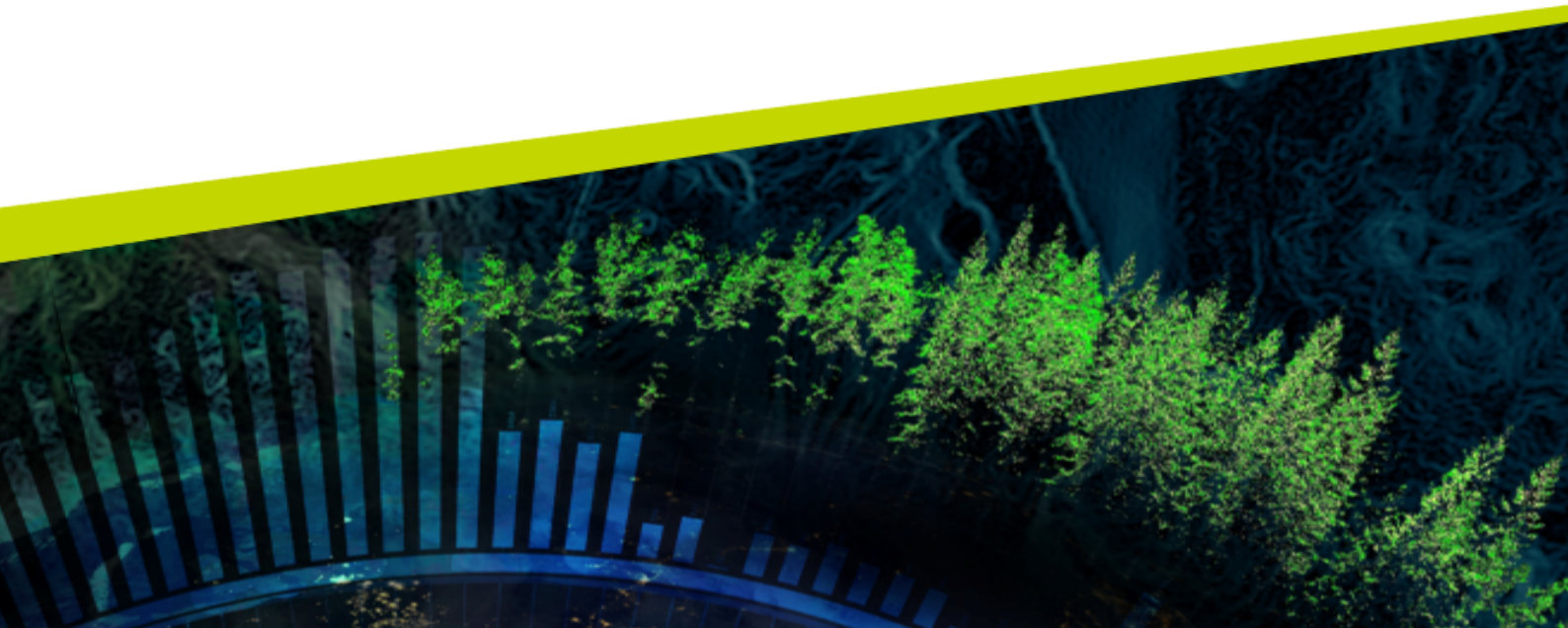
Geo-Spatial Analysis of Twitter Data

Data Mining (CS685A) PROJECT REPORT

Group 06
Nikhil Kumar Singh (19211265, ksnikhil@iitk.ac.in)
Vishal Singh (19211273, vshlsng@iitk.ac.in)
Suvasree Biswas (19111416, suvasree@iitk.ac.in)



Indian Institute of Technology, Kanpur
Dept. of CSE
December 6, 2020

1 Abstract

With the emerging advent of various social network tools and platforms, analyzing and drawing inferences from the societal response to important and emerging social issues and events through social media data is a problem of immense importance. It is inevitable, now more than ever, in this pandemic struck world that the power of the online platforms becomes a monopolistic power. However, there are numerous challenges towards realizing this goal as effectively and efficiently as one would like to perceive, due to the unstructured and noisy nature of social media data. This project tries to overcome this, with the focus being on inferring from the sentiments when spread across the demographic scenario. Our output conveys that by intertwining and developing a combination of machine learning and data management techniques, the effectiveness is escalated at an overpowering scale. The results of our project have the potential to be used in finding a solution to other interesting social issues such as building neighborhood happiness and health indicators.

2 Clear Problem Statement

The aim of our project is Geo-Spatial analysis of (a) India for Twitter data of **Farmer-Protest** and **Hathras rape** case, (b) World for Twitter data of **France and Vienna Terror Attacks** and **US Elections** case. The tweets are tweeted from different parts/locations of India and World. We will show number of tweets tweeted from different major regions of India/World, which will show which parts of India/World have major rage/support for that particular case (i.e Farmer-Protests).

We are further going to do sentiment analysis of these tweets which will also indicate that how much of them have positive sentiment and Negative sentiment for different regions in India and World.

We are also looking to form clusters from the sentiment analysis data, based on the density distribution.

3 Introduction and Motivation

Since the inception of Twitter, people have been glued to these platforms to express their opinion about current affairs, politics, business, sports, finance and entertainment. There are studies with statistics depicting that twitter is being used predominantly by people of age under 30. The unquenchable voice of young generation is pivotal in paving road to the future of any country. Ever increasing social media usage coupled with enhanced computing technologies have enabled us to analyze peoples opinion from their tweets at a large scale. An event like election that attracts the interest of crowd and has significant impact on society. Such effect is worth analyzing.

Posts by social media users provide details about the people perceptions of their immediate surrounding, and updated observations about what is trending in the real vicinity. In this approach, we look forward to exploit the relevance of Twitter as a data source for geospatial analysis as well as sentiment analysis.

These tweets also indicate the presence of active twitter users in different regions across India and World. They also indicate as to how much impact a given issue (i.e. Farmer protests) has on the the people in different regions of India and World. Moreover, the sentiment analysis of these tweets indicate the nature of these tweets.

For instance, we would like to find the impact of Farm Bills brought by the Indian Government in Sept 2020 on different different regions of India. We feel that most impacted regions should be those where Farming is a major component of the local economy. So, our expectation is that the northern plains of India which has a significant farmer population to have the maximum impact by

the farm bills [5]. Similarly, we will also look into the impact of Hathras Rape Case [1] on different regions on India. In this case, we would expect the major tweets to be from major cities especially those with high literacy levels. The previous two analysis are mainly restricted to India. We next try to look into Global issues. We try to find the impact of the unfortunate incident of beheading of a School teacher in France and the terror attacks in Vienna during October 2020. This was an event of significant importance since the opinion were highly divided across the world [4, 3]. The opinion of European nations was different from those in the Middle East especially Turkey. We will also be looking into the impact of election results in US during November 2020. This event was significant because the world opinion was highly divided in favour of Democrat candidate Joe Biden and Republican Candidate Donald Trump [2]. We already know that US plays a significant role in the global geopolitics and hence the person occupying the President post would have a tremendous impact on the world. For instance, the differences between the two candidates are more apparent with respect to issues like China (COVID, HongKong), Terrorism, Iran Nuclear Deal etc.

4 Datasets used

We scrapped data from Twitter for four different cases as mentioned below along with their hashtags:

- **Farmer-Protest :**$\#FarmersProtest$, $\#FarmBills$, $\#FarmerFirst$, $\#farmer$, $\#FarmersWithModi$, $\#FarmersBill2020$
- **Hathras rape :**$\#HathrasCase$, $\#HathrasTruthExposed$, $\#HathrasTapes$
- **France Beheading and Vienna Terror Attack :** $\#FranceBeheading$, $\#ViennaTerrorAttack$
- **US Elections 2020 :** $\#USElection2020$

Tweets are scrapped from Twitter in different days by group members as shown in Table 1, 2, 3, 4

Date of Scrapping	Number of Tweets
22-Sept	0.1K
23-Sept	8.7K
24-Sept	11K
25-Sept	37K
26-Sept	16.8K
27-Sept	10.7K
28-Sept	15.5K
29-Sept	7.8K
30-Sept	3K
Total	111K

Table 1: Farmer's Protest

Date of Scrapping	Number of Tweets
29-Sept	33K
30-Sept	64K
01-Oct	35K
02-Oct	58K
03-Oct	78.5K
04-Oct	30K
05-Oct	18.7K
06-Oct	15K
Total	402K

Table 2: Hathras Rape Case

Date of Scrapping	Number of Tweets
29-Oct	96.4K
30-Oct	18.7K
03-Nov	89.7K
04-Nov	3.9K
Total	208.7K

Table 3: France and Vienna Terror Attacks

5 Methodology

5.1 Creating Dataset

Creating twitter developer account and getting the access API keys. Then, we scrap tweets using *tweepy* python library for each day. Else, we often faced issue of key exhaustion due to twitters' restrictions on daily limit on the number of tweets that can be scrapped from an account. Then we merge all tweets date-wise and create a single CVS file(dataset). In CVS file, a row indicates a tweet and column describes its attributes as shown in Table 1.

Issues faced during creation of datasets: At first we tried to write tweepy object data for each tweet in file as a row, and when we tried to create tweepy object from file,there was no method to create tweepy object. Then secondly we used $\_json()$ function of tweepy object to get the tweet's all metadata along with data. that was successful, but when we tried to read the json data of every tweets from file, it was giving parsing error. The json format which tweepy object returning was not able to read by python json library. Throught this whole process we lost some date's tweets as tweeter does not allow to scrap tweets before seven days from current date. Finally we desing a tuple representation for each tweets, where tuple values/attributes are shown in Table 1. The tuple representation contains tweet data and also important tweet's metadata(required metadata for our analysis and some extra which may help in future extension of this project).

Merging tweets of different date to create a final tweet data file for each cases: For our four case study (Farmer-Protest, Hathras rape, France and Vienna Terror Attack, US Election 2020),tweets are scrapped in different dates as shown in table 1,2,3 and 4. For each case study, tweets of different dates are merged to their corresponding cases to get final dataset. The procedure to create final dataset is given Figure 1.

Date of Scrapping	Number of Tweets
4-Nov	109.5K
5-Nov	129K
6-Nov	49.3K
Total	287.8K

Table 4: US Elections 2020

Column	Attribute
1	When the tweet is created.
2	Tweet ID.
3	Text of the tweet.
4	Geo location
5	Geo co-ordinate of tweet location
6	place name
7	How many time that tweet retweeted
8	Users who re-tweeted
9	language of tweet
10	location of user.
11	User ID
12	User's user name
13	User's profile display name
14	User's profile description
15	When user's profile is created in Tweeter.
16	UTC offset
17	User's time-zone
18	User's Geo location is on or not.
19	User is verified user or not.
20	User's Language.

Table 5: Tweets Attribute's details

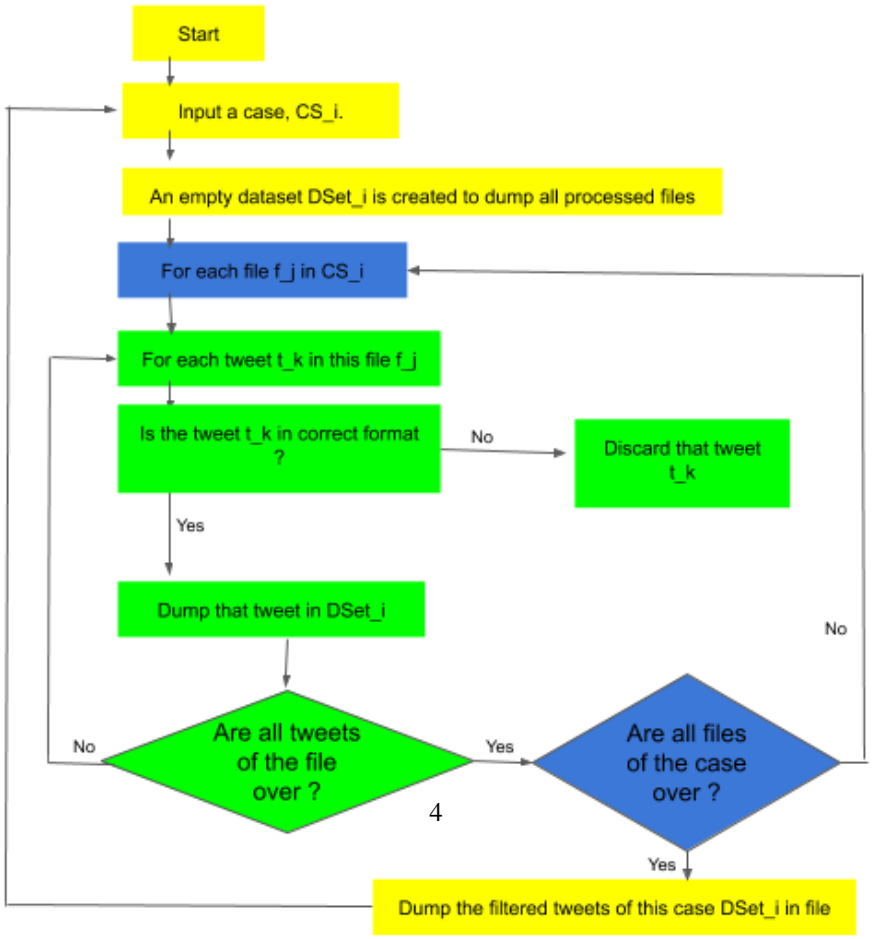


Figure 1: Flowchart showing the procedure to create dataset

5.2 Data Preprocessing and Sentiment Analysis

5.2.1 We list down some of the issues and their solution that we faced while creating the datasets for our project.

- **Issue with newline characters:** Some tweets has newline characters in their tweets which will make problem when we read the data from the csv file so it is removed from tweets.
- **Issue with delimiters:** Since Tweets can contains any special characters(a user can use that character in his/her tweet), which will make problem to the csv file. If delimiter used in csv file is used in the any tweet then during reading from csv file that delimeter will make extra column i.e row will not be read as the way it should. To resolve this problem, we are writing every column as a tuple format i.e every attributes are the values of a tuple, so using python function EVAL(), string representation of row can be parsed to EVAL() function and that will return a correct tuple format entry in valid format.
- **Names,locations and tweets are present in multiple language:** The tweets are scrapped using hashtags, so these tweets are tweeted from not only from India but also from other country in our data. So their locations are in multiple language, also there are some cases where tweets are tweeted from India for which location are not in English e.g few locations of Tamilnadu are present in Tamil language, other countries' location are present in Arabic,Chinese language etc. To resolve this issue we used geopy library. This library takes a location name in multiple languages and also if a tweet's location is partially provided it can use that to generate the proper adress in English.
- **Missing attributes:** Our project is mainly depends on location of user's tweet and text of tweet. As we already mentioned attributes which we scraped from tweeter in section "Creating Dataset", the values of "Geo location" and "Geo co-ordinate of tweets location" are not provided by tweeter i.e None. So we are completely dependent on column 10 i.e user's location which is missing for many tweets. in our case if the user's locations are not present then that tweet is considered as useless tweet for Geo-special analysis as we can't interpolate or guess user's location by ourselves. So during preprocessing stage missing location tweets are discarded. There are also other attributes are missing but they are not important for our analysis,so ignored.
- **Tweet preprocessing:** Tweet preprocessing is also the major part for our project since it will be used for finding sentiments. To do sentiment analysis, we need to remove noises from tweets. Noise in tweets refers to hyperlinks(e.g https://...),user mentions (e.g @vishal112, @nikhils, @spurveam1), Hastags, Punctuations,Emoticons, Numbers, re-tweet, invalid words etc because they are irrelevant to the sentiment. To remove noise re python library i.e regular-expression library is used.
- **Invalid locations in tweets:** Some of the locations present in data have invalid address, those tweets are removed using regular expression and geopy library.

5.2.2 Creating final processed dataset:

Our final dataset will contain geo-location and tweet's sentiment along with with positive and negetive sentiment value. A tweet in this datset will be a set of 26 attributes value as shown in Table 6. Steps to create this dataset is shown in Flowchart 2. The same procedure is applied to all four datasets.

The very first step is to remove all those tweets for which location is not present in the tweet's metadata i.e attribute/column 10. Then for the remaining tweets, removing noisy/invalid locations(in our case this noisy locations are handled using regular expression and geopy library).Now we are left

with all tweets for which tweet's location metadata are present. Some noisy locations are still present in data.

Then next step is to find the geolocation(latitude and longitude) and full address of a tweet's location.The python library geopy is used for this step. This library allows maximum 100 entries to process at one session and 1 entry processing time nearly .5 second. To reduce time, we created a set of location which will contain all locations unique.

Then to find sentiment of tweets, noises are need to removed before it is used to find its sentiment value. As we already mentioned about noises in tweets under section 5.2.1. Hyperlinks(e.g https://...) are replaced with word URL, user mentions (e.g @vishal112, @nikhils, @spurveam1) are replaced with $USER\_MENTION$, Hashtags are replaced by removing $\#$ from Hashtags, removing retweet, replacing more than 2 dots with space, striping space, Γ, from tweet, replacing multiple spaces with single space, Convert more than 2 letter repetitions to 2 letter(funnnnny with funny),removing punctuation, replacing Emoticons(Smile, Laugh, Love, Wink emojis with EMO_POS and Sad, Cry emojis with EMO_NEG), Numbers, re-tweet. Then each old tweet is replaced new preprocessed tweet to new dataset.

Then we are building Sentiment classifier to classify above tweets. We have unlabelled data i.e we do not have labels of tweets. we tried unsupervised methods to find sentiments of tweets but they perform really poorly so we used a labelled dataset. The TextBlob library is used to create the sentiment classifier model. NLTK's labelled movie-review corpus to classifier. We used Naive-Bayes based classifier to train model. Our model is soft classifier instead hard classifier because sentiment values will be used further in our analysis.

After training the sentiment classifier model we classify each tweets as either Positive(labelled as pos) or Negative(labelled as neg). Neutral tweets are those whose sentiment value is near 0.5. As we already mentioned above sentiment values are used in our analysis so Neutral tweets are easily handled in analysis. So sentiment label(pos or neg), positive sentiment value and negative value are added for each tweet in data-set.

The dataset is finally created which will be used in **section 5.3** for generating results to get Maps and clusters.

Column	Attribute
1	When the tweet is created.
2	Tweet ID.
3	Preprocessed text of the tweet.
4	Geo location
5	Geo co-ordinate of tweet location
6	place name
7	How many time that tweet retweeted
8	Users who re-tweeted
9	language of tweet
10	location of user.
11	User ID
12	User's user name
13	User's profile display name
14	User's profile description
15	When user's profile is created in Tweeter.
16	utc offset
17	User's time-zone
18	User's Geo location is on or not.
19	User is verified user or not.
20	User's Language.
21	Latitude.
22	Longitude.
23	Processed complete address.
24	Tweet is positive or negative(pos or neg).
25	Tweet's positive sentiment probability value.
26	Tweet's negative sentiment probability value.

Table 6: Tweets Attribute's details

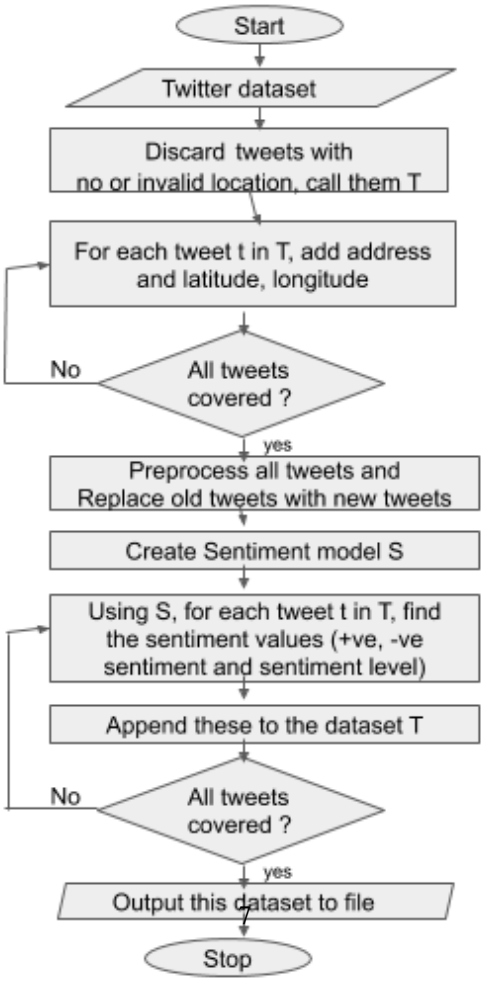


Figure 2: Flowchart showing the procedure to create final dataset

5.3 Generating Results

In this section, we are generating World maps (for France and Vienna Terror attacks and USElection) and India maps (for Farmer-protests and Hathras rape case) and displaying every locations' Normalized sentiment value on the map with different colors. sentiment value (0-0.25) is considered as highly negative and dark red color is used to indicate in map. Same for (0.25-0.50),(0.50-0.75) and (0.75-1.00) less negative, less positive, highly positive sentiment and ,green, dark green colors are used. To display the number of tweets from a location, we used a circle marker. The more number of tweets from a particular location, the more will be the circle size.

As the cluster of regions will give significant help in the analysis, we used DBSCAN (Density based) cluster algorithm. As partitioned based or Hierarchical clustering will not give much significant result compared to Density-based clustering method as partitioned based clustering methods will give convex shape cluster but in our case cluster will be of any shape and it also sensitive to noise which is more in our case. As also we do not know how many clusters will form, so hierarchical clustering is not suitable. Density based will work better as our cluster points are very close to each other within a cluster. DBSCAN uses 2 parameters, eps and MinPts. eps is the neighborhood around a data point i.e. if the distance between two points is lower or equal to 'eps' then they are considered as neighbors. MinPts is the minimum number of neighbors (data points) within eps radius.

The General structure in given below.

- Datasets generated from **Section 5.2.2**
- Normalized sentiment calculated for every location present in the dataset.
- We plot all locations on the map (for Farmer-protests and Hathras rape case : India Map, for France and Vienna Terror attack and USElection : World Map) using folium library.
- While plotting, we use several important features of folium.
- We display the count for each location along with the circle marker.
- Moreover, we also use colored markers for geo-spatial analysis as well as sentiment analysis.
- In case of geo-spatial analysis, we use color to show the density of tweets from the location.
- In case of sentiment analysis, we use color to show the magnitude of sentiment being positive/negative.
- To find the cluster of regions for analysis, DBSCAN algorithm is used.

5.4 Libraries Used

- **geopy** : This library is used to generate the proper address and Geo-location of the user's tweet's location.
- **folium** : This Library is used to generate the map on which we will show our analysis.
- **tweepy** : This Library is used to scrap tweets from Twitter.
- **nltk and textblob** : These Library is used for tokenization, lemmatization, preprocessing and Sentiment analysis.
- **DBSCAN** : This is used to run the clustering algorithm dbscan.
- **re** : This is used for working with regular expressions.
- **csv** : This is used for working with csv files.

6 Results and Analysis

In this section, we analyse the results of our four case studies. Two case studies are India specific i.e. Farmer Protests and Hathras Rape Case and the other two case studies are global i.e. Terror Attacks in France & Vienna and US Elections 2020. We will display our analysis on the map of India, world and different regions.

6.1 Farmer Protests

In case of tweets regarding Farmer protests, we observed the following:

- Most of the tweets and hence, active twitter users belong to the northern and western parts of India (Figure 3).
- From Figure 4, we observe that both positive and negative sentiments are present throughout the regions. However, tweets with positive sentiments are much more compared to those with negative sentiments.
- Major cities contributing maximum number of tweets are from Delhi, Mumbai, Jaipur, Chandigarh, Ludhiana, Ambala, Chennai, Bangalore, Hyderabad.
- Majority of the tweets are from Northern states like Delhi, Punjab and Haryana (Figure 5).
- We also observe a large number of tweets from Hyderabad region of Telangana (Figure 6) and from Tamil Nadu and Kerala (Figure 7).

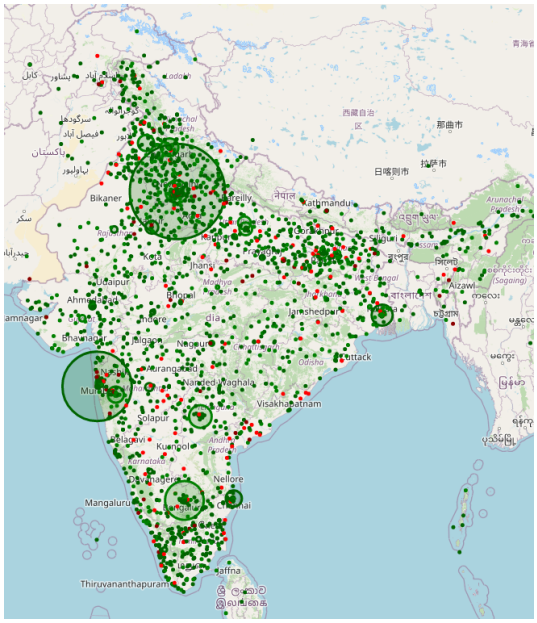


Figure 3: Tweets for Farmer Protests in India during 22-Sept to 30 Sept, 2020

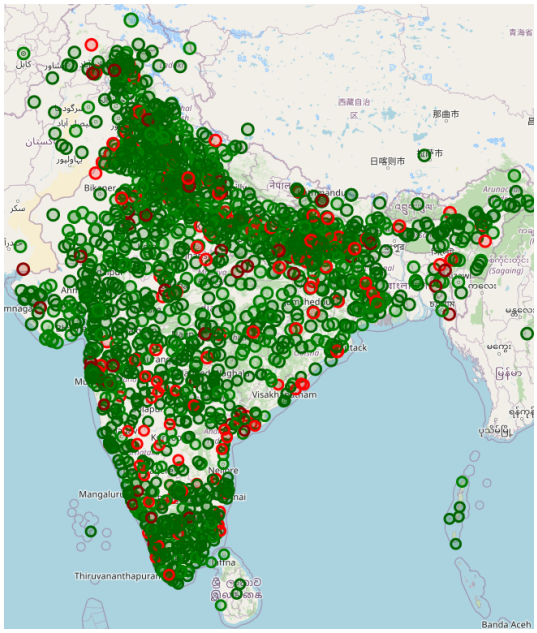


Figure 4: Sentiments of Tweets for Farmer Protests in India during 22-Sept to 30 Sept, 2020

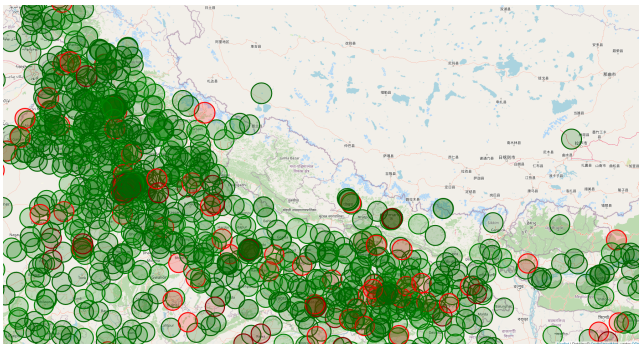


Figure 5: Sentiments of Tweets for Farmer Protests in North India during 22-Sept to 30 Sept, 2020

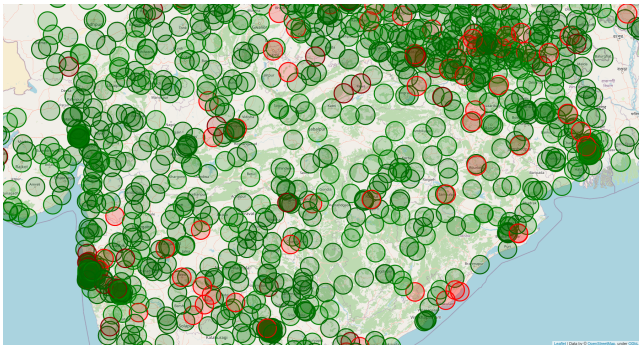


Figure 6: Sentiments of Tweets for Farmer Protests in Central India during 22-Sept to 30 Sept, 2020

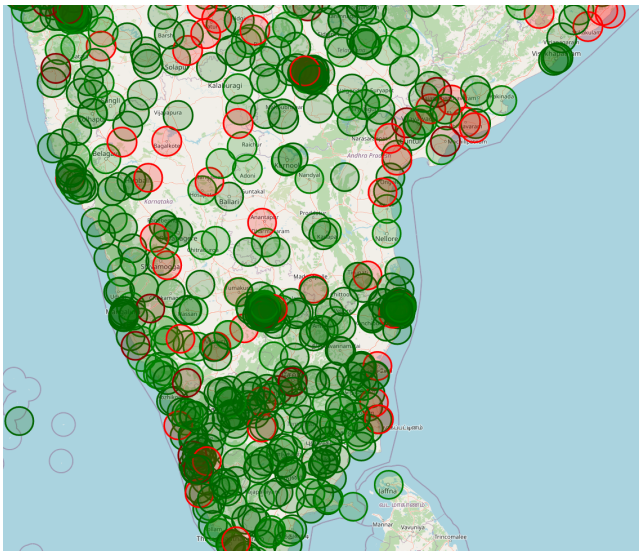


Figure 7: Sentiments of Tweets for Farmer Protests in South India during 22-Sept to 30 Sept, 2020

6.2 Hathras Rape Case

In the case of tweets regarding Hathras Rape Case, we observed the following:

- We see a lot of tweets with negative sentiment.
- In **Eastern India and Southern India**, the number of tweets with negative sentiments is generally more than the number of tweets with positive sentiments.
- Even throughout India, generally the number of tweets with negative sentiments are almost equal to those with positive sentiments.
- Major tweets are from the cities of Delhi, Mumbai, Bangalore, Kolkata, Hyderabad, Lucknow, Patna.

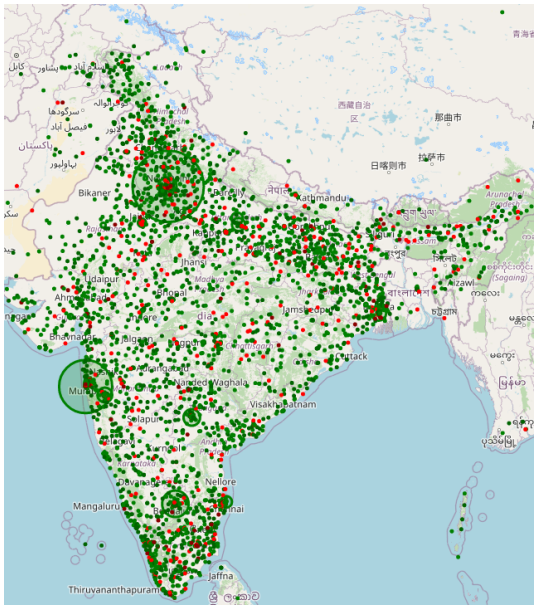


Figure 8: Tweets for Hathras Rape Case in India during 29-Sept to 6-Oct, 2020

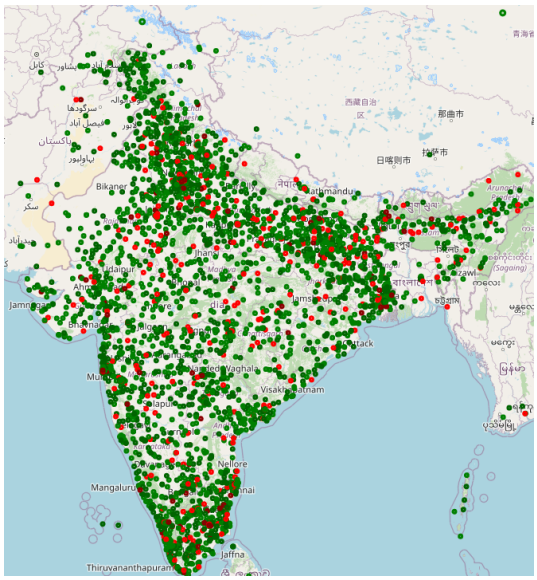


Figure 9: Sentiments of Tweets for Hathras Rape Case in India during 29-Sept to 6-Oct, 2020

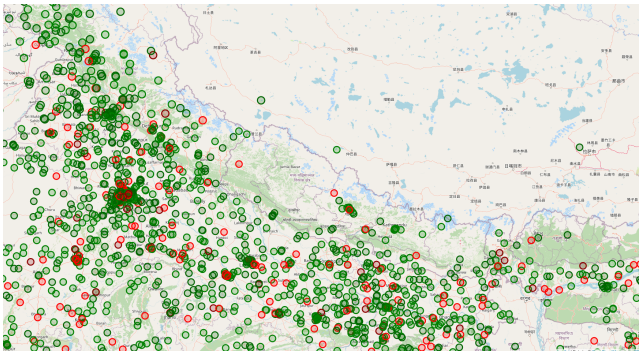


Figure 10: Sentiments of Tweets for Hathras Rape Case in North India during 29-Sept to 6-Oct, 2020

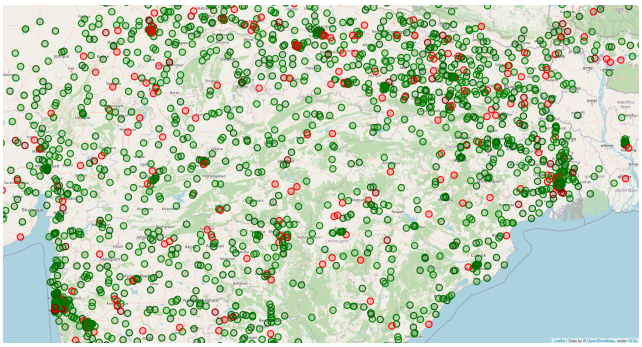


Figure 11: Sentiments of Tweets for Hathras Rape Case in Central India during 29-Sept to 6-Oct, 2020

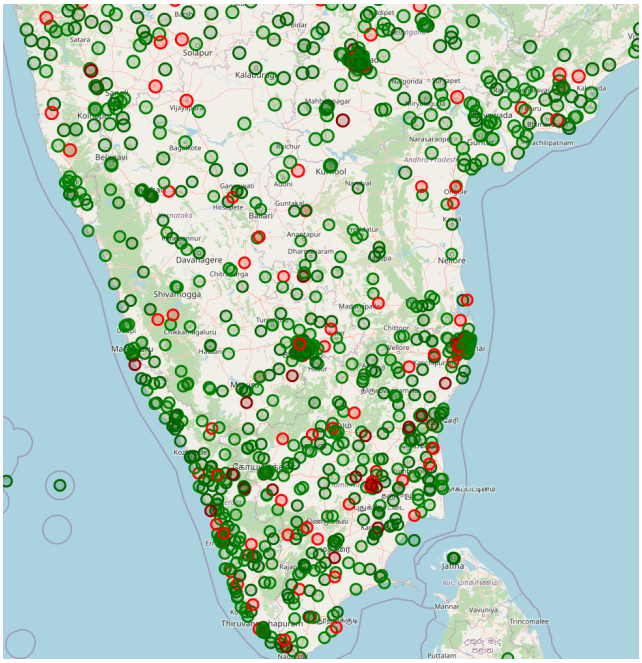


Figure 12: Sentiments of Tweets for Hathras Rape Case in South India during 29-Sept to 6-Oct, 2020

6.3 France and Vienna Terror Attacks

In this section, we present our analysis of tweets related to terror attacks in France and Vienna.

- We see a significantly lesser number of tweets from India compared to the previous two case studies. This can be because the issue was more global than the previous ones (Figure 13).
- We observe that the majority of the tweets are from Europe, India, Turkey and US. There was tremendous support for the French President from India.
- Although the density of tweets are comparatively less for Turkey, however when we take into account its population (8 crores approx), these are very significant. We observe that western turkey contributed more tweets than the other parts. Moreover, tweets with positive sentiments are more in number than those with negative sentiments.
- Another interesting country worth observing is Israel which contributed to a lot of tweets when compared to its population (88 lakhs approx). Here, we see tweets from all parts of the country.
- Although the population of Middle East (42 crore approx) is much more than that of Turkey (8 crore approx), but the number of tweets from Turkey is much more than the Middle East.
- In Europe, we see the majority of tweets from northern parts of Europe especially Britain, Germany, Brussels, France, Austria, Czech Republic, Netherlands and Switzerland.

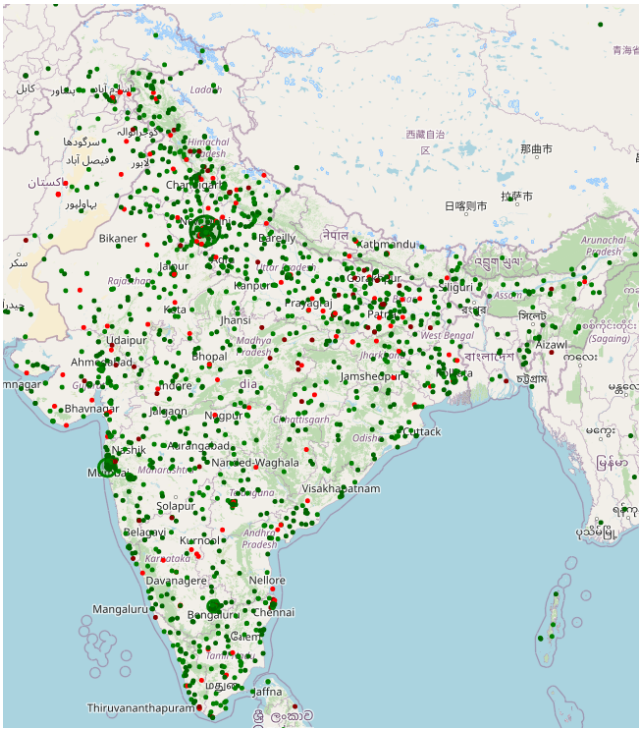


Figure 13: Tweets for France and Vienna Terror Attacks in India during 29-Oct to 4-Nov, 2020

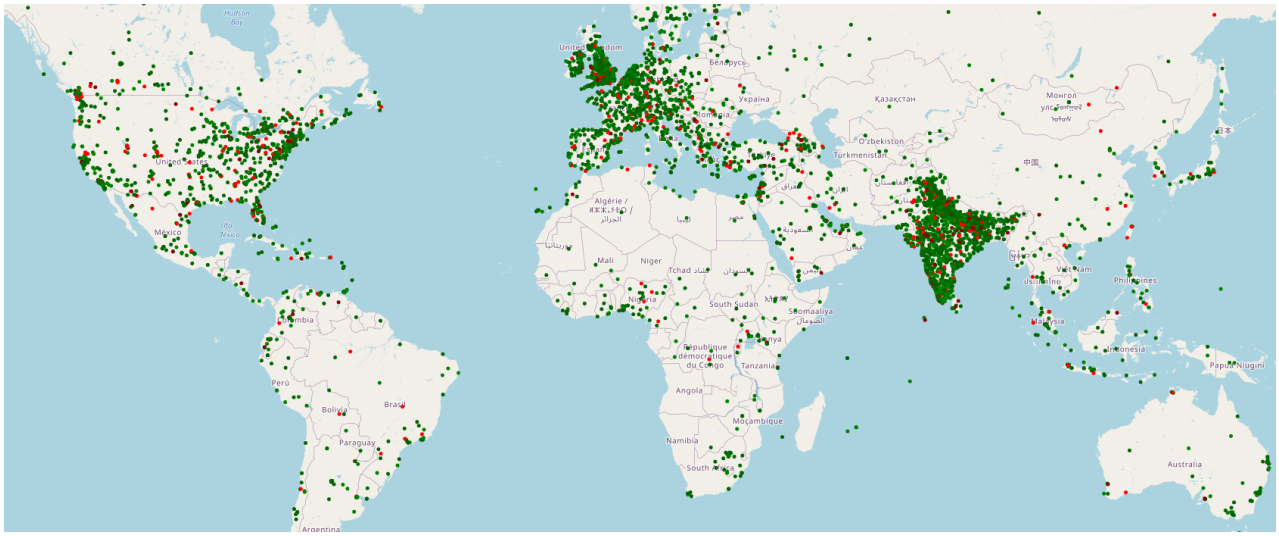


Figure 14: Tweets for France and Vienna Terror Attacks all over the world during 29-Oct to 4-Nov, 2020

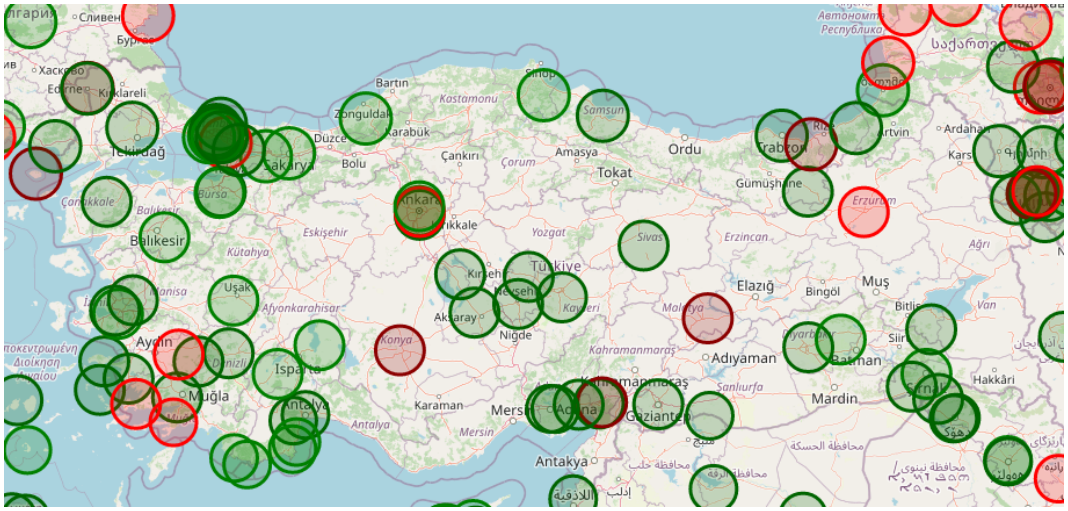


Figure 15: Tweets for France and Vienna Terror Attacks from Turkey during 29-Oct to 4-Nov, 2020

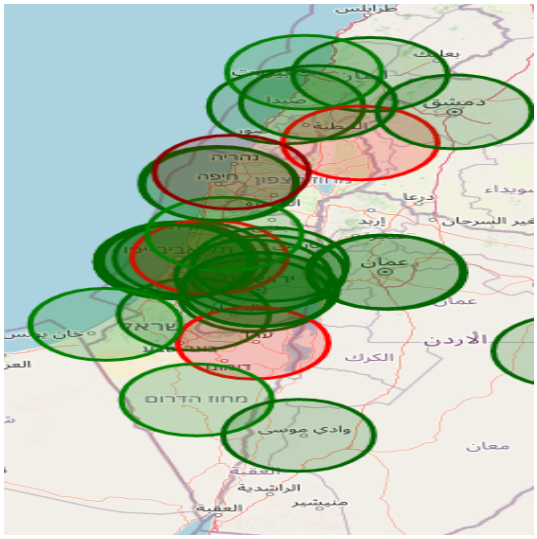


Figure 16: Tweets for France and Vienna Terror Attacks from Israel during 29-Oct to 4-Nov, 2020

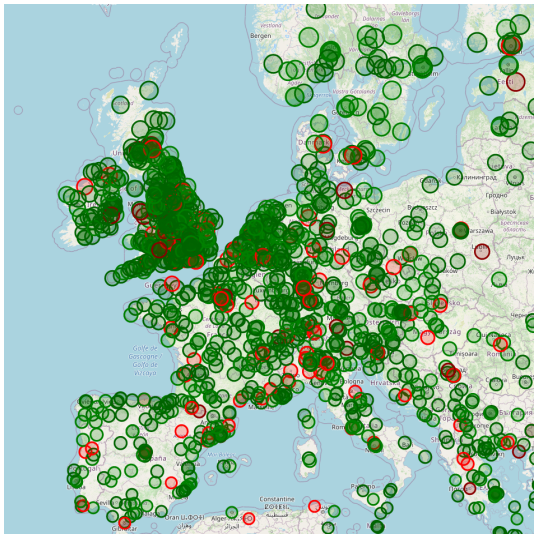


Figure 17: Tweets for France and Vienna Terror Attacks from Europe during 29-Oct to 4-Nov, 2020

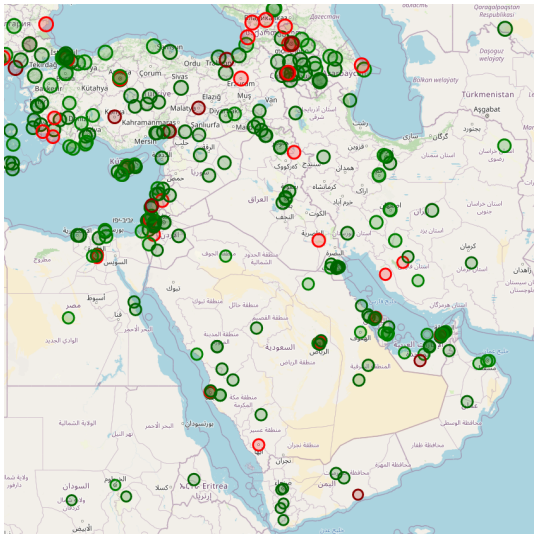


Figure 18: Tweets for France and Vienna Terror Attacks from Middle East during 29-Oct to 4-Nov, 2020

6.4 US Elections 2020

In this section, we present our analysis of tweets related to US Elections 2020.

- In case of tweets from India, we observe that tweets of negative sentiments are more (Figure 19). Also the majority of tweets are form the northern and southern parts of the country.
- In Figure 20, we observe that the tweets are much more diversified than those in case of France Terror Attacks.

- We observe that eastern part of US contributed much more than its western part. Moreover, the number of tweets with negative sentiments are more in number (Figure 21).
- In case of Asia (Figure 22), we observe that majority of tweets are from India, Japan, Philippines, Pakistan, Malaysia, Indonesia, South Korea and Hong Kong. More importantly, majority of tweets from Japan are of negative sentiments. Also, it is important to note that many tweets are observed from east China even though Twitter is officially blocked by China. However, it has been widely known that people across China use these sites though VPN and even companies like Huawei and CCTV use Twitter through a government-approved VPN.
- In the case of Australia (Figure 23), we observe that majority of tweets come from its eastern coast and are of positive sentiments. We also observe many tweets from New Zealand.
- In case of Africa (Figure 24), we observe the majority of tweets from three countries - Nigeria, Kenya and South Africa.
- In case of Europe (Figure 25), we observe the largest number of tweets compared to any other region of the world, even including US. The number of tweets with positive sentiments and negative sentiments are almost equal in number.
- In case of Middle East (Figure 26), we observe that majority of tweets are from Turkey, Israel, Georgia
- In case of South America (Figure 27), we observe that majority of tweets are from Brazil, Chile, Ecuador, Argentina and Columbia. Most of the tweets are of negative sentiments.
- We observe that the central part of USA is devoid of much sentiments. This observation is owing to the fact that these are predominantly agricultural states.

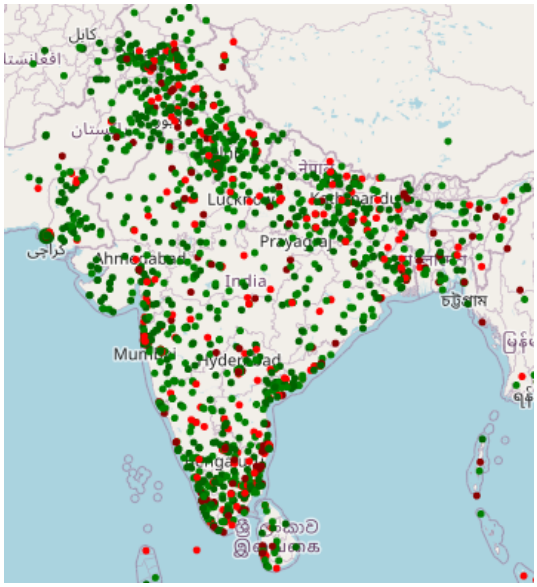


Figure 19: Tweets for US Election 2020 in India during 4-Nov to 6-Nov, 2020

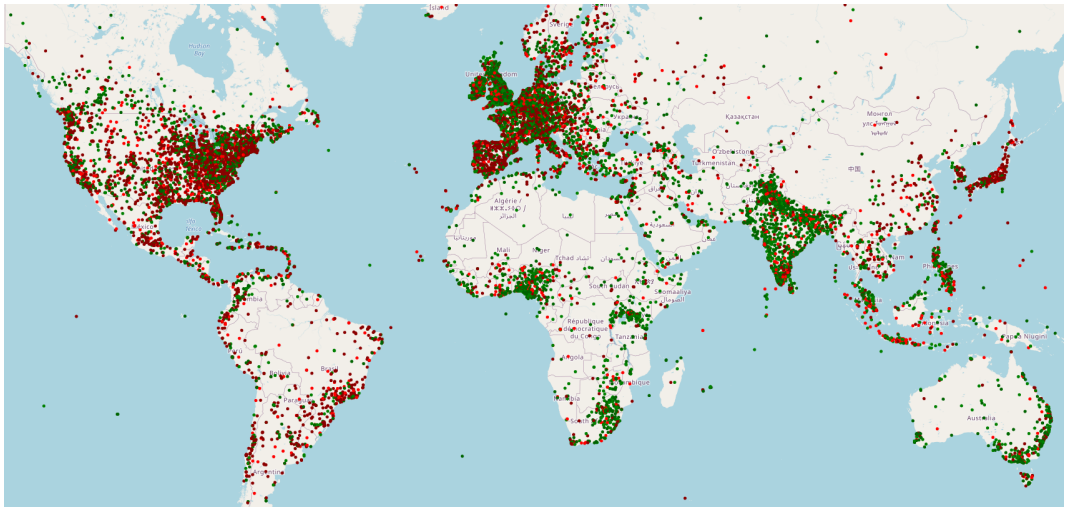


Figure 20: Tweets for US Election 2020 all over the world during 4-Nov to 6-Nov, 2020

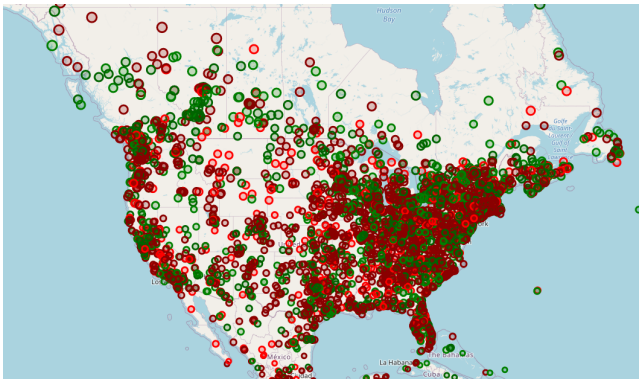


Figure 21: Tweets for US Election 2020 from US during 4-Nov to 6-Nov, 2020

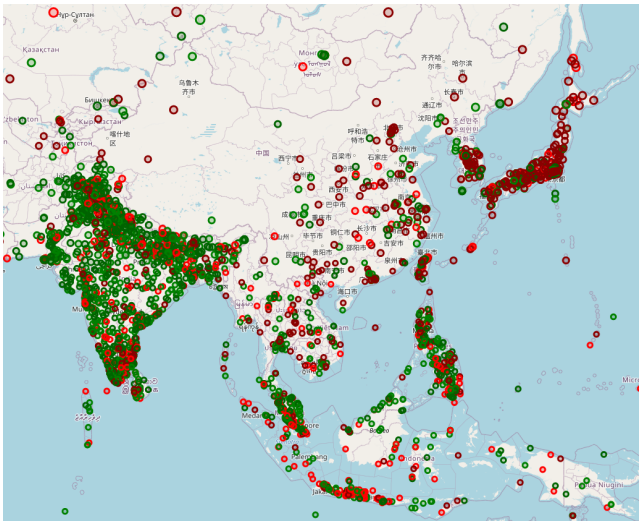


Figure 22: Tweets for US Election 2020 from Asia during 4-Nov to 6-Nov, 2020

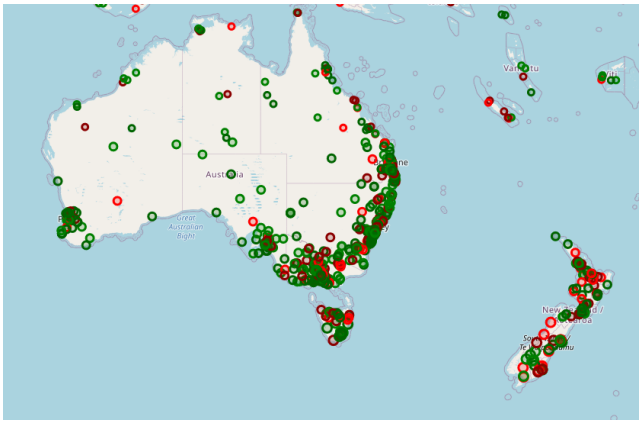


Figure 23: Tweets for US Election 2020 from Australia and NewZealand during 4-Nov to 6-Nov, 2020

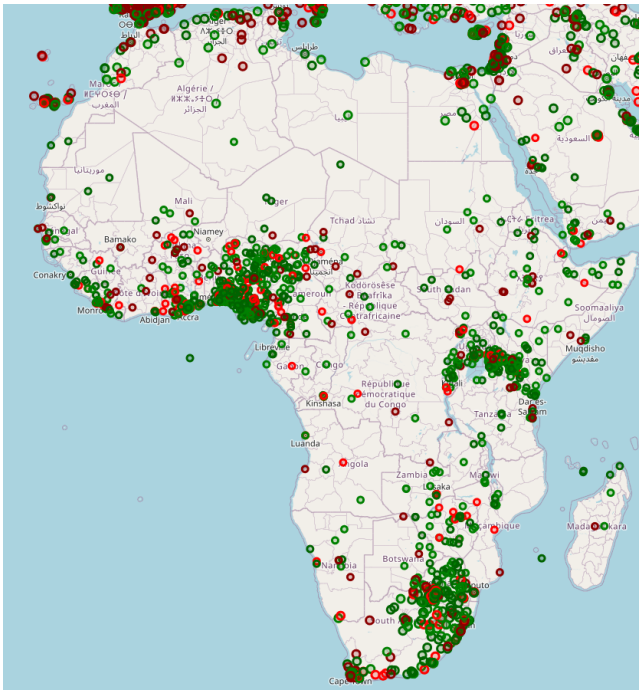


Figure 24: Tweets for US Election 2020 from Africa during 4-Nov to 6-Nov, 2020

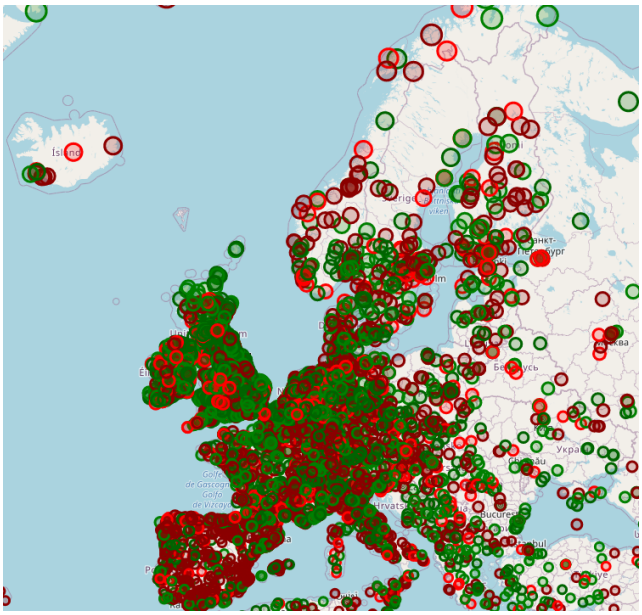


Figure 25: Tweets for US Election 2020 from Europe during 4-Nov to 6-Nov, 2020

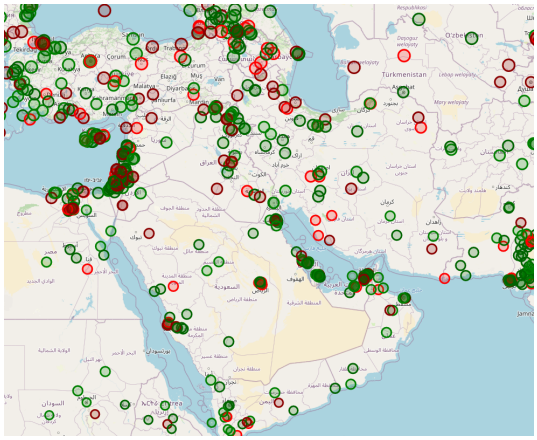


Figure 26: Tweets for US Election 2020 from Middle-East during 4-Nov to 6-Nov, 2020

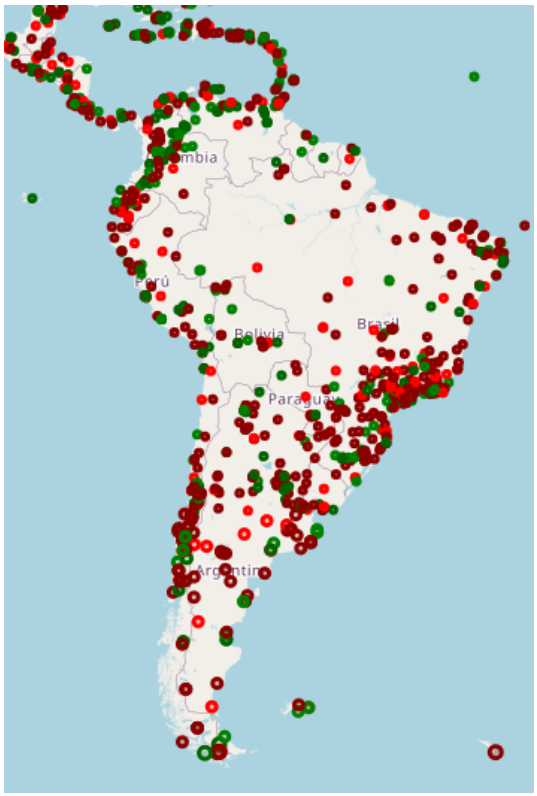


Figure 27: Tweets for US Election 2020 from Latin America during 4-Nov to 6-Nov, 2020

6.5 Clustering using DBSCAN algorithm

For the above analysis, we have used the DBSCAN clustering algorithm. The main reason for using DBSCAN is that it is extremely efficient in performing density-based clustering. DBSCAN also

works well with noisy data and Twitter data happens to have a lot of noise. Clusters are shown with different color excluding color black, color black indicates non-clusters.

The particulars of the parameters used for the DBSCAN algorithm are as follows:

1. For Farmers data: Epsilon=0.5, MinPoints=10.
2. For Hathras data: Epsilon=0.4, MinPoints=10.
3. For Vienna and France: Epsilon=0.8, MinPoints=10.
4. For US Elections: Epsilon=0.5, MinPoints=10.

The deciding factor for the selection of epsilon is to take the next point just after the peak of the curve. This is owing to the observation that for each of the epsilon value the best number of clusters is yielded. In Farmers, Hathras, France and US datasets, the number of clusters increase for a small portion and start decreasing over large portion. This implies that the dataset is well distributed across different nations.

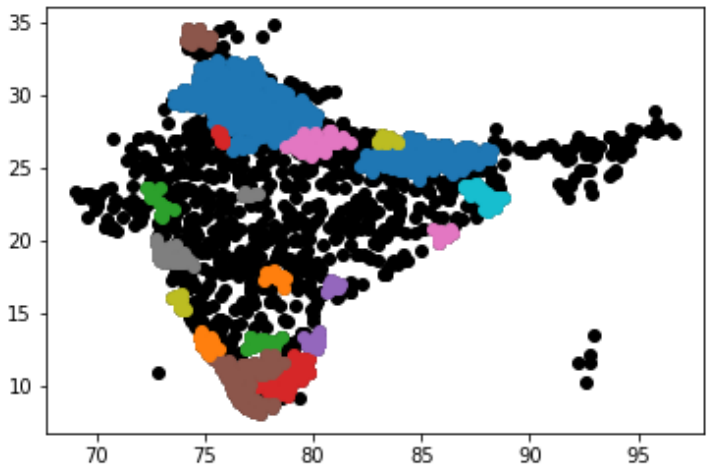


Figure 28: DBSCAN clustering algorithm on Farmer protests data

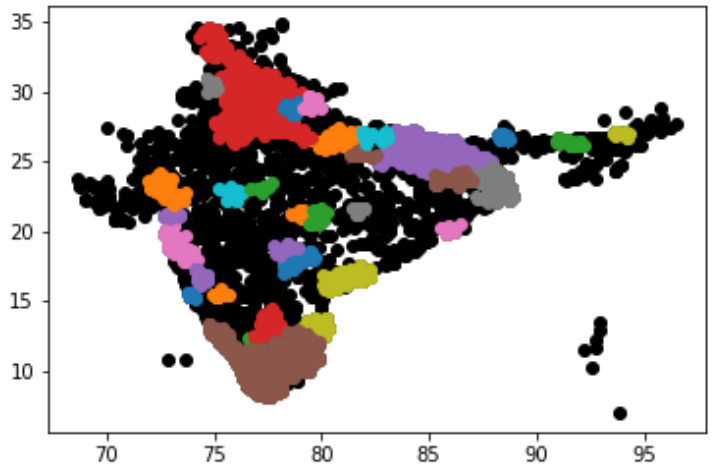


Figure 29: DBSCAN clustering algorithm on Hathras case data

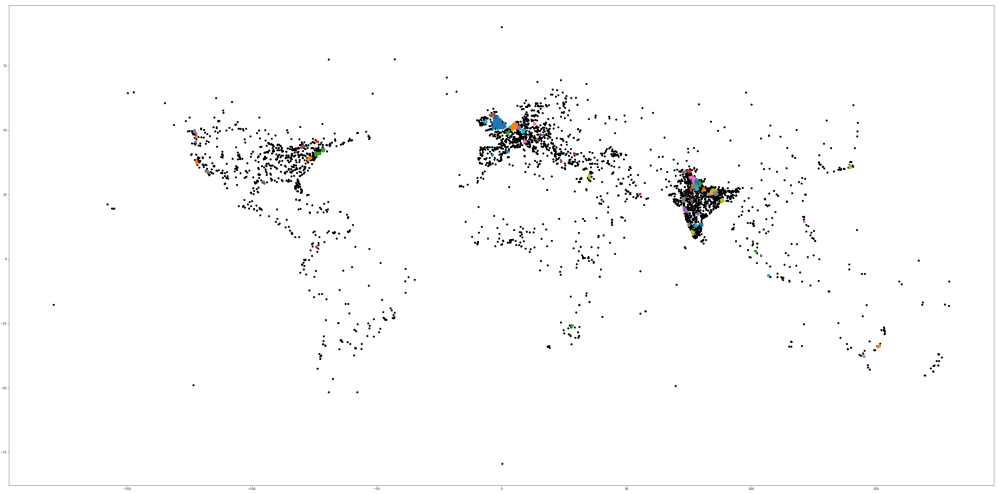


Figure 30: DBSCAN clustering algorithm on France and Vienna Terror Attack data

6.5.1 Tuning Parameters of DBSCAN

For each of the 4 dataset cases, using Elbow method, optimal number of clusters had to be obtained. This was done by tuning the two parameters: epsilon and MinPoints parameters. X axis represents the epsilon value. Y axis represents the number of clusters value. In the 3 cases of Farmers, Hathras

and USElections, we observe that after a small positive slope, it keeps decreasing abruptly. Only in case of Vienna does it have a large positive slope with no steep negative slope.

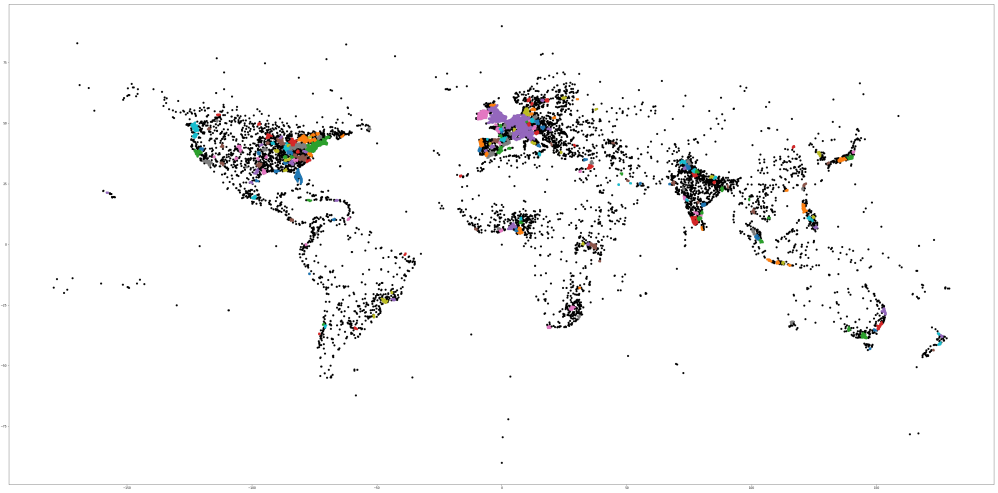


Figure 31: DBSCAN clustering algorithm on US Election data

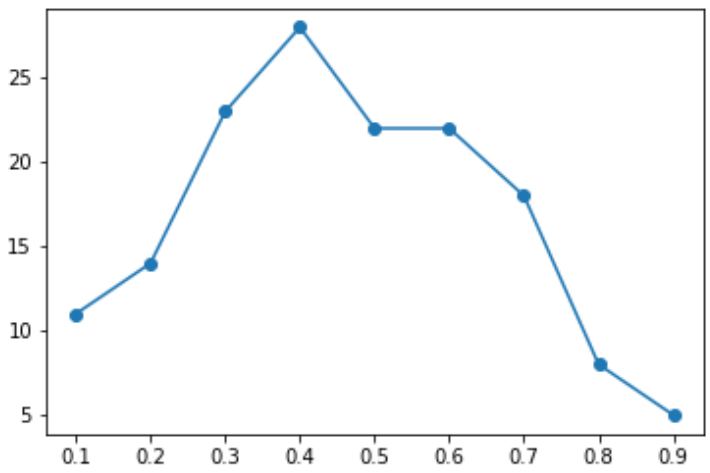


Figure 32: Number of Clusters v/s eps parameter for DBSCAN algorithm in case of Farmer's Protest

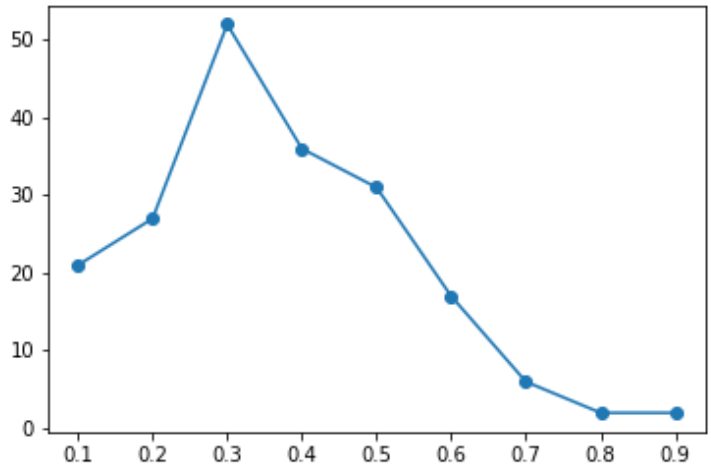


Figure 33: Number of Clusters v/s eps parameter for DBSCAN algorithm in case of Hathras rape.

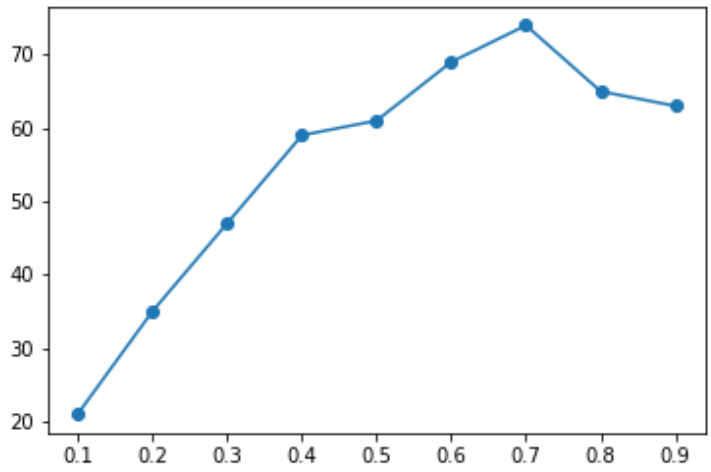


Figure 34: Number of Clusters v/s eps parameter for DBSCAN algorithm in case of France and Vienna Terror Attacks.

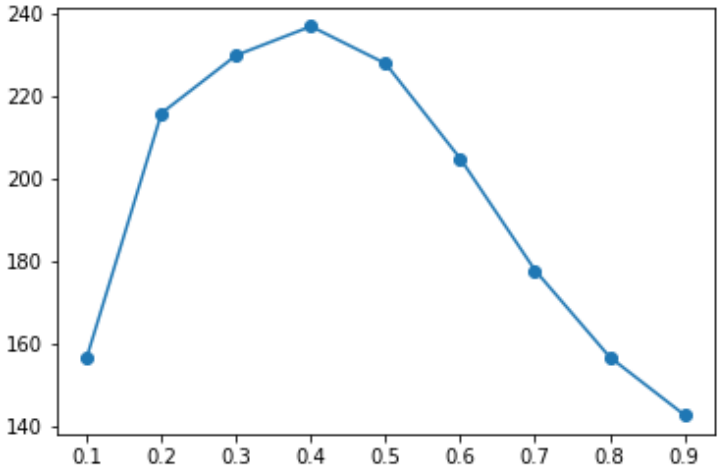


Figure 35: Number of Clusters v/s eps parameter for DBSCAN algorithm in case of US Elections 2020.

7 Conclusion and Future Scope

For this project, we mined approximately 10 lakh tweets. This number could have been even higher, if we had not lost some dataset due to the issues described in section 5.2. We could not recover it since Twitter only allows mining of upto 7 day old tweets only. We believe that Twitter by default filters some of the tweets with high negative sentiments. This we also observe from our dataset.

As an extension to this project, we can also build a **Sentiment predictor** which will predict the sentiment for a given location. We have the dataset ready with us and we can train a model do this.

References

[1] *https://en.wikipedia.org/wiki/2020*$_{H}athras_{g}ang_{r}ape_{a}nd_{m}urder$.

[2] *https://en.wikipedia.org/wiki/2020*$_{U}nited_{S}tates_{p}residential_{e}lection$.

[3] *https://en.wikipedia.org/wiki/2020*$_{V}ienna_{a}ttack$.

[4] *https://www.bbc.com/news/world-europe-54729957*.

[5] *https://www.thehindubusinessline.com/economy/agri-business/how-many-farmers-are-there-in-india-government-has-no-clue/article30614882.ece*.