for the project is New Delhi. This area is selected because it is the capital of India and was a hub for national as well as international travellers connecting all the northern region.

Figure 1: Twitter Data of Delhi

# **Literature Review**

Natural Language Processing, or NLP for short, is the automated manipulation of natural language such as speech and text.   Computational linguistics—rule-based human language modelling—is combined with statistical, machine learning, and deep learning models in NLP. These technologies, when used together, allow computers to process human language in the form of text or speech data and 'understand' its full meaning, including the speaker's or writer's intent and sentiment(Kanakaraj & Guddeti, 2015).

Naive Bayes is a very simple probabilistic model that tends to work well on text classifications and usually takes orders of magnitude less time to train when compared to models like support vector machines. It can be enhanced to match the classification accuracy of more complicated models for sentiment analysis by choosing the right type of features and removing noise by appropriate feature selection. Naive Bayes classifiers due to their conditional independence assumptions are extremely fast to train and can scale over large datasets. They are also robust to noise and less prone to overfitting. Ease of implementation is also a major advantage of Naive Bayes(Narayanan, Arora, & Bhatia, 2013).

Random forest is a supervised machine learning algorithm. It is used extensively in classification and regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. Random forest can handle continuous variable in case of regression and categorical variables in case of classification. Random forest randomly selects observations, builds a decision tree and the average result is taken(R, 2021).

Within our sentiment polarity model, SVM uses algorithms to train and classify text, going beyond X/Y prediction. It draws a hyperplane that separates the clusters, in our case, the sentiments positive, negative, and neutral (Ahmad, Aftab, & Ali, 2017).

# **Objectives**

1. To collection and management of the twitter data
2. To Extraction of tweet information for the Study Area
3. To Perform Sentiment Analysis using Machine Learning algorithms
4. To Creating a Dashboard for Visualization

# **Selection of database**

As the data contains coordinates information (point geometry), it can be treated as a spatial data. Also, the data satisfies all three characteristics of Big Data- Volume, Velocity and Variety. Hence the data can be treated as Big Geodata. There are many platforms for storing and management of big geodata such as Postgres, mongo dB, Apache spark, etc. From the literature survey, we understand that Postgres is best for handling structured data with less complex geometry. Hence, Postgres platform is used store and manage the data.

# **Selection of Dashboard**

There are many platforms for dashboard creation such as Metabase, Apache Superset, Mongo Atlas, Tableau, etc. Metabase is a java-based application. Installation and configuration of Metabase is very simple. It can be run on any system that has Java installed. It has excellent data integration capabilities, interactive visualizations, documentation & community support. Metabase was used for visualization of the data, as it is easy to integrate with Postgres database.

# **Methodology**

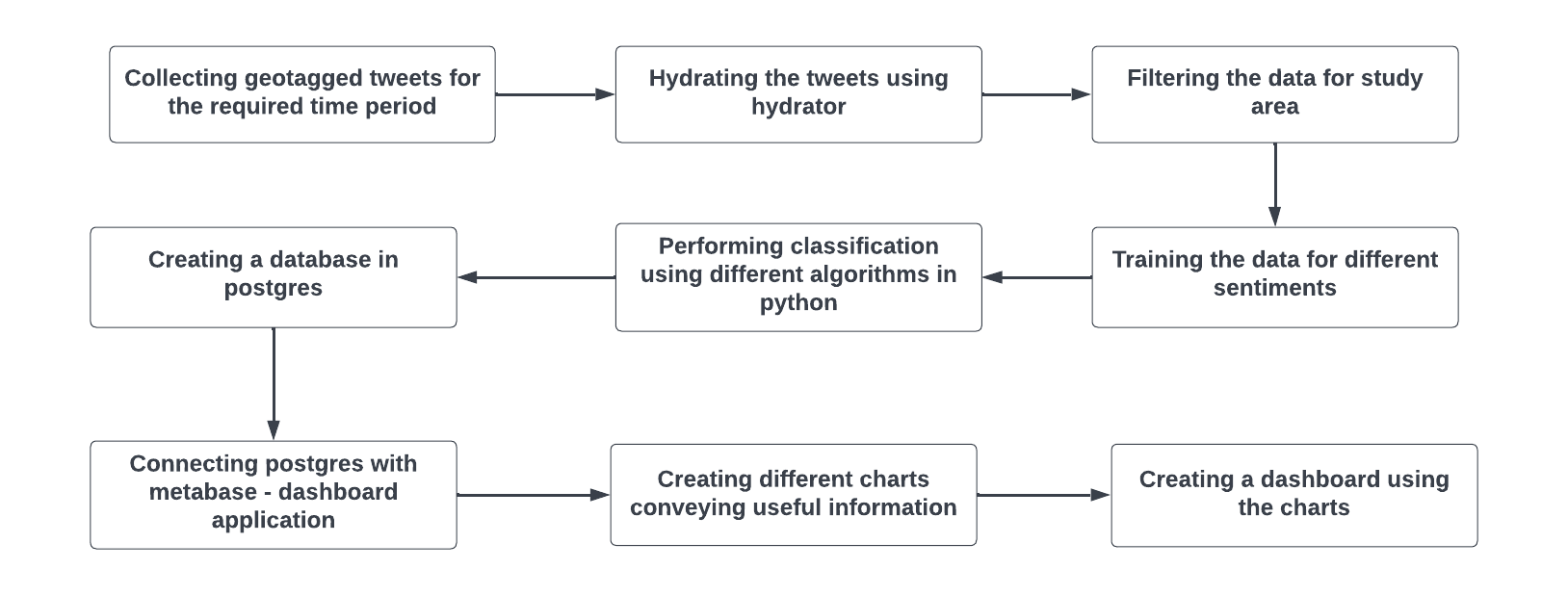


Figure 2 workflow of analysis

The collected data contains the tweet IDs in the csv format. These IDs are used to extract the

actual tweet information by a process known as hydration. The hydration is done using a tool- *Hydrator App or* by using python libraries. The initial tweet ID data is converted from csv to text using python. Hydrator app takes text file as an input and returns csv as an output. After the hydration process, around 4 lakhs data points containing information such as created date, place, text, URLs, Co-ordinates, user information such as id, name, followers etc are obtained for the entire world.

The hydrated tweets are stored in a geodatabase for further use. The database used here is PostgreSQL. It is an extensible, open-source object-relational database which can handle spatial data in an efficient manner(“What Is PostgreSQL?,” 2022).

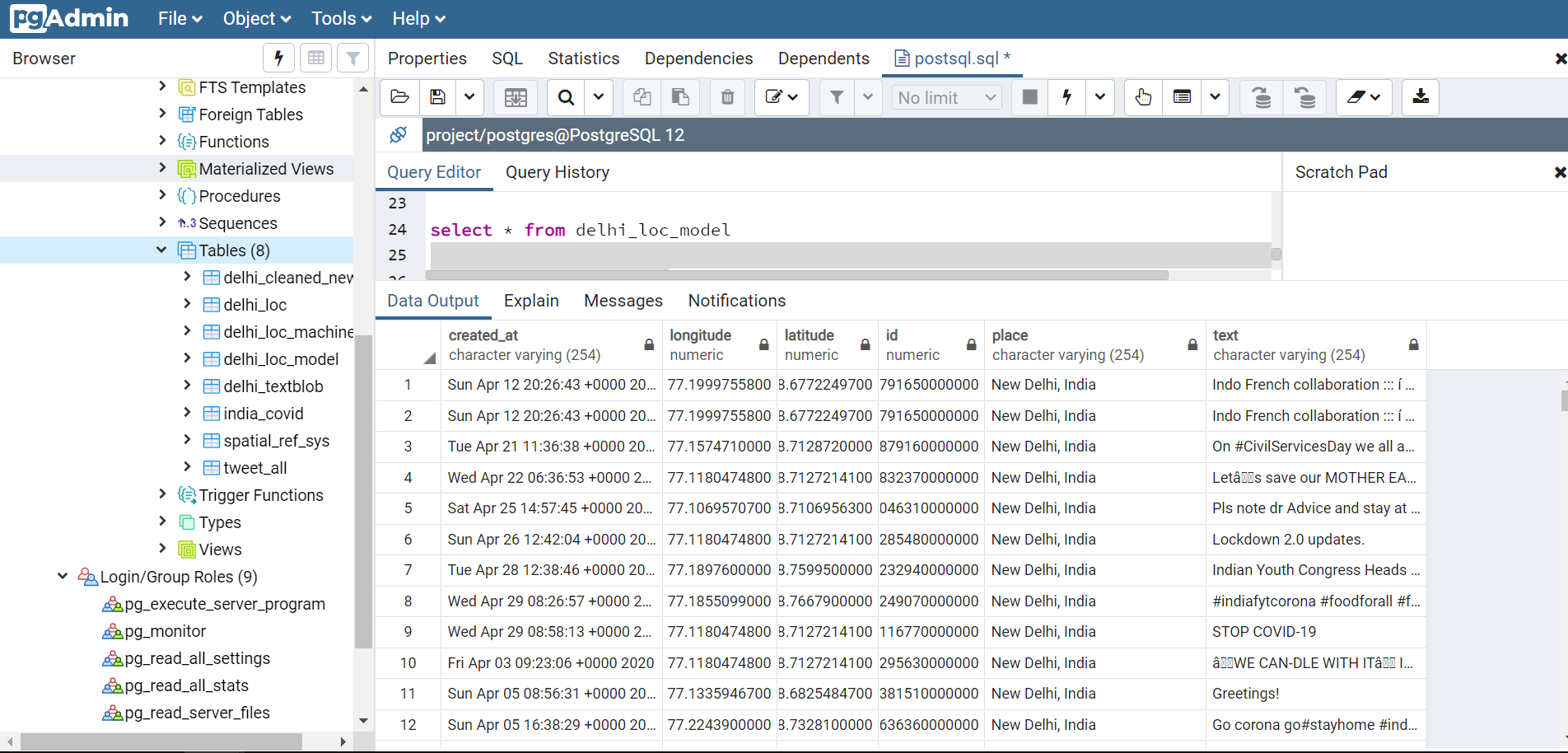


Figure 3: PostgreSQL database showing study area data

Figure 4: PostgreSQL database showing study area data

Figure 4: random Forest resultFigure 5: PostgreSQL database showing study area data

Figure 6: PostgreSQL database showing study area data

Figure 7: PostgreSQL database showing study area data

Figure 8: PostgreSQL database showing study area data

Figure 4: random Forest resultFigure 9: PostgreSQL database showing study area data

Figure 10: PostgreSQL database showing study area data

From the entire world’s data, the information about the study area is extracted using the geodatabase. There are around 3500 data points. Then this geodatabase is connected to the jupyter notebook (python) for the analysis. The text information of the data has various component as emojis, special characters, symbols, multiple languages, and punctuates. This has to be cleaned and converted into a machine-readable language for model fitting. This process has been carried out in Jupyter notebook using various libraries in python.

Once the data is stored in the database, we can divide it for the training and validation. Sentiment labels are applied adjacent to the tweets as **+1**(for positive), **0** (for neutral), and **-1** (for negative). Training and validation play a major role for the machine learning algorithm. The quality of the result depends on the trained model. The database will be integrated with the dashboard platform for visualization.

# **Result and Discussion**

The end result, that is, sentiment of the tweets is generated using three algorithms. We get satisfactory results by all the three methods. There were 3500 data points for the study area. In this, training and model fitting was carried out. Based upon the model, accuracy was generated.

## Random Forest algorithm:

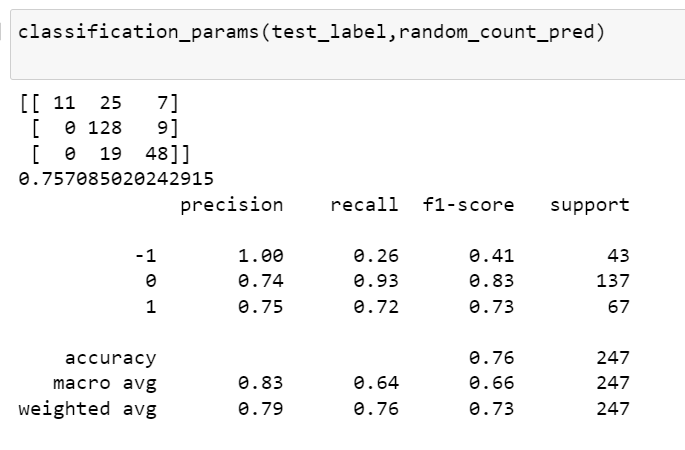
The accuracy of Random Forest Method is 75.70%

Figure 4 - PostgreSQL database showing study area data

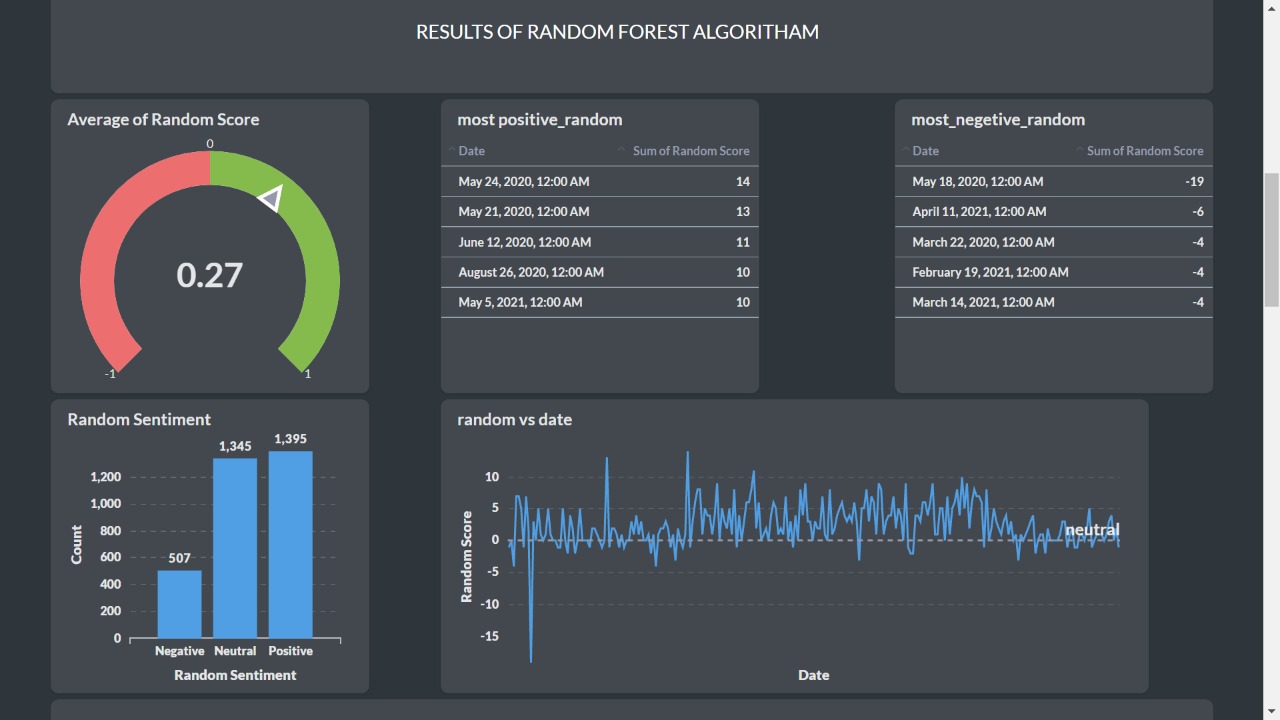


Figure 5-Dashboard of Random Forest algorithm

## Support Vector Classifier (SVC):

### SVC using Linear as kernel**:**

The accuracy of SVC-linear is 74.08%

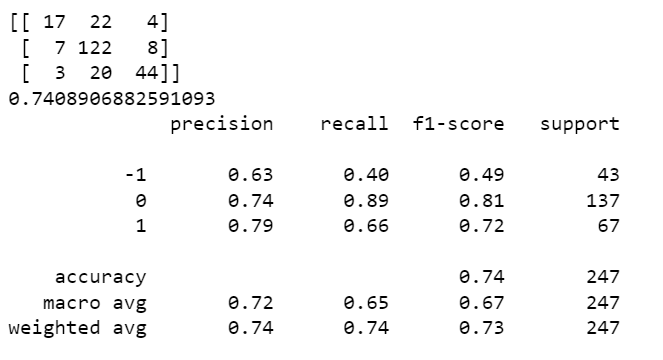


Figure 6 - Result of SVC-linear

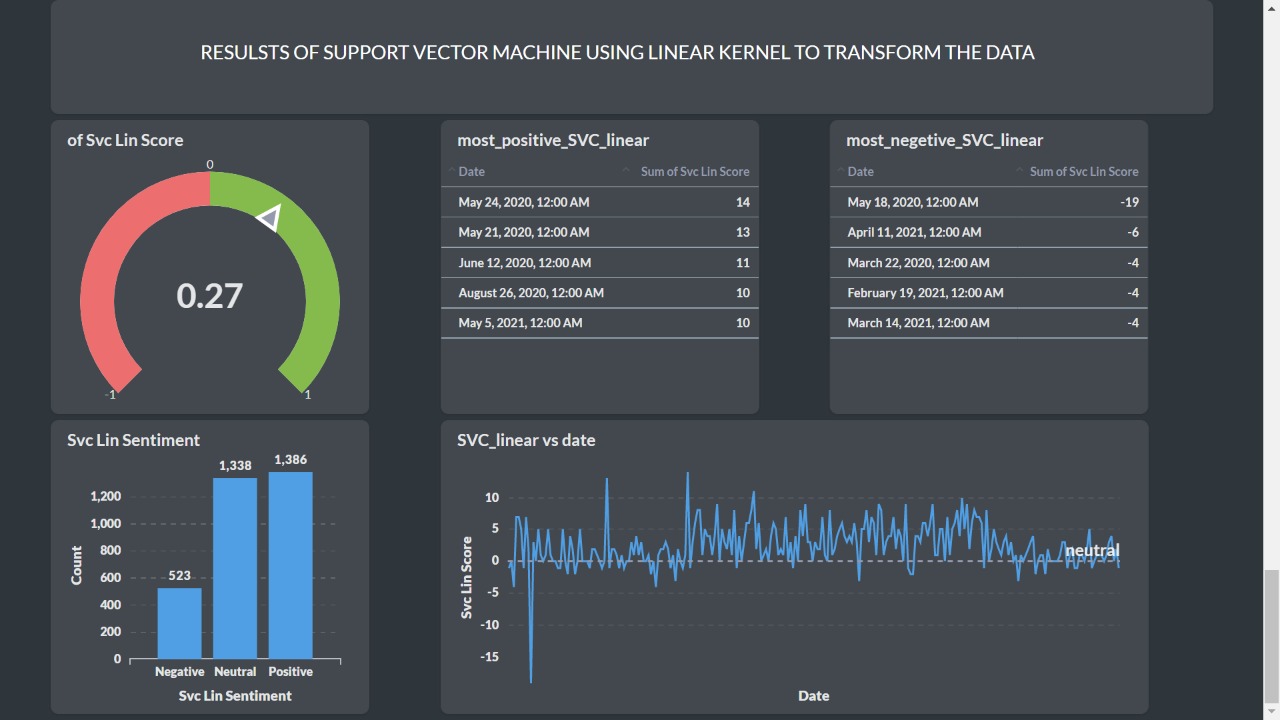


Figure 7- Dashboard of SVC-Linear

### SVC using Radial Basis Function (RBF) Kernel

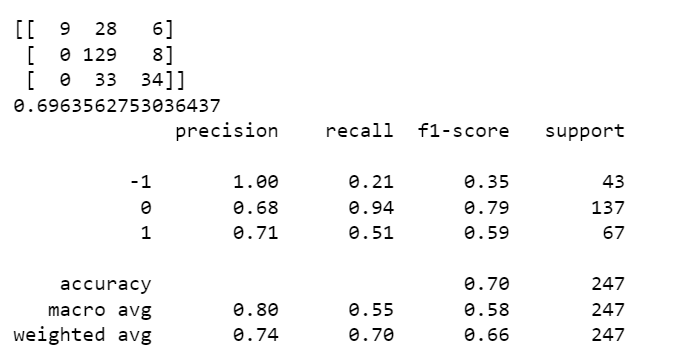


Figure 8 Result of SVC - RBF

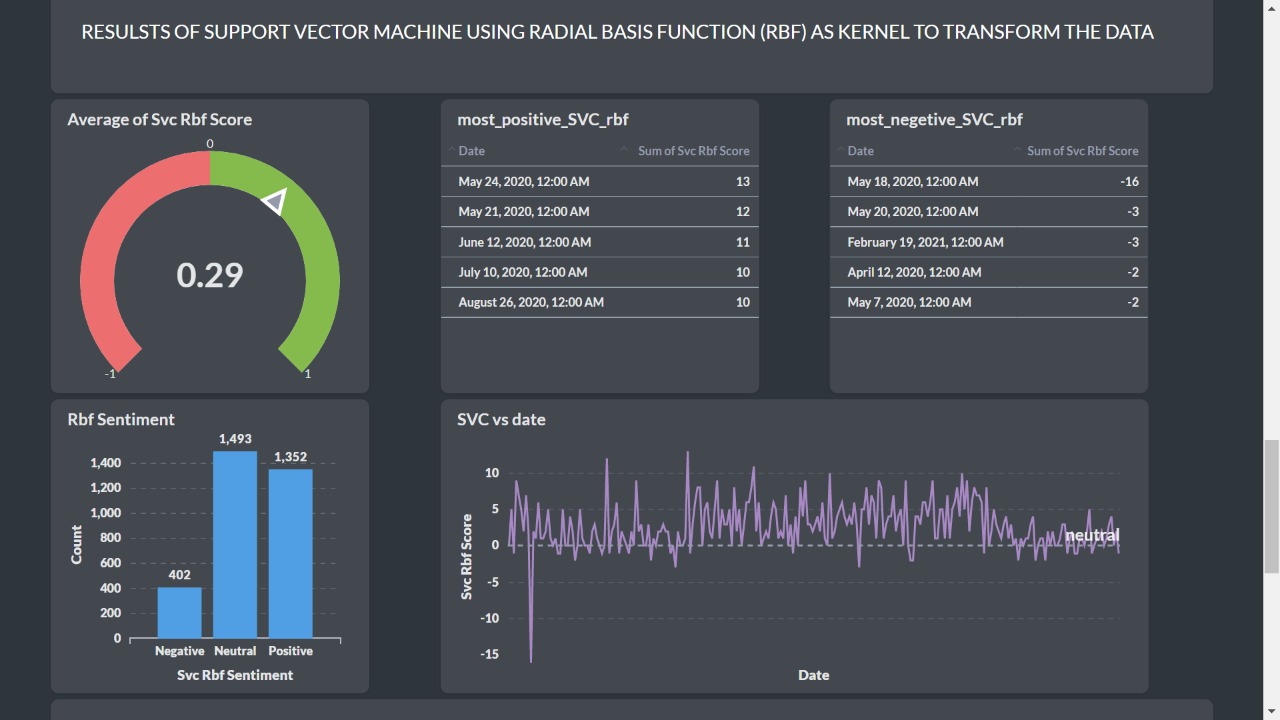


Figure 9 - Dashboard of SVC-RBF

## Naive Bayes:

The accuracy of Naive Bayes method is 67.20%

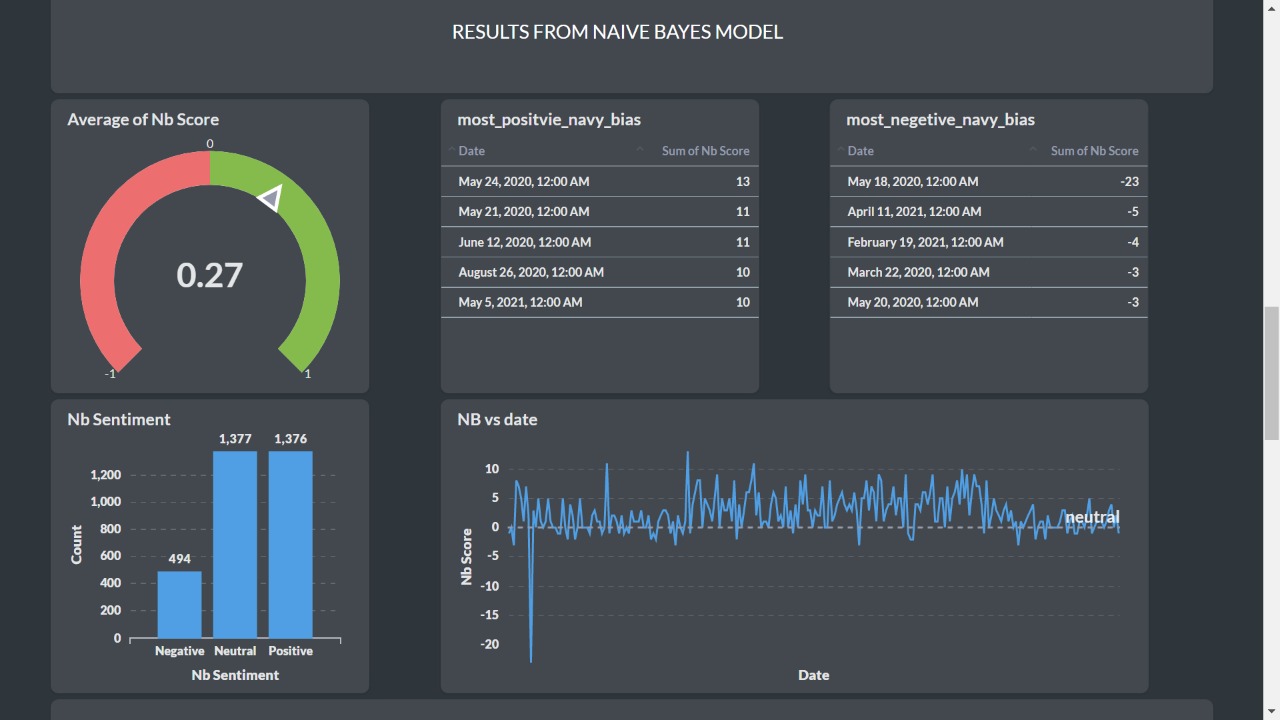
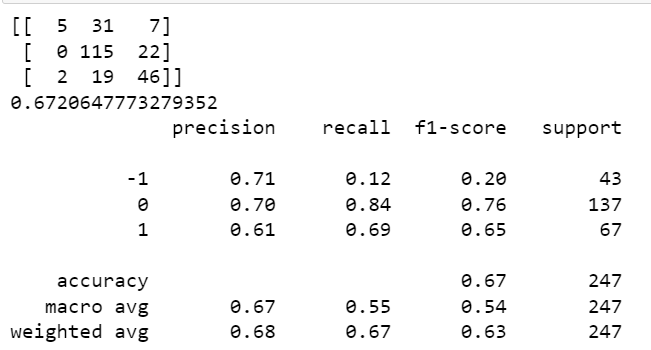


Figure 10 - Result of Naive Bayes

Figure 11- Dashboard of Naive Bayes'

# **Conclusions**

There are many machine Learning algorithms which can be used for the sentiment analysis. In the project, we used random forest, naïve bias, and SVC algorithms for getting the sentiment. All algorithms performed well considering the small size of data. Among all three, random forest gave the highest accuracy of approximately 75%. It has been observed through multiple iterations of the algorithms that accuracy depends on the size of the data. If more samples are given for training, the model gets better.

It has also been observed in the final analysis that most of the tweets in the given timeframe are positive, then neutral and least is the negative in all three algorithms.

# **Contribution of Members**

|  |  |
| --- | --- |
| **Tasks** | **Names** |
| Data Collection and Hydration | Sandeep, Aman, Nanchariah, Gangadhar |
| Literature Review | Dhanasekaran, Aman |
| Data cleaning and Analysis | Sandeep, Aman, Dhanasekaran |
| Dashboard Preparation | Nanchariah, Gangadhar, Dhanasekaran |
| Report and Presentation | All 5 |

[**Git-Hub Link**](https://github.com/sandeepkumar-kandukuri/sentiment-analysis-)

**https://github.com/sandeepkumar-kandukuri/sentiment-analysis-**

# **References**

Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment Analysis of Tweets using SVM. *International Journal of Computer Applications*, *177*(5), 25–29. https://doi.org/10.5120/ijca2017915758

Kanakaraj, M., & Guddeti, R. M. R. (2015). NLP based sentiment analysis on Twitter data using ensemble classifiers. *2015 3rd International Conference on Signal Processing, Communication and Networking, ICSCN 2015*. https://doi.org/10.1109/ICSCN.2015.7219856

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

Narayanan, V., Arora, I., & Bhatia, A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8206 LNCS*, 194–201. https://doi.org/10.1007/978-3-642-41278-3\_24

Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). Sentiment Analysis in Social Networks. In *Sentiment Analysis in Social Networks*.

R, S. E. (2021). Understanding Random Forest. Retrieved May 4, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random forest is a Supervised,average in case of regression

What is PostgreSQL? (2022). Retrieved April 5, 2022, from https://www.postgresql.org/about/