Geographical Clustering of Sentimental data on Twitter

Ujjawal Tomar[1], Vikas Tyagi[2], Vedvrat Arya[3]

[1,2,3]Students, Computer Science & Engineering Department, ABES Engineering College, Ghaziabad, U.P, India

---

Abstract - *The popularity of social media is increasing day by day and it has become a part of the daily lives of a lot of users. Social media has a large amount of data on the social web. Social data available is growing exponentially as the number of users on the social media are increasing. Twitter is a widely used platform on which users can express their views and opinions on current trends and topics. A huge amount of contextual information can be derived from Sentimental Analysis of such a popular and renowned social media platform. Further evaluations and operations can be performed to search the associated sentiments with the tweets. This research is focused on the study and classification of real time tweets to get their emotional content and performing the geographical clustering of the sentimental data to represent the tweets on the world map as per their associated sentiment.*

Key Words: **Social media, Twitter; Social, APIs, Social Web, Geo-Spatial Clustering, Density-based Clustering**

1. INTRODUCTION

Social media allows us to create connections with friends, family and customers so that we can share our views, opinions and concepts on various topics. Social media has emerged as a community where the interests of people are exchangeable and shareable. Some of the most widely used social media platforms include Twitter, Facebook, Instagram, Snapchat etc. where users can interact with others and in the meantime, a large amount of multimedia content gets created and consumed. These platforms are a part of the daily lives of a lot of users. Therefore, social media platforms including Twitter can provide us a rough picture of the thoughts and opinions of people about various trending topics in the world.

Sentimental Analysis also known as opinion mining uses text analysis and natural language processing to identify and study subjective information. The analysis of sentiments can be done at any level such as archive level determines the sentiment as positive, negative or neutral by outlining the whole record at once. On the other hand, express level deals with checking the extremity in an expression and sentence level deal with classifying the sentence in a specific class to get the sentiment associated. Sentiment analysis is an important tool that can analyze a message and tell whether the associated sentiment with the message is positive, negative or neutral. It has a number of applications in different fields such as it is highly useful in getting the customer feedback of a product such as its reviews so that the producing company can avoid those shortcomings in their future productions. It can also be used for producing suppositions about people with the help of their captured sentiments. On the industry level, it can be used to get money related reports of the organizations and surveys related to any specific item or product.

We have used hashtags and screen names to fetch the tweets. After successfully fetching the tweets and storing them in a file, data cleaning is performed to filter out the unnecessary data or images that are not contributing to our purpose. Then the polarity of the tweets is determined to know whether a tweet represents positive, negative or neutral sentiment followed by the visualization of the results in a geographic representation in the form of latitudes and longitudes. The main aim of our research is to represent the tweets on the world map after performing clustering on the basis of the location of the fetched tweets. K-Means and DBSCAN clustering algorithms are used for the clustering of tweets.

2. RELATED WORK

We have taken inspiration from a number of research papers in order to have a better understanding of the problem and to ensure that our results are more efficient and accurate than the previous researches in this field. We have tried to remove all the limitations of the previous research.

[1] This paper has shown sentiment analysis using K-Means clustering technique which landed us with the idea of using a density-based clustering technique such as DBSCAN to be more efficient with our results.

[2] This paper uses genetic algorithm in java and uses the Map-Reduce algorithm to get the best results.

[3] This paper uses a simple native method of clustering through python features due to which it has lower efficiency.

[4] This paper deals with the hashtag analysis and how fuzzy logic can be used in an unsupervised way.

[5] This paper deals with the amount of inventory that should be added as per various social media trends. Genetic algorithm has been used for this purpose. It helps in predicting the upcoming trends in the market and the stock predictions in the inventory. In this paper, the sum of squares and chi-square methods have been used for clustering. It gave us an idea of how trends and future predictions can be made in the field of social media.

[6] This paper deals with the data that contains partial class information and fuzzy logic has been used in doing so.

[7] This paper deals with how different hashtags can be represented with the help of fuzzy logic. As people generally use the same hashtags in different situations with different intentions, so clustering has been used to record sentiments and identify all such hashtags.

3. MODEL ASSUMPTIONS

Twitter is an American social networking and microblogging website on which users use tweets to interact with each other and share their views, thoughts and opinions on various topics. On Twitter, unregistered users can only read while the registered users can post, like and retweet the tweets. Users can group the posts by using hashtags and a hashtag which is used by more users in their tweets starts trending on Twitter. Twitter is a popular platform, so it has a very large number of users and hence a huge amount of information gets collected. As this information is distorted, so various operations need to be performed to get contextual and meaningful information from Twitter.

3.1 Platform Setup

We have used Python language in our research which is a high-level programming language having a simple and more understandable syntax as compared to the other programming languages. Python is widely used in software as it can easily connect to a database and create workflows. A large number of APIs are available in Python which plays an important role in communication between different components of software. It is considered to be highly useful in the field of Artificial Intelligence and Machine Learning which are among the most trending technologies in the world at the current time.

Firstly, we need to have a developer account of Twitter which provides access to the Twitter API from web and desktop applications. OAuth can also be used for this purpose. OAuth is a standard protocol that is highly secured. Twitter provides several models for fetching the tweets so that further analysis can be performed on them. Some of these models are as follows:

- Application authentication- It is entirely application based as the particular application which requires access to the Twitter API will make a request. For using this model, we need to have a consumer key and a secret key with the help of which a bearer token will be generated.

- User API authentication- In this model, a user uses an application and that application will act on behalf of the user. It requires access key and secret key by the user and a consumer key and secret key by Twitter.

3.2 Simulation Environment

The environment and packages that we have used in our methodology are as shown in the table below:

Table: Packages and Environment

Simulation	Value
DBSCAN algo	Ballt_tree
DBSCAN metric	Euclidean
DBSCAN, Eps	0.05
DBSCAN, minimum sample	5

3.3 Model Libraries

[1] Tweepy: Twitter API can be accessed by using the Tweepy OAuth method.

[2] TextBlob: It is a python library for providing access to its methods and perform basic NLP tasks.

[3] Numpy: Numpy is a python library used for working with arrays and it is also used as a container of efficient data.

[4] Matplotlib: Library which helps in the representation in the form of graphs.

[5] Pandas: Pandas is a library required for efficiently using data frames and data structures.

[6] GeoPandas: Python library for finding the coordinates in the form of latitudes and longitudes.

[7] Mplleaflet: Python library used for representing the obtained coordinates on the web map.

3.4 Dataset Generation

In order to fetch the tweets, we need to have access to the Twitter API and to do so first we need to create a developer account and after creating a developer account, Twitter provides us a consumer key and secret key and access key and access token are generated consequently. By using these generated credentials, Twitter API can be accessed. Real-time tweets are then fetched and stored in a separate file for further processes. In our research, we have fetched around 20,000 tweets that were posted using the hashtag #coronavirus. The attributes of Tweets that we have fetched from the Twitter API mainly includes ID, Tweet body, source of tweets, date of the tweet, length of the tweet, number of likes on the tweet, number of times the tweet has been retweeted and the location of the tweet. These attributes are also depicted in the table below. After fetching all the tweets, data cleaning is performed to remove unnecessary and redundant data if any. After data cleaning has been performed, we are left with 12,226 tweets. The location attribute that we have fetched is processed for finding the respective latitude and longitude which on using the mplleaflet library are plotted on a map as per their respective coordinate values.

Table: Attributes of the fetched tweet

ID	Unique Id of the user
Source	Device that has been used
Tweet	Text that has been used
Length	Length of the tweet
Location	Location of posting the tweet
Date	Date of tweet
Likes	Number of likes on tweet
Retweets	Number of retweets of tweet
Sentiment Value	Analyzed value that is 0, -1 or 1

4. IMPLEMENTATION & INTERPRETATION

We have implemented our proposed model for finding the sentiments of tweets in the following way:

4.1 Basic Framework

- We have used real time tweets for our research.
- In the first step, tweets are fetched by using the Twitter API.
- Data preprocessing is done to clean and filter out unnecessary or redundant data.
- Clustering algorithms are used to form various clusters of the cleaned data as per their obtained sentiments.
- Mplleaflet library is used for visualization of tweets on web maps.

The following diagram depicts the basic framework has been followed.

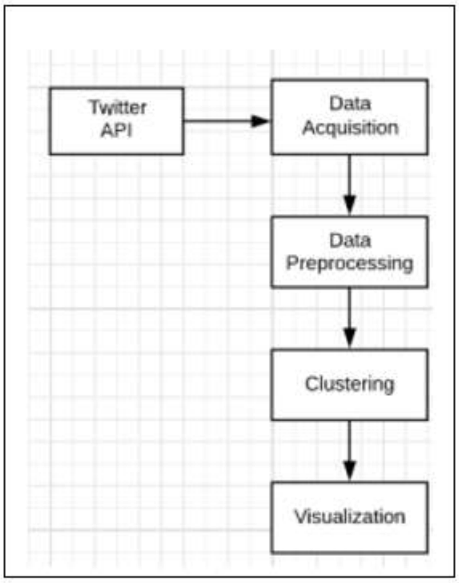

Figure 1- Basic Framework model

4.2 Execution Process

- **Data Acquisition** -In order to fetch the tweets, we need to have access to the Twitter API and to do so first we need to create a developer account and after creating a developer account, Twitter provides us a consumer key and secret key and access key and access token are generated consequently. By using these generated credentials, Twitter API can be accessed. Then hashtag or screen name can be used for fetching the tweets.

- **Data Preprocessing** - After fetching all the tweets, data cleaning is performed to remove unnecessary and redundant data if any. Links and special characters in a tweet are removed by using the utility functions. Then, data is transformed to ensure that it contains only the dictionary words for analysis. Locations that are available to us in the form of latitudes and longitudes, are converted to coordinates and finally plotted on the map.

- **Clustering**- We have fetched the location of the tweet as an attribute, that location is initially available to us in the form of latitudes and longitudes which is then converted to coordinates using the GeoPandas library. Now various clustering techniques such as K-Means, Hybrid can be applied to the cluster as per the location of the tweets. In this way, we have created the geospatial representation of the tweets.

We use K-Means clustering to form geographic clusters of tweets otherwise it would be impossible to know where the sentiment of a particular tweet is located. K-Means also provides improved efficiency and performance over the native approach that is why we have preferred to use it for clustering. The clusters formed allows us to know the regions from where maximum tweets on a particular topic have been posted along with their sentiment which indicates the trending topics of a region. K-Means clustering partitions n number of observations in k clusters by assigning a data point to its closest cluster and then determining the new cluster center by computing the average of assigned points.

We use the DBSCAN algorithm for density-based clustering as it groups together the points that are closely packed and also it is less prone to outliers. The main advantage of using the DBSCAN algorithm is that it can identify the clusters with random shapes and with great efficiency. In DBSCAN, for each point of a cluster, the neighborhood should contain at least a minimum number of points. It also marks the outlier.

K-Means is an unsupervised Machine Learning Algorithm as it tries to classify the data without having been trained with the labeled data first. On implementing the algorithm, new data can be easily assigned to the most relevant cluster. In K-Means clustering, the elbow method is used for finding k number of clusters. In the elbow method, we have to select the value of k at the elbow that is the point after which the distortion or inertia starts decreasing at a linear rate. Basically, we try to find the value of the sum of square error by running the K-Means clustering algorithm for the values present in the dataset. The main aim of the elbow method is to find the optimum number of clusters for a given dataset. In this way, we find the clusters for 12,226 tweets left with us after the process of data cleaning.

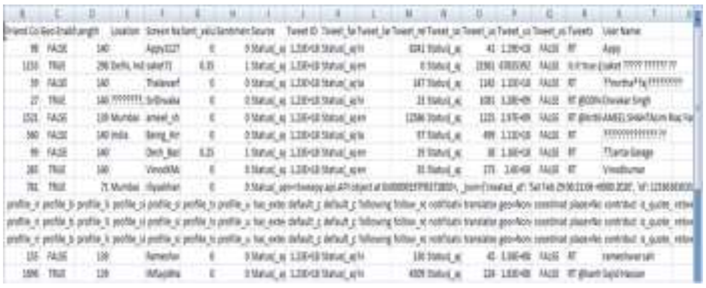

Fig -2: Sample data collected from the fetched tweets

- **Visualization** – Figure 2 shows a sample of the data that we have fetched using the Twitter API. We have represented the sources from which the tweets were posted on a pie chart shown in Figure 3 and the sentiments of the fetched tweets that is positive, negative and neutral are represented in the form of a pie chart shown in Figure 4. The coordinates we have found after converting the latitudes and longitudes are now plotted on the map using mplleaflet library.

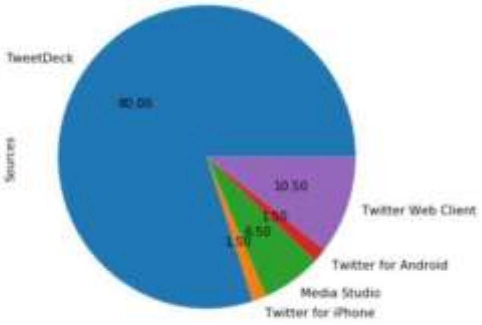

Fig -3: Pie chart for generating sources of tweets

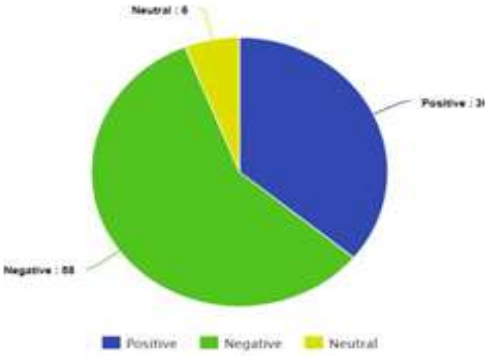

Fig -4: Pie chart representing negative, positive and neutral sentiments

4.3 Results & Outcomes

The main aim of our research is to find the coordinates for clustering rather than limiting our research to the longitudes and latitudes so that we can obtain a precise location from where the tweets were posted and represent them with the help of heat maps. Clusters that we have formed using DBSCAN and K-Means clustering algorithms have been visualized. Finally, we have represented the tweets as per their sentiment that is positive, negative and neutral on the maps as shown in the adjacent figures. Here we have shown the results of #coronavirus fetched from the Twitter API.

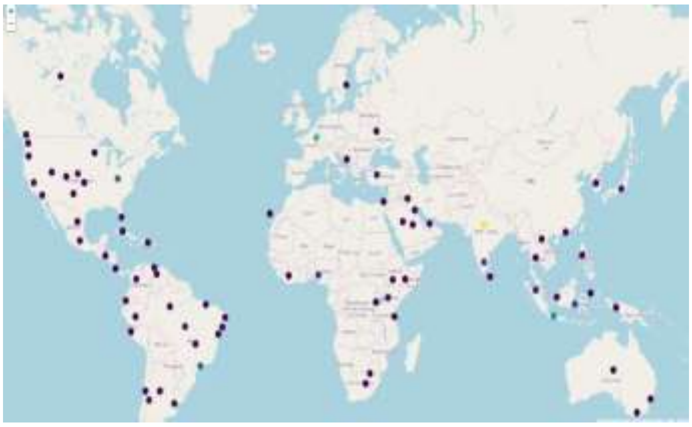

Fig -5: Representation of Neutral Tweets

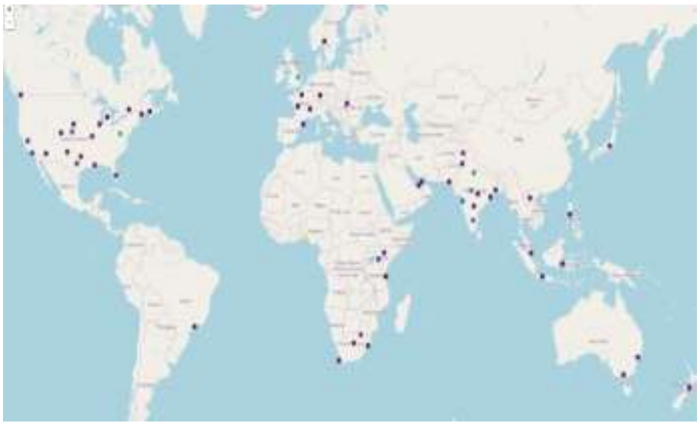

Fig -6: Representation of negative tweets

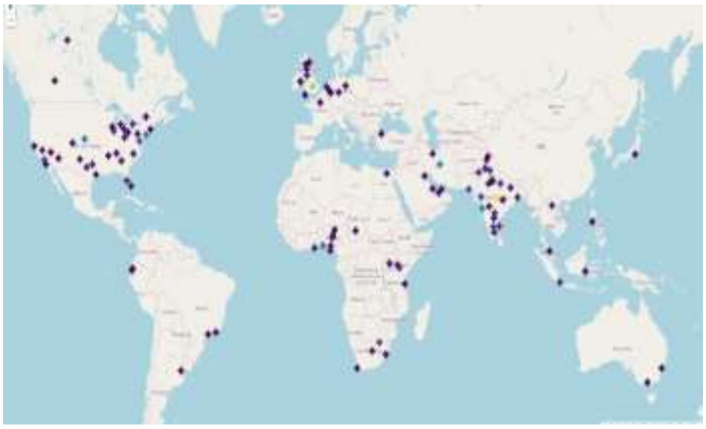

Fig -7: Representation of positive tweets

Figure 5 shows the representation of neutral tweets on the map, Figure 6 shows the representation of negative tweets on the map, Figure 7 shows the representation of positive tweets on the map. Figure 8 shows the total twitter activity of the users on the map.

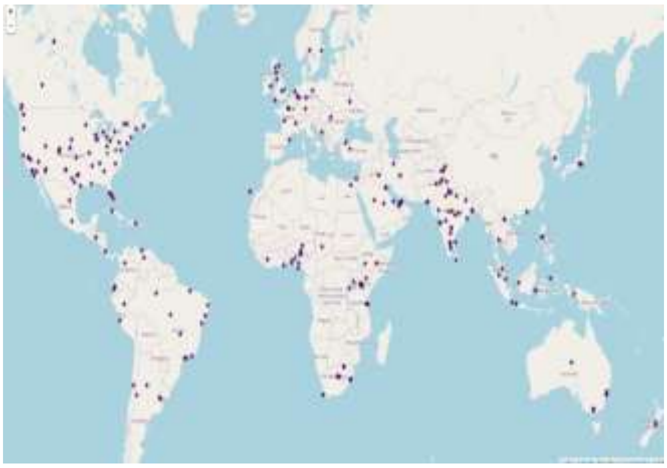

Fig -8: Representation of total twitter activity of users

REFERENCES

“

[1] “Social media generated big data clustering using genetic algorithm,” 2017 ICCCI, Coimbatore, pp. 1-6, 2017.

[2] “Real-time clustering of tweets using adaptive PSO technique and MapReduce,” by A.P. Chunne, U Chandrasekhar in Global Conference on Communication Technologies (GCCT), pp.452-457, 2015.

[3] “Clustering and sentiment analysis on Twitter data,” by S. Ahuja and G. Dubey in 2nd International Conference on Telecommunication and Networks (TEL-NET), pp. 1-5, 2017.

[4] “Analysis and Visualization of Twitter data using K-Means clustering” by N. Garg and R. Rani in ICICCS, pp. 670-675, 2017.

[5] “Social Media generated big data clustering using genetic algorithm” by P. Sachar V. Khullar in ICCCI, pp 1-6, 2017.

[6] “An unsupervised fuzzy method for Twitter sentiment analysis” by S. Suresh and Gladston Raj S.in CSITSS, pp 80-85, 2016.

[7] “ST-DBSCAN: An algorithm for clustering spatial-temporal data,” Data& Knowledge Engineering, Volume 60, pp 208-221, ISSN 0169-023X, 2007.