SMART MONITORING OF THREE FARM LAW USING TWITTER

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering

By

Vardhan Agarwal (181255)

Under the supervision of

Mr. Surjeet Singh

To

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

CANDIDATE'S DECLARATION

I hereby declare that the work presented in this report entitled "**Smart Monitoring Of Three Farm Laws Using Twitter**" in partialfulfilment of the requirements for the award of the degree of Bachelor of Technology inComputer Science and Engineering/Information Technology submitted in the department ofComputer Science &; Engineering and Information Technology, Jaypee University of InformationTechnology Waknaghat is an authentic record of my own work carried out over a period from January 2022 to May 2022 under the supervision of **Mr. Surjeet Singh,Assistant Professor(Grade II).**

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Vardhan Agarwal
181255

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Mr. Surjeet Singh
Assistant Professor(Grade II)
Computer Science & Engineering
Jaypee University of Information Technology, Waknaghat

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the project work successfully.

I am really grateful and wish my profound indebtedness to Supervisor **Mr. Surjeet Singh, Assistant Professor(Grade II)**, Department of CSE, Jaypee University of Information Technology, Waknaghat. Deep Knowledge & keen interest of my supervisor in the field of "**Machine Learning**"to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to **Mr. Surjeet Singh,**Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Vardhan Agrawal

Table of Contents

LIST OF ABBREVIATIONS

Py	Python
Np	Numpy
Pd	pandas
NLP	Natural Language Processing
Pck	pickle
DTM	Document Term Matrix
TS	Transcripts
BS	Beautiful Scope
STL	ScrapsFromTheLoft
SW	Stop words
TW	Top words
SA	Sentimental Analysis
TM	Topic Modelling
TG	Text Generation

LIST OF FIGURES

ABSTRACT

Nowadays, the government across the globe is becoming more and more dependent on the opinion of the public for formulating and implementing the various laws and policies towards the wellbeing of the common man. Social media plays a significant role in this upcoming trend. Back in the day, the public's lack of participation for policy making decisions turned out to be a significant hurdle particularly during the formulation and evaluation of such policies. But with time, the tremendous rise in social media platform usage by the common public has turned out to be a wider insight for the government to overcome this long pending challenge. E-governance based on emerging technologies is being put to use due to availability of IT infrastructure along with changes in mindset of advisors of government in realizing the numerous policies in the most favourable manner.

The project presents an approach that is efficient in binding the capabilities of bsocial media and cloud towards efficient analysis of governmental policies through involvement of the public. This project comprises a collection of dataset using Twitter APIs and thereafter cleaning of this dataset. Further with the help of various analysis algorithms we cover many aspects of public opinion. Firstly we implement sentiment analysis. Wherein we extract the sentiment from the dataset and formulate meaningful observations from it. Secondly, we perform topic modeling which helps in identifying the main topic from the tweets captured. For this we use the Latent Dirichlet Allocation algorithm; it takes into consideration each document as some topics in a fixed proportion. And each topic comprises some keywords, in a fixed proportion. Then we apply Geo-Location Analysis which gives us the idea about location based public opinions. The method used has provided us with good results,the test was done for public opinion on the Three Farm Bills issued by Indian government and in return the retaliation by the farmers is known as Farmers' Protest. It was hence established that it can be used for efficient analysis of policies .

Chapter 01: INTRODUCTION

1.1 Introduction

Back in the day, the public was not given the chance to be a part in the decisions for making of policies made by the government and therefore the administrators had their own say. This prompted a sharp decrease in conviction just as dependence of general society towards the government by and large and its strategies specifically. The discernment fracture between the government and general society expanded dramatically and right now both are attempting to keep an ideal concordance concerning their relationship. Electronic-Government (E-Government) is an influential practice that has all the credentials to have a better relation between the public and the policy makers as it devotes greater importance on having a transparent system and increasing the public participation. To solve this, governments everywhere have started using social media for getting a fair assessment from different parts of the country regarding the effectiveness of fresh reforms being formulated from the general public's point of view. With a large number of users actively participating on social media, the data volume generated is very large. The common systems are not enough for handling this amount of data, due to the need of an infrastructure to gather and formulate this much data. This project proposes an efficient way to unite the capabilities of the emerging technologies like cloud computing and social media analysis to analyse government decisions and agenda for the common public.

DS Model (Data Science Model):

Programming: That's computer and computer science and knowing how to code. This is the most basic skill required by the data scientist.

Math's and Stats: This includes some linear algebra and calculus some statistics used to solve the maths algo and machine learning problems

Communication: After we done all that number crunching and coding can we wrap it all together in a story and communicate our insights and there's this one part here that we just want

to mention that is the danger zone of data science so if we are really good at programming and we also have communication skills, but we don't have the math and stats background this is what we call the danger zone

And all three depend on each other directly and indirectly. And we studied on them in brief with the help of calculations and the visualizations in this project

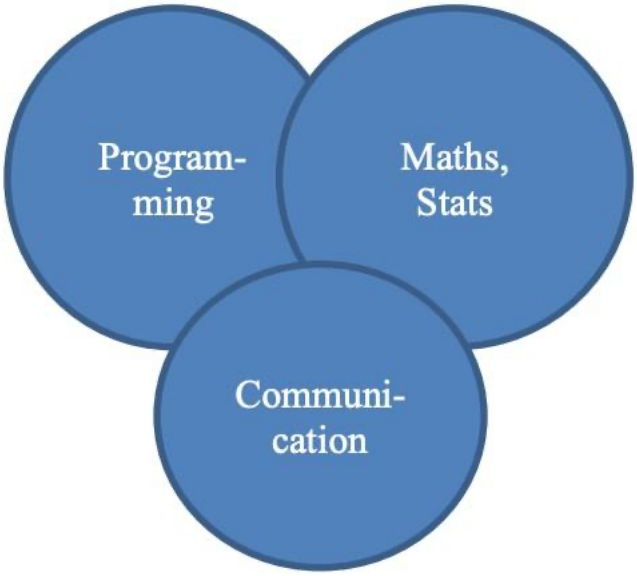

Fig. 1 DS Model

1.2 Objective of the Major Project

The main objective of this project is to analyse the opinion of the public for different government agendas. A lot of people these days are engaging in political discussion on social media leading to huge amounts of data being generated. This calls for a need for a cloud system which can make use of this huge amounts of data and collect some useful opinion of the public in meaningful format of proposals, their questions, effective solutions, issues and merits and demerits aimed towards a public reform at

an initial stage where change can be made so some steps can be taken towards pleasing the public for whom this policy is being made. This project seeks to unitize the advantages of cloud computing and social media analytics for effective examination of the policies. This project will enable us to answer the following question and draw meaningful conclusions:

Can social media platforms make up as a good tool for opinion analysis on important national level issues?

1.3 Motivation of the Project

Governments everywhere are thinking of ways to promote a transparent system to make it easy for its citizens to flourish under them. E-Governance is a mechanism that allows the government to carry out its every task and deliver a good service to its people. E-Government makes use of IT innovations to grant full access to efficient, quick and good public service to employees and its people. This requires skilled people to be implemented and also requires a good infrastructure for the services to be provided .Cloud computing can be used to address these challenges surrounding the masses with greater degree of satisfacton. Cloud computing involves usage of a large pool of computer resources that enables features like pay-as-use plus on-demand scalability to add a few. These advantages play a very decisive role in inspiring the policy makers of various big countries to make easier switch from the more orthodox and costly E-Government to a more efficient and cost effective cloud computing based E-Governance mode for better services.

Generally, making of the policies is mostly based on official statistical data which is formulated by agencies of government and international bodies all over the world. However, decision makers generally report flaws in such forms of statistics as sometimes there is publication delay in the process, top down approach and less interest in the topic etc. So for overcoming such a traditional problem of collection of data, political scientists and makers of policies moved in favour of data driven platforms like

Twitter and Facebook, to generate further trustworthy data that is collected in real time. Therefore, these platforms has become an effective tool for greater transparent working of the government and also decreasing the communication barrier between policy makers and citizens by producing more authentic statistical data. In the present generation, social media has become a very important part in everyone's life, not taking into consideration the individual's status. These online platforms act as an effective place for all the people of the world to discuss various agendas of common interests such as entertainment, sports and even political affairs. Taking politics into consideration , a healthy 33% of social media accounts comment, post and discuss politics on social media platforms. Even the government has come to realize the potential social media holds in the present world. Consequently, various agencies of the government have started seeking help of these online platforms to engage and connect to the common man. It has helped immensely in increasing the communication between the government and the public, it is trying to increase the participation of the public. In the modern era, people have different views regarding government policies and therefore share their thoughts on the same which in turn many a times turns into an agenda. This information helps the government for formulation of more effective policies for the public and framing and conveying better services while taking into consideration the opinion of its citizens.

1.4 Language Used

This project is designed using Python language. The selection of Python was based on the availability of numerous inbuilt libraries, packages and functions. It's syntax is simple, concise and clear. Python provides a direct tweepy package to directly handle twitter API. It has a simple graph plotting function like matplotlib to handle a vast variety of graphs.

1.5 Technical Requirements

The project is carried out in Python programming language and the data is stored on Google cloud platform. The project makes use of various data analysis libraries like numpy, pandas, matplotlib and pyLDavis . The algorithms are implemented by making use of libraries like textblob and gensim. The data is extracted by making use of tweepy API and then preprocessing it for further analysis.

1.6 Deliverables of the Project

The project makes use of data analysis to get the opinion of the people.The project gives the following results:

- Analysis of opinion of the people regarding a specific policy by telling the positive, negative or neutral response about the policy.
- The major topics which are identified in the theme and hence giving an idea about what things should be focused on regarding a specific agenda.
- Get specific insights of a certain region by making use of location to help the local government to get an idea of what people expect from the government.

CHAPTER 2: LITERATURE SURVEY

S.No.	Literature	Topic Discussed
1.	Purohit, H., Hampton, A., Shalin, V. L., Sheth, A. P., Flach, J., & Bhatt, S. (2013). What kind of# conversation is Twitter? Mining# psycholinguistic cues for emergency coordination. Computers in Human Behavior, 29(6), 2438–2447.	Data science, data science solution, big data.
2.	Y. Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Medicine, 2, no. 10, e267	Data visualization, visualization, data analysis,databases.
3.	Yi, S., Li, C., & Li, Q. (2015). A survey of fog computing: Concepts, applications and issues. In Proceedings of the 2015 workshop on mobile big data (pp. 37-42). ACM.	Optimisation,Big data, data visualization, data mining.

4.	Stieglitz, S., & Dang-Xuan, L. (2013b). Social media and political communication: A social media analytics framework. Social Network Analysis and Mining, 3(4), 1277–1291.	Data mining, clustering algorithms.
5.	Severo, M., Feredj, A., &Romele, A. (2016). Soft data and public policy: Can social media offer alternatives to official statistics in urban policymaking? Policy & Internet, 8(3),	Comedian segmentation, clustering, data mining, pattern clustering.
6.	Jha, R. (2018). Regional inequality and indirect tax reform in India. In Facets of India's economy and her society volume II (pp. 119–148). London: Palgrave Macmillan.	Unsupervised machine learning, customer segmentation,spendingbehavior, data models
7.	S. IBM (2019) [Online], Available: h from the scrpas from the loft IEEE (pp:8975) , Ref -09873	Comedian Transcripts , customer segmentation, predict algorithms,logistics

8.	Yi Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). Polarization and acculturation in US election 2016 outcomes–can twitter analytics predict changes in voting preferences. Technological Forecasting and Social Change. https://doi.org/10.1016/j.techfore. 2018.09.009.	Demand response, data mining, load management,feature extraction

2.1 Proposed system

The main proposed system that we have used here is more focused on the new tools and technologies introduced in the past few years. Data science, Machine Learning, Deep Learning, Artificial Intelligence, Natural Language Processing, and other technologies are among them.

We have followed the data science process of initially finding the dataset, analyzing it and employing the model we developed after training and testing to fuel the growth of any business In the proposed system, instead of just letting the humans do all the work, most of the analysis is automated which is being done by Jupyter Notebook, Python libraries that we imported like Keras, tensorflow, numpy, panda, scikit, sklearn, and using these libraries to do automation and build a model to visualize how the change is happening in the given dataset.

2.2 Feasibility Study

Any vital stage in the transcript improvement process has been procured. Permits engineers to get a functioning item that has been tried. Alludes to item exploration that may be done as far as transcripts results, application execution, and specialized help expected to utilize it. A potential examination ought to be completed dependent on an assortment of conditions and conditions.

CHAPTER 3: SYSTEM DEVELOPMENT

3.1 Design and development

In any case, most importantly, why do we do division?

Since you can't treat each user the same way with a similar choice and views. They will find another choice which comprehends them better. Some methods we used in this project are the Sentiment Analysis, Topic Modeling and Text Generation etc.

- Sentiment Analysis: Let’s say you are a manager of a company that sells hats and also shirts and you want to know what your customers are thinking about your hats and shirts do they have positive feelings about them or negative feelings about them so then you go to your call center and you see that a bunch of people have called in without your hats and your shirts and you could go through all of these and listen to every single message but that would take you a really long time so instead you can use an NLP technique to automatically tag these as positive or negative calls and then at the end of the day you can figure out that people tend to think that your hats are pretty good and that your shirts are not very good so this concept is called sentiment analysis. In this we use text blob sentiment analysis in which we give polarity to the word in range of -1 to +1 in order to decide which one of them are negative and which are positive.

- Topic Modeling: The task of discovering themes that best characterise a set of documents is known as topic modelling. Only throughout the topic modelling process will these themes arise (therefore called latent). Latent Dirichlet Allocation is the topic modelling technique (LDA) we used in our model.

There are some other methods also that we have used.

Classification: Classification is the process of identifying a function that aids in the classification of a dataset based on several factors. A computer programmed is trained on the training dataset and then categorizes the data into distinct classes based on that training. The classification algorithm's goal is to identify the mapping function that will convert the discrete input(x) to the discrete output(y) (y).
Algorithms for classification can be further classified into the following categories:

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machines
- Kernel Support Vector Machine
- Naive Bayes
- Decision Tree Classification

Naive Bayes- It is a classification algorithm that may be used to classify binary and multiclass data. It is a supervised classification technique that uses conditional probability to assign class labels to instances/records in order to categories future objects. It plays an important role in this project.

Regression: The technique of discovering correlations between dependent and independent variables is known as regression. It aids in the prediction of continuous variables such as market trends, house values, and so forth. The Regression algorithm's goal is to identify the mapping function that will translate the continuous input variable (x) to the discrete output variable (y) (y).

Regression Algorithm Types:

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial Regression

- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

3.2 Algorithms

Machine learning algorithms which we've used are:

Linear regression: Linear Regression is a simple machine learning algorithm that is used to solve regression problems and falls under the Supervised Learning technique.
It is being used to anticipate a measured process variable using control variables. Linear regression is used to find the best-fit line for predicting the outcome of a continuous variable. When only one regression analysis is used, simple linear regression is being used. If there are more than two types of variables to predict, the Multiple Regression Model will be used. By picking the optimal fit line, the algorithm sets up the correlation between the dependent as well as relationship between the independent variable.

Among many only well Machine Learning techniques used during Directly controlled Learning approaches is logistic regression.. It can be used to solve both classification and regression problems, however classification is the most typical application. Logistic regression is used to predict the categorical dependent variable using independent factors. A Logistic Regression problem can only have two possible outcomes: 0 and 1. When calculating the probability between two groups, logistic regression can be used. For example, whether it will rain today or not, whether it will rain today or not, whether it will rain today or not, true or untrue, and so on. Probabilistic prediction is often used in logistic regression. In this case, the observed data should be considered the most plausible. In logistic regression, we pass the weighted sum of inputs through an activation function that can transfer values between 0 and 1. A type of activation

function is the sigmoid function.K-means clustering-based unsupervised machine learning method Using K-Means for Cluster analysis

3.3 Model Development

Data wrangling: The process of cleansing and integrating chaotic and complicated data sets for easy access and analysis is known as data wrangling. With the amount of data and data sources continuously increasing and expanding, it is becoming increasingly important to organize vast amounts of data for analysis. In most cases, this procedure entails individually converting as well as mapping data through one numerical form to the other in order to facilitate data consumption and association.

The goals of Data Wrangling are to accumulate information from diverse sources in revealing "profound intellectual capacity." Reduce the time it takes to collect and organize unorganized data before it can be used. Allow data scientists and the data analysts to concentrate on data analysis rather than the data wrangling. Senior executives in an organization should be encouraged to improve their decision-making skills.

Crucial Steps in Data Wrangling

- Data Acquisition: Locate and gain access to the information included in your sources.
- Data integration is the process of combining altered data for the future assessment including using.
- Data cleansing entails reorganizing the data into a more useful and functional manner, as well as correcting or removing any incorrect information.

Feature engineering: When developing a predictive model using machine learning or statistical modeling, feature engineering refers to the process of leveraging domain expertise to choose and convert the most important variables from raw data. The purpose of feature engineering and selection is to make machine-learning (ML) algorithms perform better. The construction, transformation, extraction, and selection of features, also known as variables, that are most conducive to constructing an accurate ML algorithm are all part of feature stuff

Hyper parameter tuning: A mathematical model containing a number of parameters that must be learned from data is referred to as a Machine.

Hyper parameters, on the other hand, are a type of parameter that cannot be learned directly from the standard training procedure. They are normally fixed prior to the start of the training procedure. These parameters describe crucial aspects of the model, such as its complexity and learning rate.

EDA (Data Exploration Analysis): EDA is a data assessment strategy that employs a variety of (mostly diagrammatical) methods to optimize comprehension of a data set. This apart from fitting the infrastructure to available information, we can fit the same parameters of the model. by building the classifier with existing data Recognize underlying structure, extract significant factors, detect outliers and anomalies, exam fundamental assumptions, construct parsimonious models, and identify the optimal factor settings.

3.4 Requirements on Major Project

3.4.1 Functional Requirements

- The project provides an analysis of the tweets for a given topic.
- The twitter api used has to be authenticated by using an access token and consumer key provided by the developer account of twitter.
- The sentiment analysis should provide results in the form of a pie chart which states polarity as positive, negative or neutral.
- The topic modeling should result in a graph which provides the main topics of the tweets regarding a particular then show it using visualization methods.

- The location analysis of a particular region and sentiment of people belonging to a certain place.

3.4.2 Non-Functional Requirements

- The authorization key should be kept a secret to prevent misuse of app and regenerated if being used.
- The LDA needs the initializing of hyperparameters by the user for most efficient results.
- The tweets for location analysis should be collected from locations with high population density to get more accurate results.

3.5 Use Case Diagram of the Project

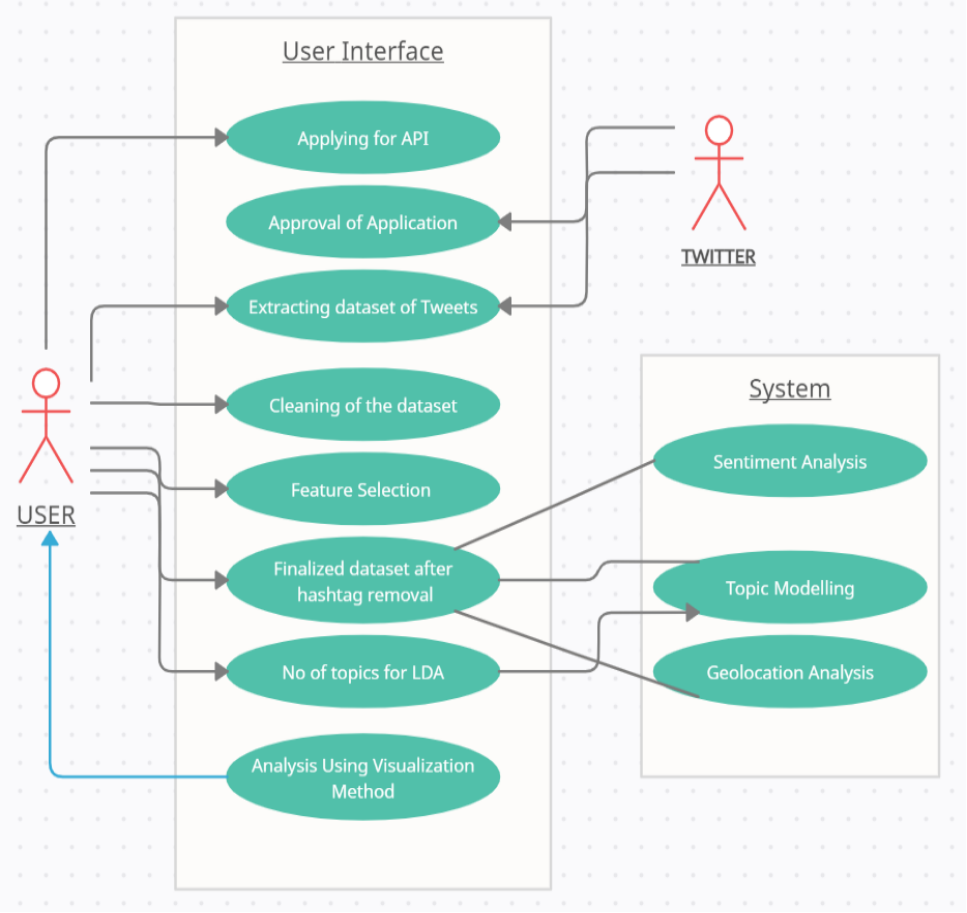

Fig. 2 Use case diagram

3.6 DFD Diagram of the Project

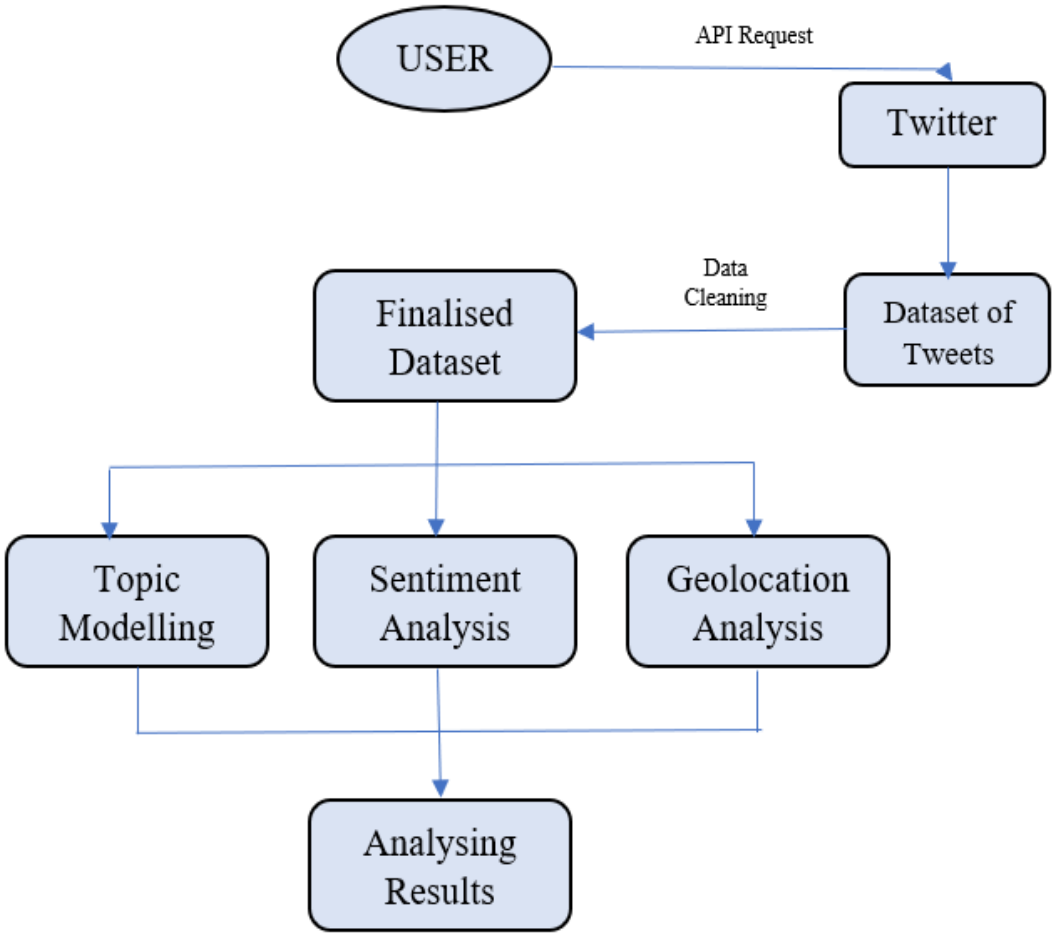

Fig 3. Data Flow diagram

3.7 Technologies used

- Anaconda : Jupyter Notebook
- Python Programming language
- Data science process
- Libraries like numpy, pandas, scikit, matlab, pyplot etc.
- Machine learning models/algorithms

Libraries and the packages used in bit brief-

Numpy: NumPy is a Python module which allows users to interact with the arrays. Numpy even has capabilities for trying to deal with algebraic expressions but also linear advanced mathematics. NumPy is referred vas the Numerical python.

Why NumPy:

We have lists in Python those acts like the arrays, however they are pretty slow to process. NumPy intends to deliver a 50-fold quicker array object than ordinary Python lists. NumPy's array object is called ndarray, and it has a lot of features.of helper functions to make working with it a breeze. In data research, when speed and resources are critical, arrays are widely employed.

Pandas: Pandas is one of the most popular and well-liked data science tools for wrangling and analyzing data in the Python computer language. In the real world, nowadays data is inherently messy. When it comes to cleaning, transforming, manipulating, and analyzing data, Pandas is a game changer. Pandas, basically, assist in the cleanup of the mess completely.

Matplotlib: Matplotlib is a Python graphical interface as well as diagrammatical plotting package which is a statistical enhanced version NumPy which keeps running. As a result, it provides an open source alternative to MATLAB.

Seaborn is a Scripting language visualisation kit that's also premised on matlab. It has a highlevel interface for creating visually appealing and instructive statistics visuals.

Plotly: Plotly allows the users to study and visualize the data by importing, copying and pasting, or streaming it. . Plotly allows you to save, share, and collaborate on Python scripts.

Datetime: The datetime module is used for the manipulation of the dates and times.

Sklearn(Computational tool) is by far the most functional and reliable pattern recognition open source Library. It makes advantage of a Python consistency interface to provide a collection of machine learning capabilities. And statistical modelling, such as classification, regression, clustering, and dimensionality reduction.

Keras: Basically in the predicting sales module we concentrated on the Long Short-term Memory (LSTM) approach, which is a prominent Deep Learning method. In order to implement LSTM in our project, we used Keras. Kera's is a Google-developed high-level deep learning API for implementing neural networks. It is built in Python and is used to make neural network implementation simple. It also allows for the computation of numerous neural networks in the backend. Tensorflow is one of the frameworks that Keras supports.

Tensorflow: It's a free artificial intelligence programmed that creates models using data flow graphs. It enables programmers to build large-scale neural networks with multiple layers. Some of the best uses of tensorflow are Classification, understanding, discovery, predicting and creating.

Chapter 04 : PERFORMANCE ANALYSIS

4.1 Data Set Used in the Project

In the light of this protest, social media users have been very active in voicing their opinion about the matter. "#FarmersProtest" is very prevalent on Twitter, with thousands of users tweeting thousands of tweets with the hashtag.

The data for this project has been collected from Twitter API known as Tweepy. Although the project has the capacity of making use of various platforms of social media for data collection , but in this scenario we made use of Twitter data only. As stated previously the data was fetched based upon specific agenda (FarmersProtest, FarmBills). The Tweepy API used for collection of tweets provided us with different parameters for searching such as, tweet's language , type of the tweets and date range in which the extraction operations using API is being carried out. Using these kind of searching parameters, fetching operation can be carried out in a more efficient and effective way.

As the data is fetched by making use of APIs, the data is returned in an unstructured manner i.e. json format. For this data to be useful in further studies it is required for us to convert this json format into structured csv format.

4.2 Date Set Features

4.2.1 Types of Data Set

The dataset is in the form of text which is extracted from the twitter api and then various text analysis techniques are applied to it to extract the information.

4.2.2 Number of Attributes, fields, description of the data set

The twitter api returns the data in form of json format which contains 18 sets of attributes in it.

In this project we made use of 3 attributes which were as follows : the id of the tweet, the text of the tweet and the place_id of the tweet. These attributes helped us to perform sentiment and location based analysis on the tweets collected.

The screenshot demonstrates how a tweet looks in a json format.


```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter AP
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news, updates & event
  },
  "place": {
  },
  "entities": {
```


Fig 4. Tweets in json format

The following screenshot shows the dataset for some tweets after processing and performing sentiment analysis on it.

RT @Chanpre87917253: ????@@ MODIJI WANTS TO SNATCH FARMER'S LAND BY THESE LAWS & GIVE IT TO ADANI AMBANI & OF COURSE #BRITISHERS WHO IS...	positive
Never cease to amaze us A party having only leaders from one family, which only wants to worsen the situation may i... https://t.co/8zQnn8A4EY	positive
...in Ibadan, said ranching remains the most viable solution to the herder-farmer clashes in Nigeria. The forum al... https://t.co/mBuT8BSICO	positive
Never cease to amaze us A party having only leaders from one family, which only wants to worsen the situation may i... https://t.co/xLyYoTtkQt	positive
RT @shaunattwood: Maxwell, Bill Gates & Leon Black: Ryan Dawson https://t.co/UtEH9v9SDV via @YouTube My latest #epstein video with Ryan Da...	positive
@people_stfu Bill ke baare mai kuch paata nhi or aajate hai muh utha kr. Koi ek modi ke against kuch bhokta hai to... https://t.co/BJawiXeH6O	neutral
RT @Chanpre87917253: @#@# MODIJI WANTS TO SNATCH FARMER'S LAND BY THESE LAWS & GIVE IT TO ADANI AMBANI & OF COURSE #BRITISHERS WHO IS BE...	positive
RT @Chanpre87917253: @*#*@* MODIJI WANTS TO SNATCH FARMER'S LAND BY THESE LAWS & GIVE IT TO ADANI AMBANI & OF COURSE #BRITISHERS WHO IS...	positive

Fig 5. Text of tweets after cleaning

4.3 Design of Problem Statement

Governments throughout the globe are responsible and answerable to their citizens and society at large, since they are responsible for preserving the interests of the common public.

Due to the rising citizen prospects and the need for innovation in ruling policies of the government, social media has become an important component of electronic government in a very short span of time.The government in some places has evolved and engaging with general public via social media,some governments have also started campaigning virtually as they understand how active the general public is on such platforms.

The problem statement of this project "Smart Monitoring of Three Frams Laws Using Twitter. " is designed in such a manner to bring active participation of the people of the country. It involves collection of the views of the people on some particular government policy through social media. Then through various analysis techniques an overall result is framed. This will tell us whether the policy is going to be a success among the people or a failure.

4.4 Algorithm / Pseudo code of the Project Problem

Pseudo code:

Step 1: Calling Twitter API to download a dataset of tweets.
Step 2: Cleaning of the downloaded dataset.
Step 3: Storing the dataset on a cloud based platform.
Step 4: Applying feature selection on the dataset.
Step 5: Applying Sentiment Analysis on the cleaned dataset.
Step 6: Plotting results of this analysis as a pie chart
Step 7: Applying Topic Modelling on the cleaned dataset.
Step 8: Selecting a suitable number to break topics into.
Step 9: Using Latent Dirichlet Allocation (LDA) to extract the major topic.
Step 10: Performing Location based Analysis on the cleaned dataset.
Step 11: Analysing the results using various visualization methods to get an idea of how the public feels about the given policy.

The main algorithms used in the projects are sentiment analysis and topic modeling by making use of LDA , these are as follows:

Sentiment analysis:

- Tokenization - The data is divided into individual words. This process is known as tokenization.
- Data Cleaning - The special characters like ! and ? are removed as these are not needed to analyse for sentiment . Stop words like ‘the’ was and ‘him’ are removed to make the analytics more efficient.
- The words which are remaining are then classified into positive , negative and neutral.

  This is done by using pre trained libraries like textblob. These libraries are trained on a large set of words by making use of neural networks.
- The result gives a percentage whether a text is positive, negative or neutral and the one with the highest percentage is declared as result.

LDA

- The parameters of the LDA are intialized ,these include the dictionary that has to be the number of topics on what to get and also the number of times the algorithm should iterate.
- The words are then assigned the topics randomly.
- The algorithm then iterates again and again and then checks whether the word in the document occurs in which topic and dictionary and then rectifies the wrong allocations again and again.
- The iteration is carried again and again until the topics make sense to get the results.

4.5 Flow graph of the Major Project Problem

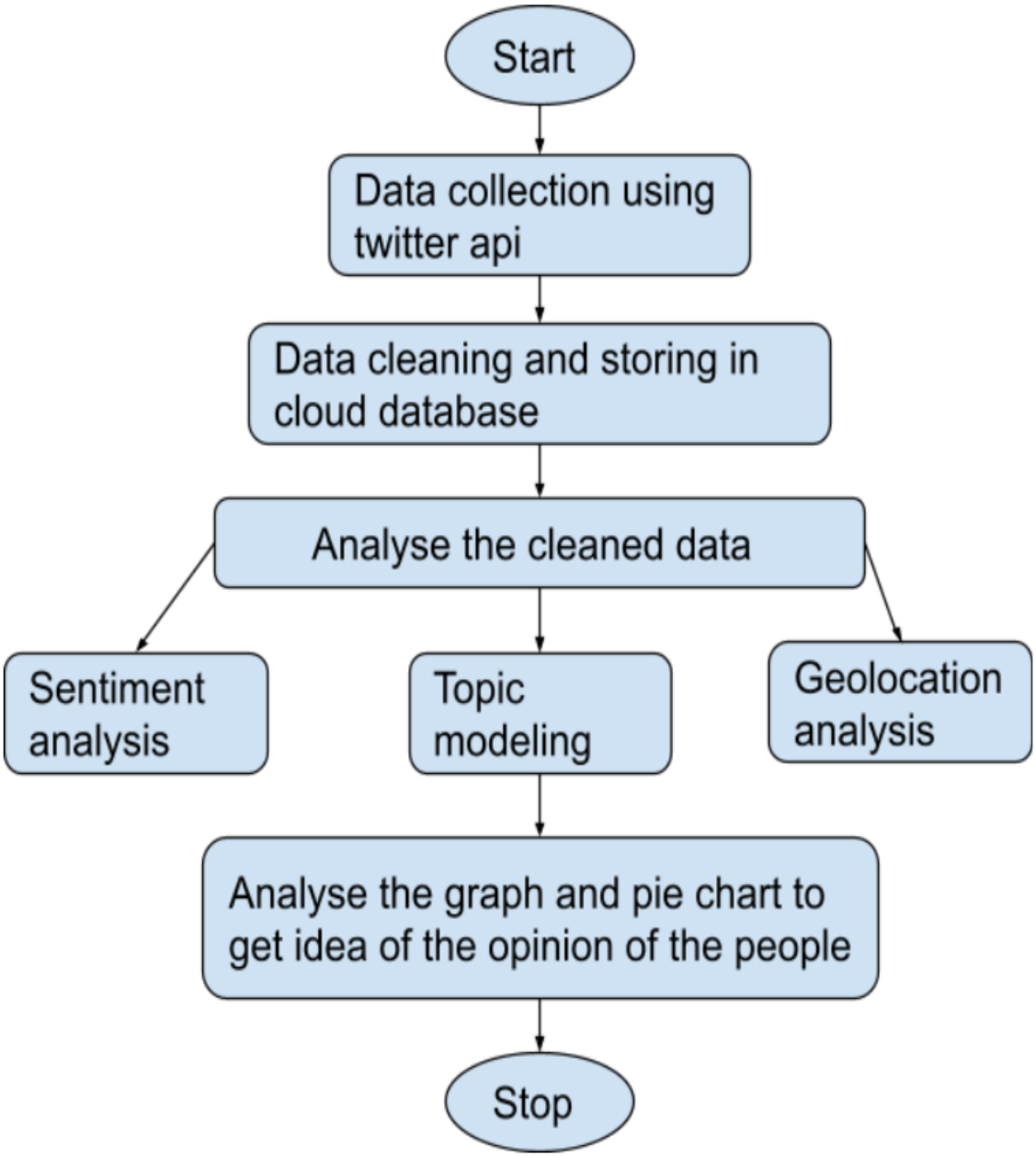

Fig 6. Flow chart

4.6 Outputs At Various Stages

The different stages of the project are as follows:

- The packages to perform the operations are imported. The tweepy API used requires authentication for making use of it.The following image demonstrates the authenticationprocess.

```
try:
    # create OAuthHandler object
    auth = OAuthHandler(consumer_key, consumer_secret)
    # set access token and secret
    auth.set_access_token(access_token, access_token_secret)
    # create tweepy API object to fetch tweets
    api = tweepy.API(auth)
    print("Authentication successful")
except:
    print("Error: Authentication Failed")
```

Fig 7.. Authentication phase

- Since the data obtained is in json format and needs cleaning the tweets are passed through a function containing a regular expression which performs the cleaning procedure then stores the needed attribute in a dictionary . The cleaning process is demonstrated below:

```
#function for cleaning of tweets
def clean_tweet(tweet):
      return ' '.join(re.sub("(@[A-Za-z0-9]+)|([^0-9A-Za-z \t])|(\w+:\/\/\S+)", " ", tweet).split())

def get_tweets(fetched_tweets):

       # empty list to store parsed tweets
       tweets = []
       try:

           for tweet in fetched_tweets:
               # empty dictionary to store required params of a tweet
               parsed_tweet = {}


               parsed_tweet['text'] = tweet.text
```

Fig 8. Data pre-processing phase

- After the data preprocessing sentiment analysis is applied and results are classified into positive ,negative and neutral categories.

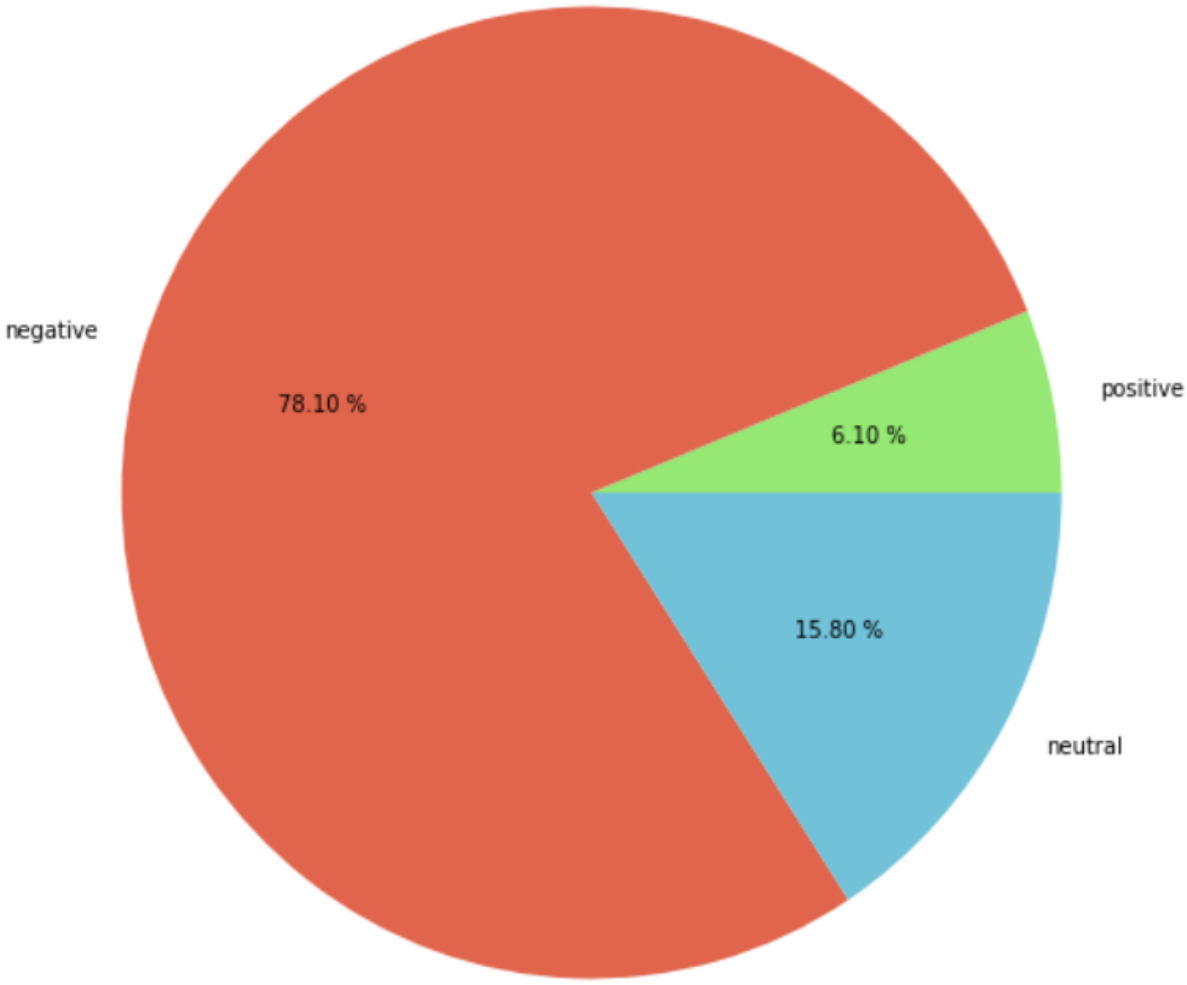

Fig 9. Sentiment analysis of tweets

- Followed by this we perform topic modeling on the data which has been preprocessed and for this we make use of LDA(Latent Dirichlet Allocation) which provides us with the topics which were present most in the agenda being taken into consideration.To have an idea through we made use of visualization , the image shows the result of the analysis performed:

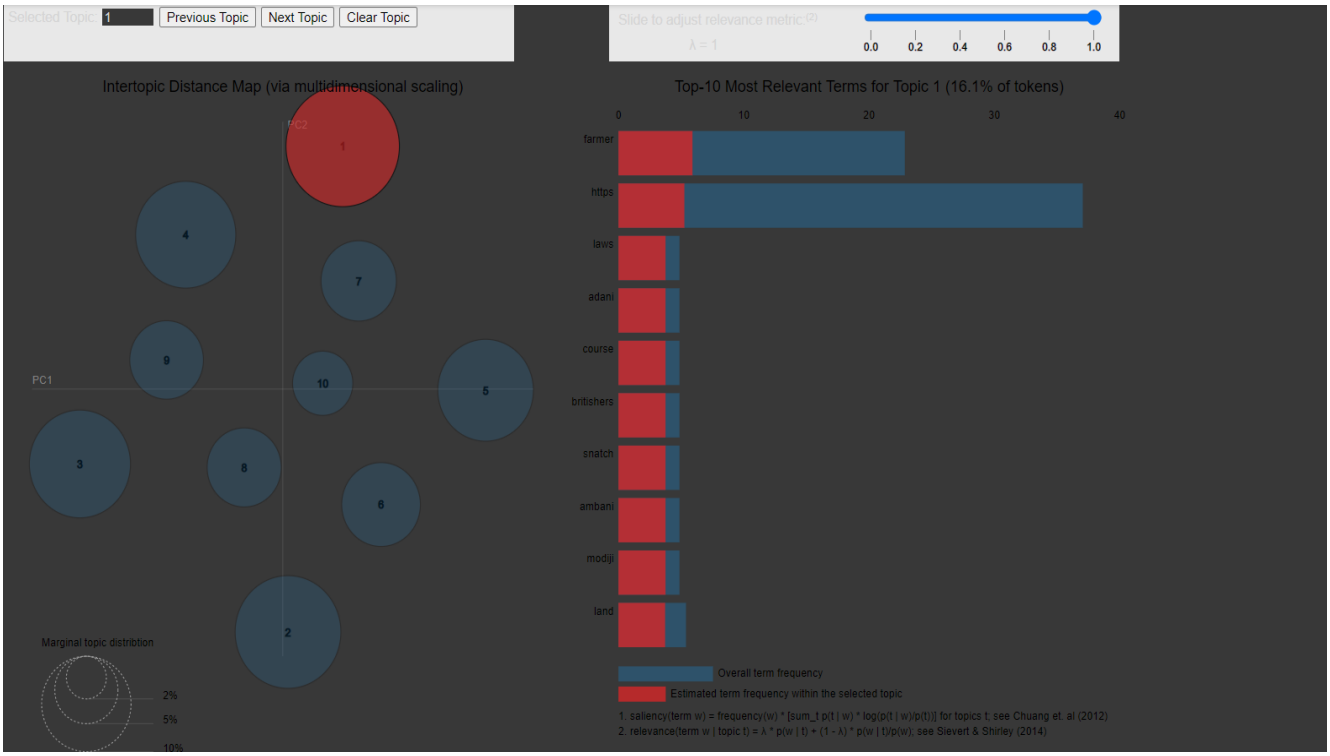

Fig. 10 Visualization of Topic modelling

- In the last stage we make use of location object provided in the json format of the data set and then try to analyse the tweets of a specific location .This helps in making informed decisions regarding the policies keeping in mind the opinions of the local people. The figure below provides us with the result for metropolitan cities like Delhi, Mumbai, Kolkata and Chennai.

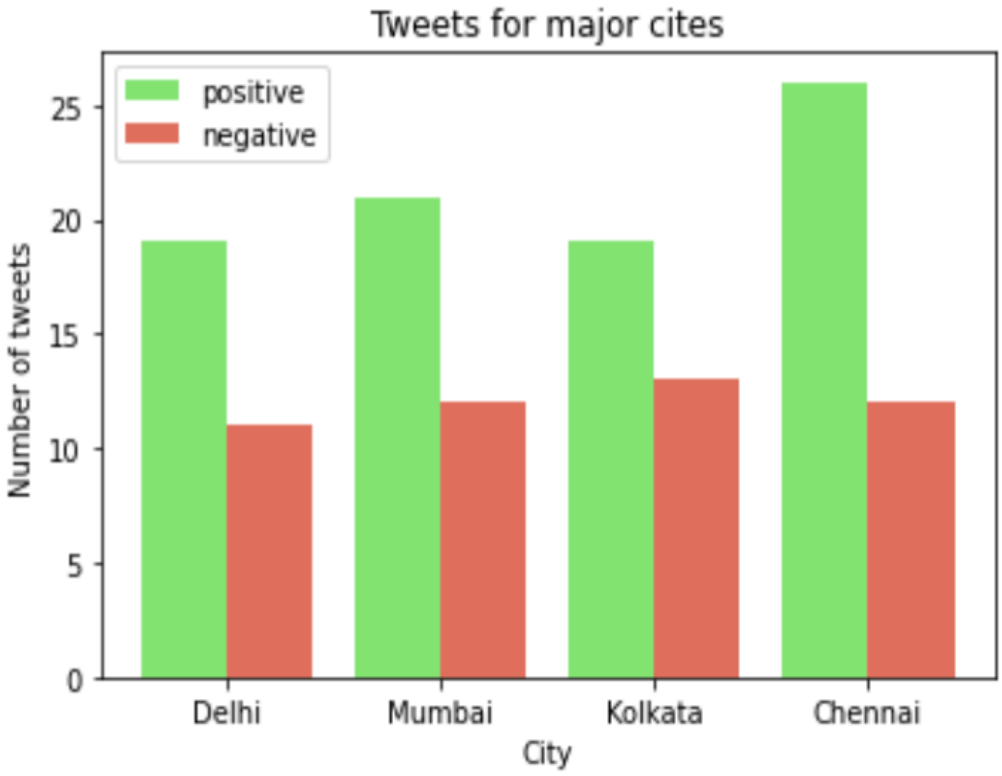

Fig. 11 Location based analysis of tweets

Chapter 05: CONCLUSION

5.1 Conclusion

Innovation is progressing with a quick speed and different governments are likewise proactive as far as embracing more up to date advancements for giving admittance to improved administrations for the service of its residents. A number of efforts are being made to close down the gap in perception which is prevalent between the policy makers and the common public, as it is them who are affected by these policies made by the government. Social media and cloud platforms have turned out to be two very powerful technologies for the government throughout the world to provide the best possible service to the general public. But in many cases, the studies conducted till now have only managed to utilize the merits of these tools independently and lacks any sincere effort for combining such emerging technologies. This project was started with the hope to combine some selected capabilities of social media analytics and cloud computing for storage towards efficient analysis of the public policies being implemented. We have implemented a cloud based approach, which takes in consideration the public's opinions via data gathered from twitter about some new policy which is being implemented. During analysis of these government reforms numerous analytical techniques are being used on the stored dataset that in this case are the tweets. Depending upon the reports provided by the techniques used on these social media sites, an appropriate result was provided by making use of different visualization techniques. In this project the testing was conducted with data collected on the Three Farm Bills imposed by the Indian government.

After performing various analysis methods, we came to a conclusion that The Three Farm bills gave a very negative impact on the public. From the sentiment analysis we found out that 78% of people are negative regarding this law. The Geolocation analysis gave us the major idea of the regions where mainly the negativity was coming from. So we safely concluded through analytic visualization that the Farm Bills law passed by the government was heavily disliked by the public. These implementation results suggest that the proposed methodology can be used to efficiently analyse the various public policies.

5.2 Discussion on the Results Achieved

Technically, this project provides us with a detailed cloud oriented system inside which user generated tweets are stored. The project also conveys about the way in which this large data set can be used in effective surveillance and controlling of government policies with the help of various online platform analytics which include descriptive analysis, content analysis and geospatial analysis.

The project also tries to analyze and report the opinion of the public towards Farm Bills law implemented by the Government of India. The process of collecting this data is done on Twitter with the help of hashtag "#FarmBill and #Farmers Protest". Numerous analytics involcing these platforms were performed in a cloud based environment to look deeply into the Farm Bills from the perspective of the public. The study of tweet statistics provided much help in comprehending the aftermath and about the magnitude of the issue happening because of the fresh reforms implementation. This detection of trends is carried out using (#)hashtag analysis. Moreover, it gave us better understanding about the association between various hashtags. Polarity and emotions linked with tweets were given by sentiment analysis, which provided much help in understanding the people's opinion regarding the Farm Bills. The sentiment analysis also acts as a valid indicator for the calculation of threshold value, depending on the particular warning signals that can be generated for the policy makers. Topic modeling carried out the process of identification of the theme. Finally, geo-location analysis helped in detecting targeted audiences that had negative opinions regarding the proposed three laws. This warning signals based on threshold and location based analysis are the tools for mapping the unhappy target public. These observations were the defined and distinct contributions made to this study.

5.3 Application of the Project

The application of this project can be categorised into two agendas (a) The government will formulate some new policy to be implemented (b) General Masses will be affected by the policy being implement. They are discussed briefly below.

1. **Government:**

One of the main stakeholders in this project system is the government.This is because it will be the one to introduce the policies to be followed and thereafter monitor it by making use of the proposed social media and cloud system. The most important step to be taken is to be taken by the government to spread appropriate awareness among the public before final implementation of any policy so that people start discussing the proposed plans on twitter, facebook and other virtual platforms. As data collection is done using virtual platforms, and if people are not active on them, then there won't be any data available for monitoring for the government.

To do this, a Facebook page can be created by the government, or the advantages of the new policy can be highlighted via tweets on Twitter. Immediate discussion and trolling or appraisal on these platforms will aid the policy makers to enhance monitoring of the policy which is to be implemented. This is done by considering numerous recommendations highlighted on social media by the public.If there is a case that the government receives too much negativity from the policy towards that particular policy, then the government has to consider postponing the implementation of that policy or maybe consider some expert opinion for improving the policy. As it was in the case of Three Farm Bills, if the government receives a lot of negative opinions towards a specific policy after the policy is implemented , then the government need to take the steps necessary for overcoming the problems that the general public is facing.

2. **Public/ People:**

Common people are also a major stakeholder during this project. We all know that every policy by the government is implemented for the public. As shown in our project results, the public should engage themselves in social media discussions in light of the new policy. Discussions on social media platforms like twitter will help the government in better understanding of merits as well as demerits of the law passed in an enhanced manner. Moreover the public can address their concerns and even give some recommendations which will definitely help the government in better framing

of the new policy and even making some improvements in the already existing policies.

5.4 Limitations of the Project

Although, the system proposed has provided us with good results but still it requires certain changes to fulfil the limitations. First of all, data (Tweets) for the experiment was taken only from one platform i.e. Twitter, this leaves room to take into consideration other platforms like Facebook for more data which can lead to more accurate results. Then second, only actual tweeted posts were taken into consideration for data analysis, and re-tweets were not considered. Re-tweets make up one-third portion of the entire traffic on Twitter, therefore we ignored a considerable amount of Twitter traffic during the analysis. Lastly, in our system detection of bots was also not performed. Bots on these platforms can affect the analysis to a large extent by leading to the results being biased. Therefore, it is necessary to consider them.

5.5 Future Work

The next step in this project will be to cover other social media platforms. Uptill now we have collected our data using Twitter APIs only. Further we would like to extend and vasten our database by collecting data from Facebook. It is another great social media platform where people present their views regarding the hot topics which in our case will be the ongoing government policies.

Also we can model a website for our project. Through which the user can directly feed his needs into the website and it will be fed to our model. And after processing all the data and implementation of the algorithms it will return the public views regarding the policy. The results can be displayed on the web page itself.

REFERENCES

1) Abascal-Mena, R., Lema, R., &Sèdes, F. (2015). Detecting sociosemantic communities by applying social network analysis in tweets. Social Network Analysis and Mining, 5(1), 38.

2) AlAlwan, A., Rana, N. P., Dwivedi, Y. K., &Algharabat, R. (2017). Social Media in Marketing: A review and analysis of the existing literature. Telematics and Informatics, 34(7), 1177–1190.

3) Attu, R., &Terras, M. (2017). What people study when they study Tumblr: Classifying Tumblr-related academic research. Journal of Documentation, 73(3), 528–554.

4) Grover, P., Kar, A. K., Dwivedi, Y. K., & Janssen, M. (2018). Polarization and acculturation in US election 2016 outcomes–can twitter analyt- ics predict changes in voting preferences. Technological Forecasting and Social Change. https://doi.org/10.1016/j.techfore. 2018.09.009.

5) A. S. (2018). Sharing political content in online social media: A planned and unplanned behaviour approach. Information Systems Frontiers, 20(3), 485–501.

6) IBM (2019) [Online], Available: https://www.ibm.com/cloud/learn/ benefits-of-cloud-computing

7) Dwivedi, Y. K., Rana, N. P., Janssen, M., Lal, B., Williams, M. D., & Clement, R. M. (2017a). An empirical validation of a unified model of electronic government adoption (UMEGA). Government Information Quarterly, 34(2), 211–230.

8) Mishra, N., & Singh, A. (2016). Use of twitter data for waste minimisation in beef supply chain. Annals of Operations Research, 270(1–2), 1–23. https://doi.org/10.1007/s10479-016-2303-4.

9) Rana, N. P., Dwivedi, Y. K., & Williams, M. D. (2013). Analysing chal- lenges, barriers and CSF of egov adoption. Transforming Government: People, Process and Policy, 7(2), 177–198.

10) Walther, M., &Kaisser, M. (2013). Geo-spatial event detection in the twitter stream. In European conference on information retrieval (pp. 356-367). Berlin: Springer.

11) Yuan, H., Xu, W., Li, Q., & Lau, R. (2017). Topic sentiment mining for sales performance prediction in e-commerce. Annals of Operations Research, 270(1–2), 1–24. https://doi.org/10.1007/s10479-017-2421-7.