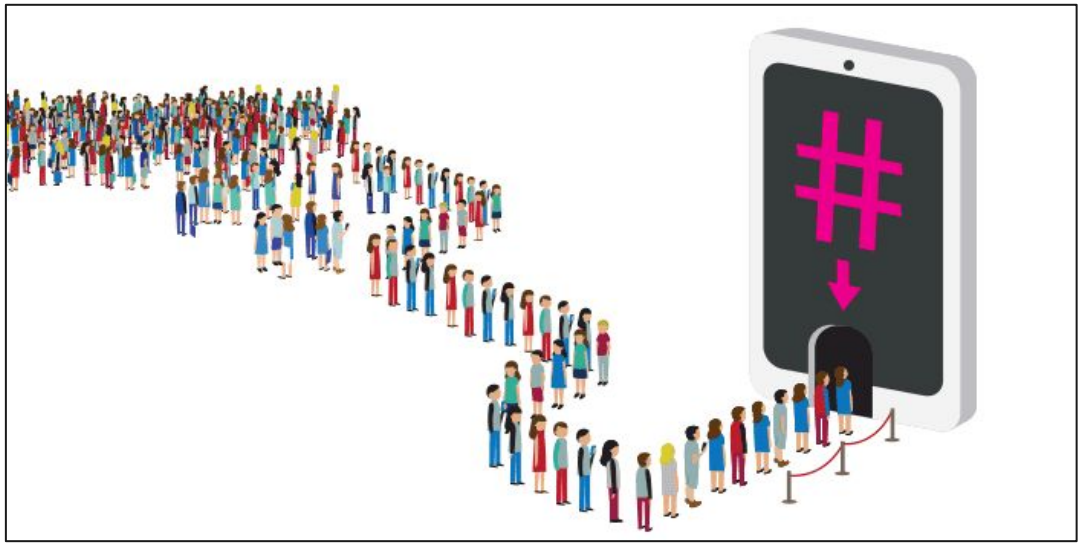
Geo Spatial Analysis of Twitter Data

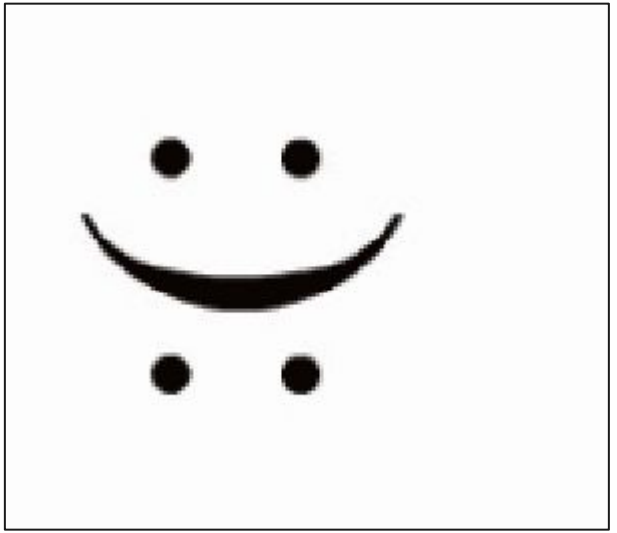
Group 06
Nikhil Kumar Singh(19211265, ksnikhil@iitk.ac.in)
Vishal Singh(19211273, vshlsng@iitk.ac.in)
Suvasree Biswas(19111416, suvasree@iitk.ac.in)

Abstract

WHY TWITTER ?

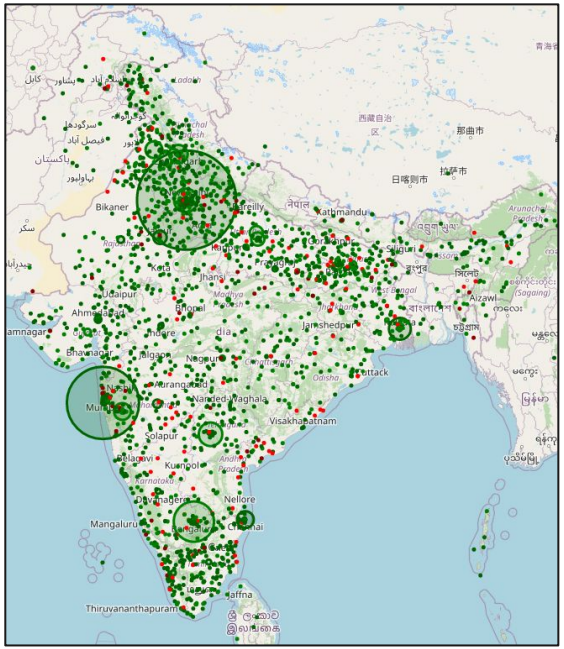


+ve / -ve

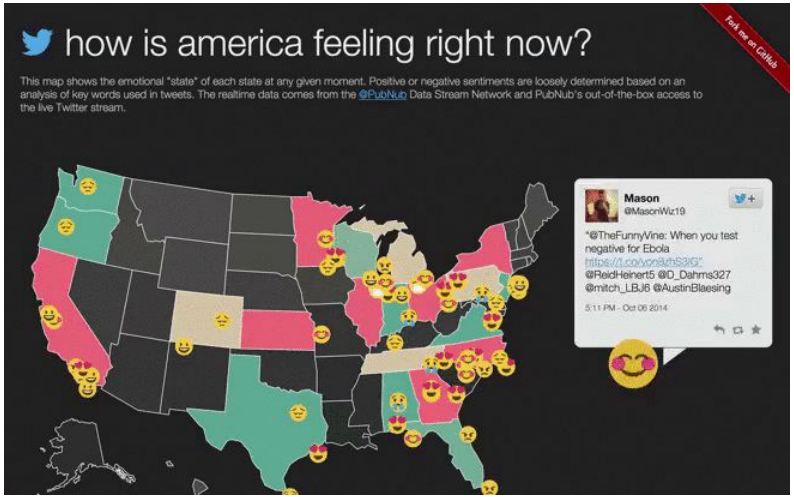


SENTIMENT

Problem statement



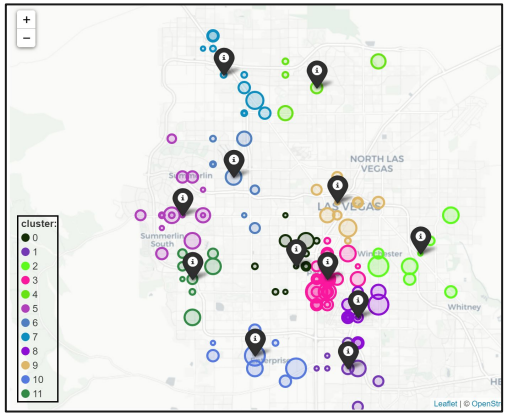
How is India feeling



Introduction

Power of Social Media

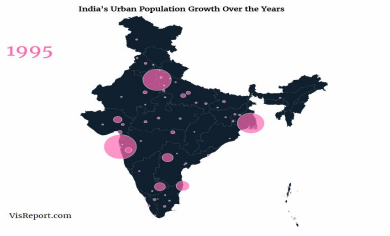
Goal1: - segregate +ve / -ve sentiment



Goal2: - according geospatial data clusters

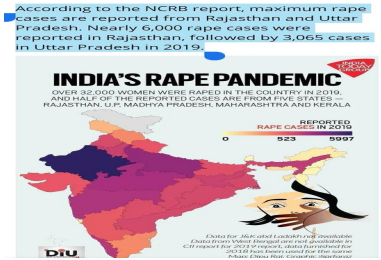
MOTIVATION

National Level



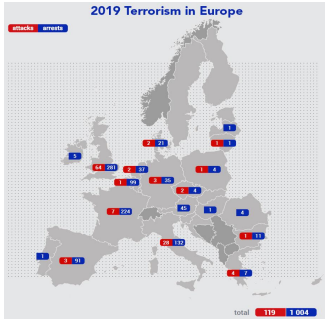
Urbanisation

Farmers Protest



if Hathras would have been the only case

Global Level



Terror attacks

US elections

Datasets Used

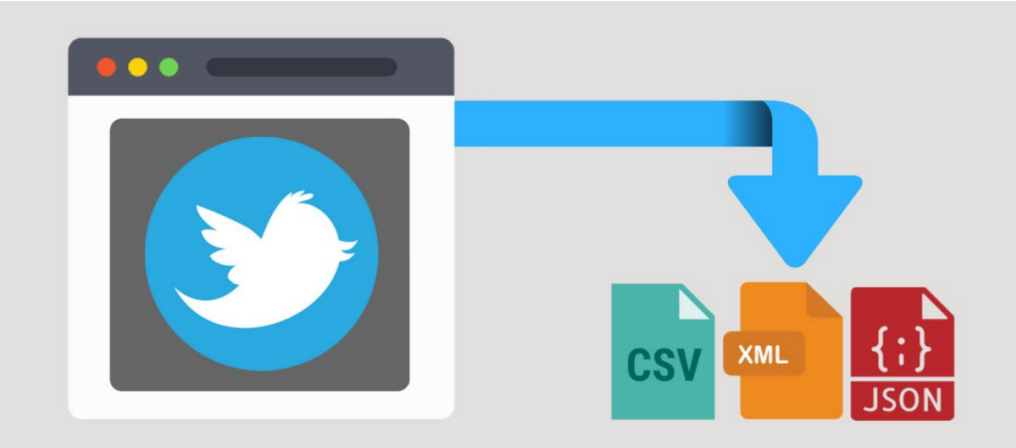
NATIONAL

Case 1	Case 2
HathrasRapeCase	FarmersProtest

GLOBAL

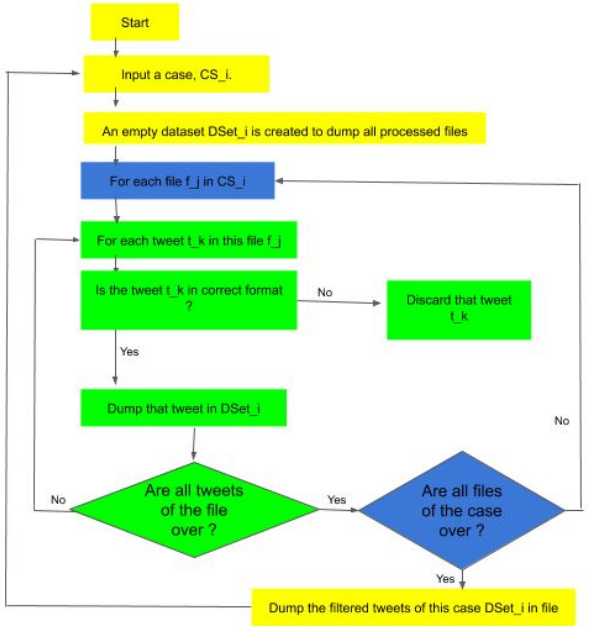
Case 1	Case 2
FranceAndVienna TerroristAttack	USElections2020

Scrapping



tweepy

Creating Datasets



Flowchart to filter and combine all scraped data

METHODOLOGY

- Data comes in all shapes and sizes.
- Process it to make it usable
- Discard non-usables

Flowchart Overview

Creation of dataset and sentiment analysis

Take each dataset T

Discard invalid tweets of the dataset T

For each tweet, append address, latitude, longitude to T

Create Sentiment model S

Preprocess Tweets in T

Use S to find Sentiments of T

Append Sentiments to T

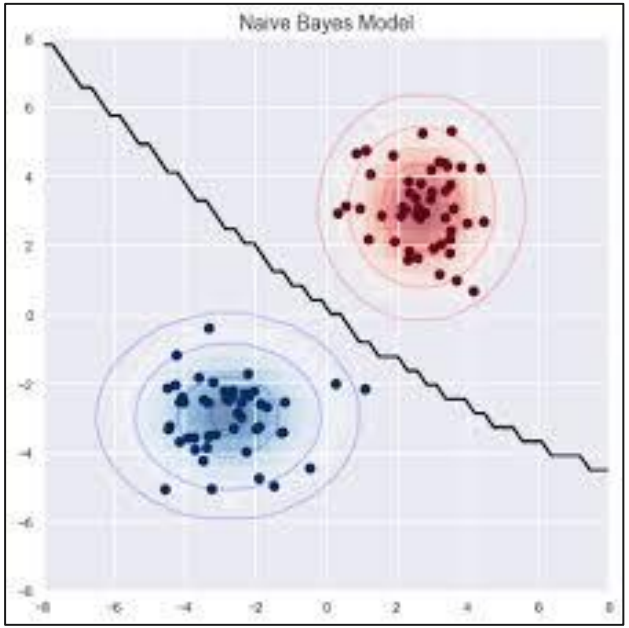
Analysis Phase

Generate Maps with data plots

Clustering

Analysis

Objective 1: segregating into +ve / -ve sentiment



Naive Bayes Classifier - to determine the probability of being +ve or -ve sentiment

Creating final dataset

Tweet attribute format

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.

Sentiment Analysis

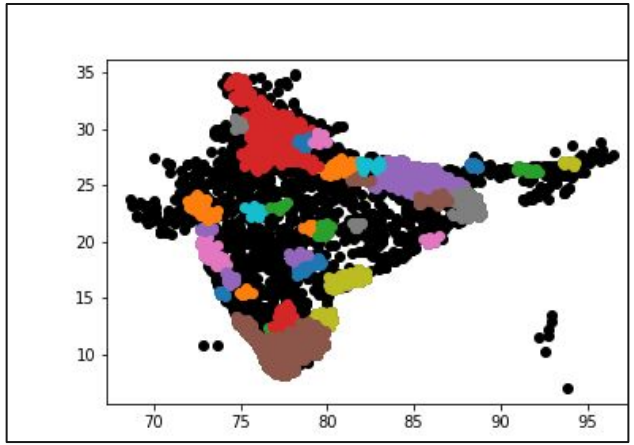
Appending Location

Final dataset created

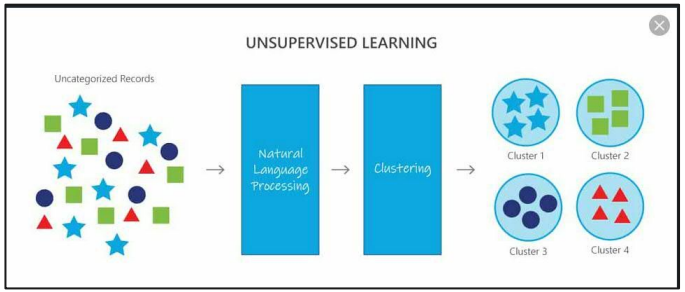
Column	Attribute
1	When the tweet is created.
2	Tweet ID.
3	Preprocessed text of the tweet.
4	Geo location
5	Geo co-ordinate of tweet location
6	place name
7	How many time that tweet retweeted
8	Users who re-tweeted
9	language of tweet
10	location of user.
11	User ID
12	User's user name
13	User's profile 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 negetive(pos or neg).
25	Tweet's positive sentiment probability value.
26	Tweet's negetive sentiment probability value.

Objective 2: forming clusters

What we look to have ?



Which Method to chose ? Why ?



Our Dataset : Unlabelled data

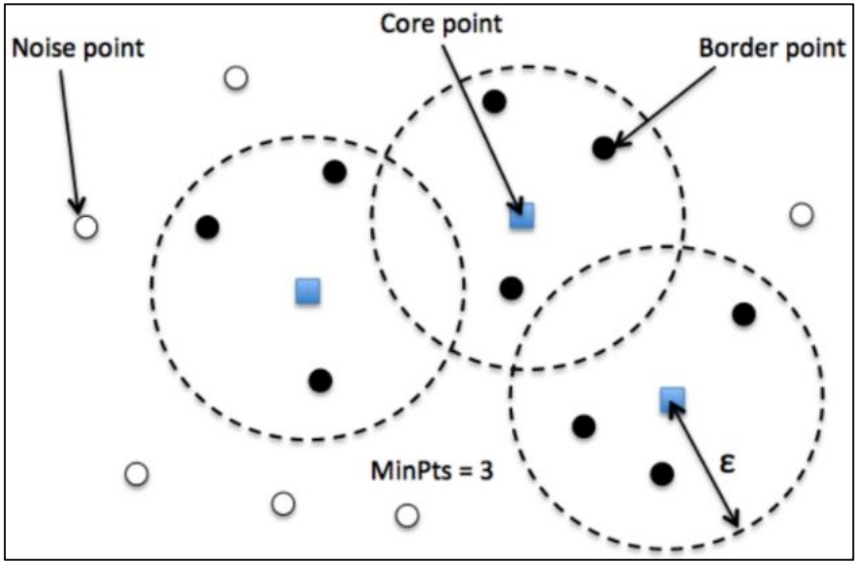
↓

Choice of Method: Unsupervised Learning

↓

Choice of Algorithm : Density Based Clustering

Technique Opted for

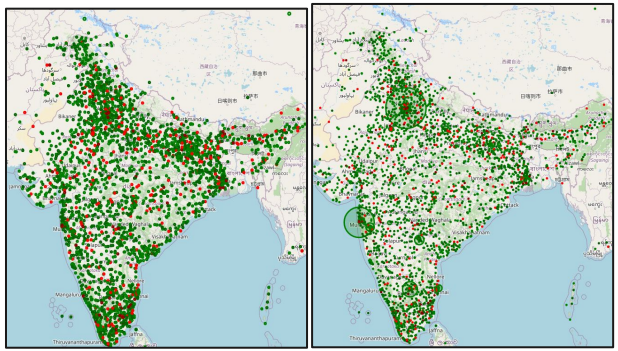


Why DBSCAN ?

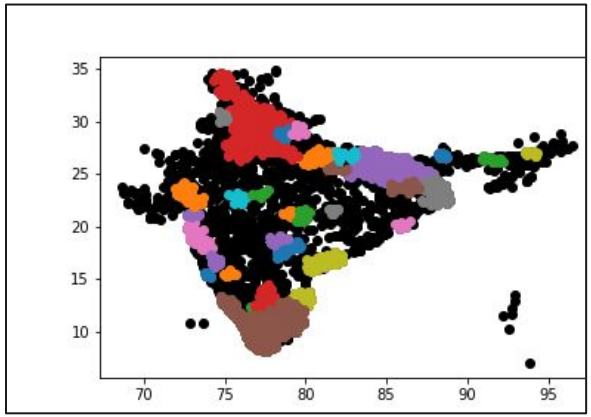
- DBSCAN works good with noisy data. Twitter data has a lot of noise.
- Our use case of density data fits well
- Our scenario of uncategorical data - is also also suited well for DBSCAN

Result I

HathrasRapeCase



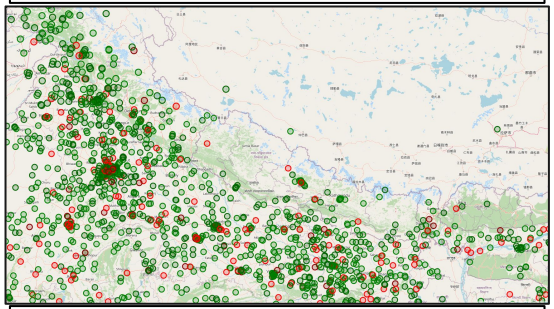
After sentiment Analysis



After Clustering

E =0.4
MinPoints=10

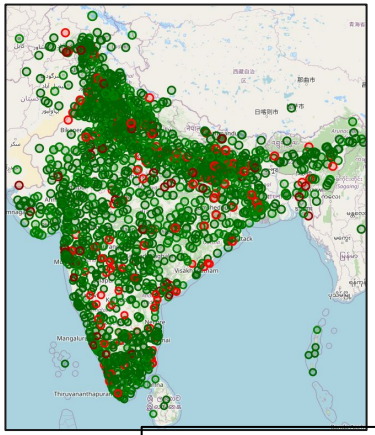
Special Mention:

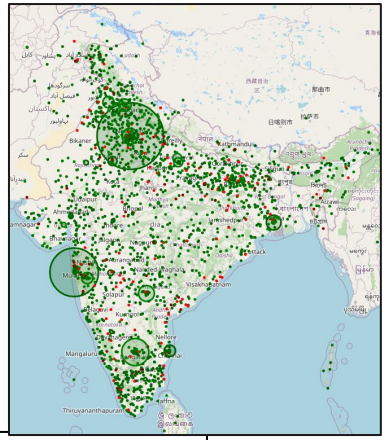


Biggest cluster in whole of UP, Delhi region

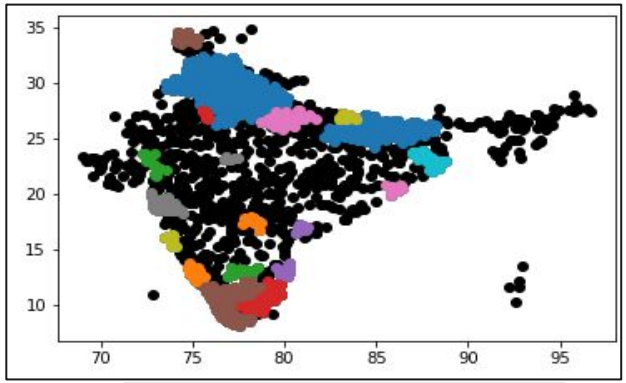
Result II

FarmersBillProtest





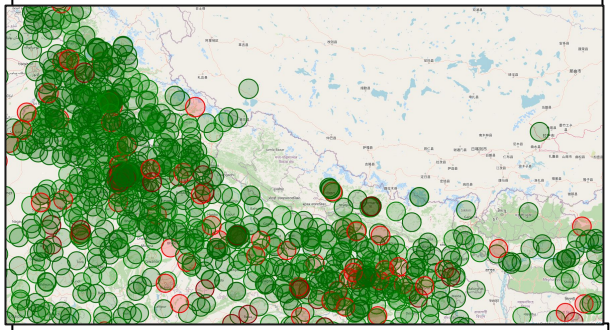
After sentiment Analysis



E =0.5
MinPoints=10

After Clustering

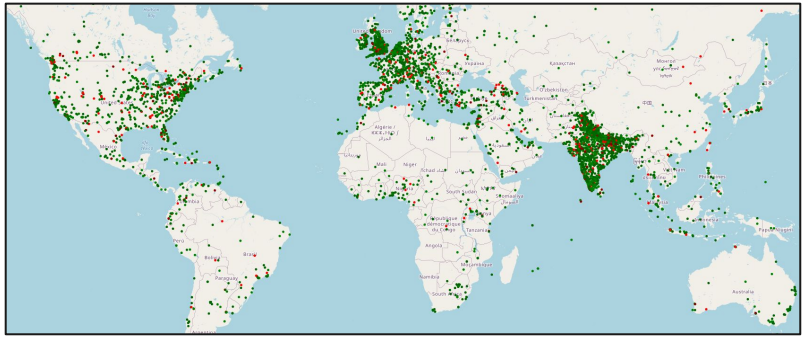
Special Mention:



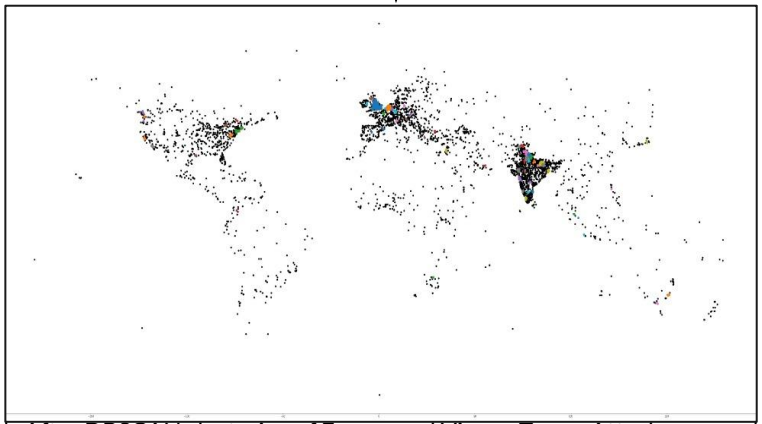
Northern india has the highest density of tweets

Result III

France&ViennaTerroristAttack



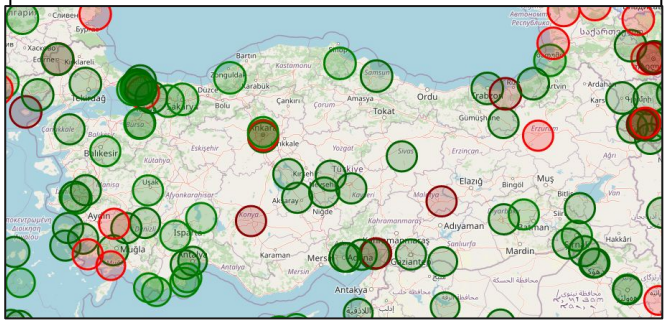
Sentiment analysis of France And Vienna Terror Attack



After DBSCAN clustering of France and Vienna Terror Attack

eps =0.8
MinPoints=10

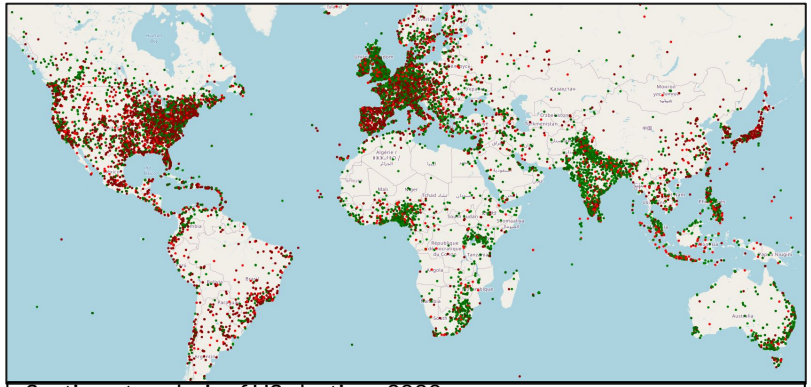
Special Mention: Turkey



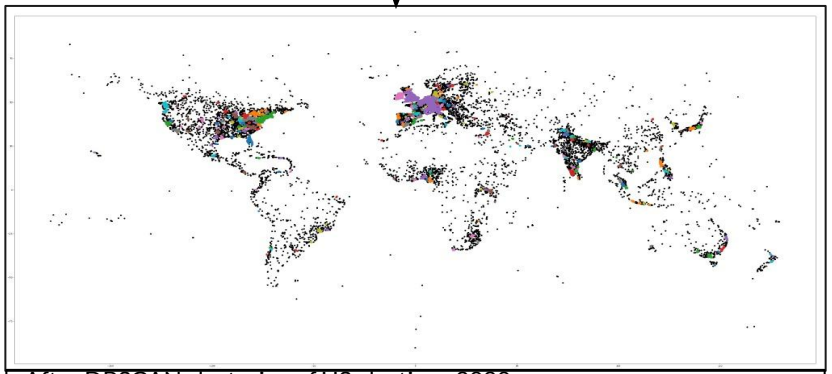
Turkey, having protested against France, do show negative sentiment tweets

Result IV

USElections2020



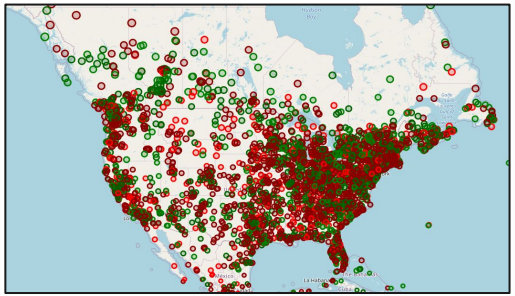
Sentiment analysis of US elections 2020



After DBSCAN clustering of US elections 2020

eps=0.5
MinPoints=10

Special Mention:



Its striking to note that negative sentiments tweets have higher density

Questions ?

Thank you