

An Accurate method for clustering Hindi tweets on a Geo-spatial area

YashviMalhan¹ and NamitaGoyal²

¹Information Technology, Maharaja Agrasen Institute of Technology, India

²Information Technology, Maharaja Agrasen Institute of Technology, India.

Abstract: The social networking sites have tremendously captured online communication over the social web. With the growth in number of users on social networks, the social data has also grown exponentially. One of the predominantly used social networking sites include Twitter. It is one of the most authenticate social platform that allows users to express their views on current trends and topics. The language should not be a hurdle to get correct and accurate results. The project will now deal with not only English tweets but also Hindi tweets. Sentimental analysis of such dynamically changing user behaviour upholds huge amount of contextual information. It will also notify the sarcasm and irony to give the optimum results. I am trying to find these emotions and sentiments of the people on Twitter. My research is focused on pre-processed analysis and classification of real-time tweets, based on the emotional content. The novel approach applies density- based clustering with longitudinal locations from the tweets to reveal social communities for sentimental analysis.

KEYWORDS: Social Media; Twitter Science; Social Communities; Geo-Spatial Clustering; Density-based Clustering; Hindi Tweets; Sarcasm Detection



Check for updates

DOI of the Article: <https://doi.org/10.46501/IJMTST0706016>



Available online at: <http://www.ijmtst.com/vol7issue06.html>



As per **UGC guidelines** an electronic bar code is provided to seure your paper

To Cite this Article:

YashviMalhan and NamitaGoyal. An Accurate method for clustering Hindi tweets on a Geo-spatial area. *International Journal for Modern Trends in Sceicen and Technology* 2021, 7, 0706101, pp. 92-98. <https://doi.org/10.46501/IJMTST0706016>

Article Info.

Received: 9 May 2021; Accepted: 2 June 2021; Published: 11 June 2021

1. INTRODUCTION

The emergence of online social media platforms allows creation of social connections with family, friends, acquaintances, customers etc. to enable sharing of opinions, views, thoughts and ideas. Such complex and large-scale communication interconnection structure evolves as social network. The growth of social media has introduced new types of social interaction, that users will affiliate themselves into the community where their interests are often shareable and exchangeable. An on-line social networking website such as Facebook, Twitter, Instagram, Tumblr etc. is a place where a user can interact with other people and can create and consume a significant amount of multimedia content. Many users have incorporated these websites into their daily lives. Sentimental Analysis[12] is a procedure of analyzing social trends and breaking down the information in view of one's sentiments, surveys and musings. Sentimental examination is frequently called as opinion mining as it mines the critical highlights from individual assessments and points of view. The analysis should be possible at any level. Like, in archive level, outline of the whole record is taken first and afterward it is examined whether the opinion is positive, negative or neutral. In express level, examination of expressions in a sentence is considered to check extremity. In sentence level, each sentence is then also characterized in a specific class to dissect the feeling. This has different applications in numerous fields. It is utilized to produce suppositions for individuals of on-line networking by examining their sentiments or opinions which they give in type of content. Such investigative results could be further applied to different applications over various domains. It is utilized to produce suppositions for individuals of on-line networking by examining their sentiments or opinions which they give in type of content. Consequences of one space can't be connected to another area. It is utilized as a part of some genuine situations, to get surveys about any item or motion pictures, to get the money related report of any organization, for expectations or promoting. In this project, the tweets will be fetched by either using hashtag, handle or screen name and are refined, analyzed and visualized in a geographic representation using precise location, in form of longitude and latitude from the fetched tweets. The main aim of the research is

to visualize the tweets in Hindi on the basis of the location information, thereby clustering sentiments over geographical span. My project will also notify the sarcasm and irony to give the optimum results. The tweets are classified according to k-means[9] and DBSCAN[10] clustering techniques. They are then compared to justify the technique for this research as optimum. But because twitter is full of Hindi tweets and sarcasms we have to find out more accurate clusters is important.

2. RELATED WORK

The literature survey of this research consists of many research papers from reputed journals which have inspired to do this work in a better and effective way. All the shortcomings and the benefits of all these papers have collectively led to the final output depicted by the paper. The paper mentioned in [1], acted as main motivation for us as they had shown the sentimental analysis using k-mean clustering, which gave us an idea over how we could take it further by adding DBSCAN[10] to it and showing the more optimum one out of these two. In [8], a fuzzy logic to investigate how each of the clusters represents a set of differently behaving hash tags is represented. Basically, we have a lot many different hash tags used by people that bears similar interpretation but, have been used differently as hash tags. So, to get a record of all types of hash tags this clustering was used to record sentiments. In [5] the paper deals in simple data analysis and clustering using different python features, and hence results in clusters. and hence assesses its optimality by F-measure and execution time. And in [3], the paper was used to predict the stock of inventory to be added, according to the trending social media responses using the genetic algorithm.

This works in the benefit of business market which predict how much market is in trend and how will it be in coming future thereby, giving major stock predictions in inventory. For clustering, chi-square and sum of square distances has been used. Aditya Joshi and Bhattacharya[10] devised a three step model for SA that incorporated in-language sentiment analysis, machine translation and the development of a lexical resource called Hindi SentiWordnet(H-SWN). Mittal et al(2013) later used the H-SWN for classifying a corpus of Hindi reviews which incorporated negation handling

and discourse analysis as well as a method for augmenting the original H-SWN. Sentiment analysis in Hindi is still a fledging work, and no work has been done on SA of Hindi Social network Data.

3. MODEL ASSUMPTIONS

Twitter is a free social networking microblogging website that users use to post and communicate through messages, called tweets. The measure of information collected on twitter is extremely enormous. This information is unstructured and written in common dialect. The users will broadcast tweets and follow different users; tweets through multiple platforms and devices. Twitter science allows to keep track of information diffusing rapidly that can be fetched from any geographical location within the world.

A. Platform Setup

In this paper Python language is used. Python is an interpreted high-level programming language for the general-purpose programming. The syntax of this language is very simple as compared to English language. The use of python is diverse in the field of software to create workflows, can be used to connect to the database, to handle big data and perform complex mathematics. It is also used in Big Data and Cloud Computing. Also for accessing Twitter API, OAuth[14] is used in the research. OAuth provides authorized access to the Twitter API[13] in a simple and standard technique from web, mobile and desktop applications. OAuth is an open source secure and standard protocol. API stands for Application Programming Interface, could be a set of methods which helps in the communication between different software components. For accessing and gathering the tweets, twitter provides various API authentication models.

User API authentication model- An application acts on behalf of a user, as the user. User authentication requires access to consumer key and secret from the twitter application and access key and secret from the user the application is trying to act on its behalf.

Application only authentication- An application makes the request on behalf of its own. In order to use this method, a bearer token is generated by passing the consumer key and secret through POST OAuth endpoint.

Model Libraries

Numpy: It is a fundamental package for computing scientifically.

Pandas: It is an efficient library for using data structures and data frames efficiently.

Tweepy: This is a library for using Twitter API.

vaderSentiment Module: It is a lexicon and rule-based analysis tool that is specifically attuned to sentiments expressed in social media, and used for analyzing the sentiments of Hindi tweets.

Matplotlib: It is a 2D library, which helps in effectively representing the graphical presentation in forms of graphs. Basically, for plotting graphs.

TextBlob: Python library for processing textual data. It provides a simple API for diving into NLP tasks.

Geopy: Python library for finding the geo- locations and hence the coordinates.

Google API[10]- This is an Application Programming Interface developed by Google which allow communication with Google services and other services. They help in accessing user data and provide functionality to data analytics and Machine Learning processes.

C. Dataset Generation

Real-time tweets are streamed with the help of Twitter streaming API. To fetch the twitter data, the user should have a twitter account. The tweets are acquired by using Twitter API. To authenticate the account a developer profile is created through which the consumer key and consumer secret and the access key and access token are generated. With the help of these credentials Twitter API can be accessed to acquire the data. The data that is collected is around 10,000 tweets. The data retrieved is in excel format. Later, excel file is created in which the acquired data is stored and is also appended simultaneously by removing the data duplication. The data is acquired with the input of a screen name or hashtag, which hence represents the data of tweets. The tweets also carry a location with them which helps us to fetch the clusters in our model. These locations fetched are processed to find their Latitudes and Longitudes. These Latitudes and Longitudes are embedded into Google API, which give us their coordinate value. These coordinates process up on maps and form clusters relating to tweets.

4. IMPLEMENTATION & INTERPRETATION

The following section includes the implementation and interpretation of our proposed model for depicting social sentiments on the basis of Twitter comments and posts.

D. Basic Framework

- As already mentioned the first step here would be fetching real time data from twitter.
- The next step would be data acquisition through twitter API.
- This is followed by data filtering of Hindi Tweets.
- After that sentiments of the tweets are found out.
- Then using a sarcasm detector machine, sarcasms in the Hindi Tweets are detected.
- Then locations of the users are fetched which we get in the form of latitudes and longitudes.
- Hence, the data is being used for making various clusters using different algorithms.
- For visualization of tweets on Google maps we represent them as heat maps.

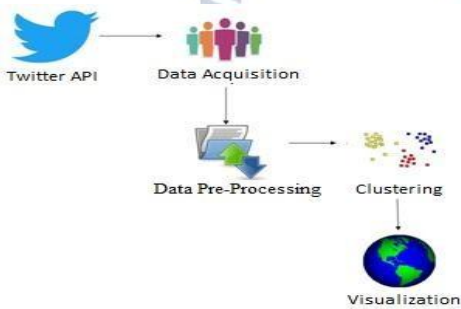


Fig. 1: Flow Diagram of Proposed Model

E. Execution Process

- Data Acquisition - Real time tweets are streamed over internet with the help of twitter streaming API.
- Data Filtering of Hindi Tweets- Under data filtering, the data is chosen to show only the particular results i.e. Indian languages like Hindi for further analysis. Textblob library is used for data filtering. Data is hence filtered to show the tweets only in the Hindi language.
- Analysing the Sentiments of Hindi Tweets- If during analysis polarity of the tweets is found to be greater than 0, then it is classified as a positive tweet, if polarity is equal to 0 then neutral tweet else negative tweet. Compound polarity is used to indicate tweet is a mixture of some part of positive or some part of negative or some part of neutral or all.
- Analysing Sarcasms - Human language has a lot of variations in it. In order to express themselves people

tend to use sarcasm or mixed feelings. While analyzing tweets, we need to understand this sarcasm in order to find the accurate sentiments.

Coordinate Finding & Clustering-Twitter allows its user to provide its location while posting a tweet. Each tweet’s location consists of the latitude and longitude to indicate its location precisely, which are converted to coordinates to help further in clustering. There are various techniques such as k-Means, Hybrid are used to cluster the data according to the tweet’s location to create the geospatial representation. In this research, k - Means[9] and DBSCAN[10] methods are used. DBSCAN stands for Density Based Spatial Clustering of Application with Noise. Based on a set points, groups together the points that are close to each other based on the distance measurement and minimum number of points. It also marks the outliers the points that are in low- density region. The k-means is an unsupervised learning algorithm, which is used to solve the clustering algorithm.

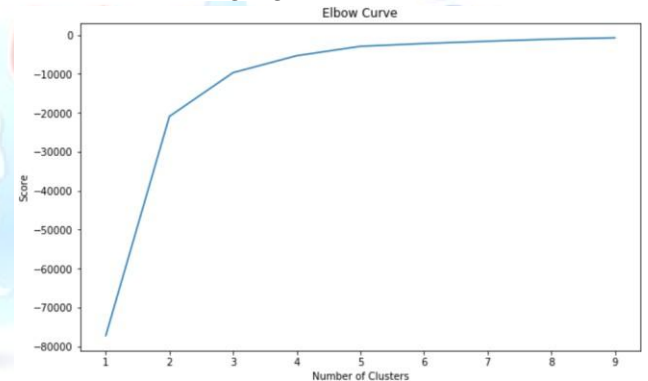


Fig.2. Experimental Value of k with Elbow Method

	Tweets length	ID	Source	likes	Retweets	date	location	sent_value	sentiments
0	विक्रमी सरकारों ने बिल लाने पर श्रेष्ठ सम्मान से उल्लेखित	139 1024126705172570250	Twitter for Android	0	0	2018-07-31 02:57:02	Haldwani	0.0	0
1	RT @PMOIndia: देश के पूर्व उपसर्गों अर्थात् बि...	139 10241267051236600	Twitter for Android	0	1513	2018-07-31 02:56:59		0.0	0
2	RT @rajkumar0973: #ManKBeath...सम्मानित	140 1024126744821784576	Twitter for Android	0	247	2018-07-31 02:56:57		0.0	0
3	@LalGanguli @cuttingedge2019 Employees of Kin...	137 1024126740531044352	Twitter for iPhone	0	0	2018-07-31 02:56:56		0.8	1
4	RT @smath_dhiveedi: Breaking news... 10/15 अर्थात्...	140 102412673688077093	Twitter for Android	0	20	2018-07-31 02:56:56	राजमहाराजपुरा	0.0	0

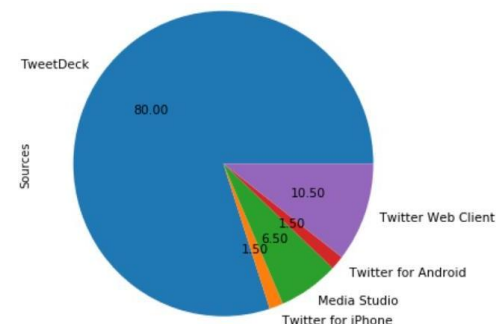


Fig.3. Classification of Tweets as per Generating Sources

For finding , i.e the number of clusters we use elbow method. So, as we have a big data, we have the arbitrary approximation ratio, probability of success, or distribution of the samples. To find the most its main aim is to classify the dataset through a certain number of clusters (k-clusters). Each observation of the dataset belongs to the cluster with its nearest mean. The learner must be able to learn the concept given any accurate number of centers for the clusters we have used the elbow method. Observations as shown in figure 2, where we have found the value of k which is equal to 3 and 4. The data which we have is used to find clusters is around six thousand five hundred tweets which form clusters according to the density.

□ Visualization - A data frame in Figure 5, is showing all the necessary information of the tweet along with sentiments of the tweet as well as location in form of latitudes and longitudes. Figure 3 depicts a pie chart that shows the sources from what device the tweet has been done. Figure 4, represents the pie chart representing the positive, negative and neutral tweets. The pie chart is been shown by using a pie function of python. The clusters are displayed on python console using pie function. The two pie charts represent the percentage of sentiments of tweets as well as the percentage of sources from which the tweets have been done. The coordinates we had found now help in showing the geo-locations on Google maps in form of heat maps.

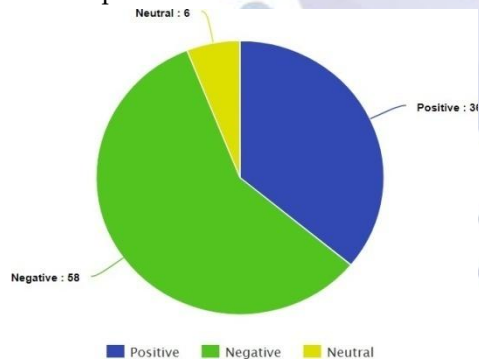


Fig. 4. Classification of Tweets as Positive, Negative Neutral Sentiments

F. Results and outcomes

The screen name of Hindi Divas is used for fetching our data and sentimental analysis. The result is first a graphical representation of all the tweets we have gathered while the whole process. Also, then it is followed by the visual clusters formed by using K-means algorithm. The new idea behind my project is

to find sentiments of Hindi Tweets and then find sarcasm behind these tweets. Later while in place of finding the latitudes and longitudes, we find the coordinate points for clustering; Hence, we get an accurate location without breaching any secure data of locations in terms of sentiments and locations both. Later the results were based on all the analysis being done. Heat maps thereby depict the sentiments and map in fig.6. Shows the clusters at all the locations according to their responses of the users.

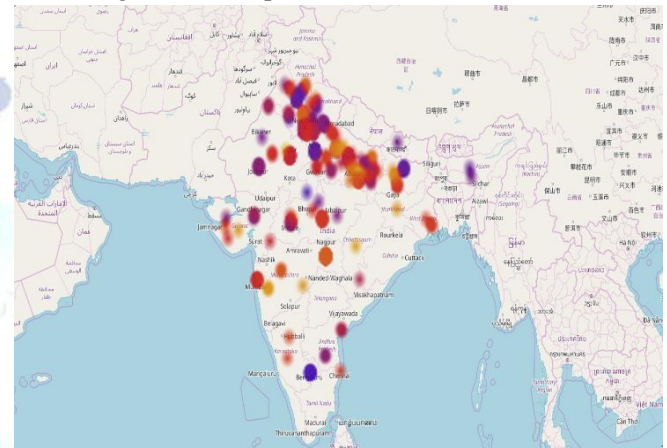


Fig.6. Geo-Spatial Highlights of NeutralTweets

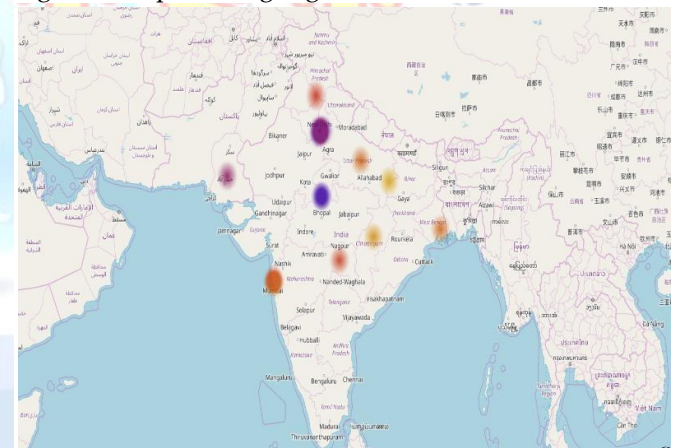


Fig.7. Geo-Spatial Highlights of NegativeTweets

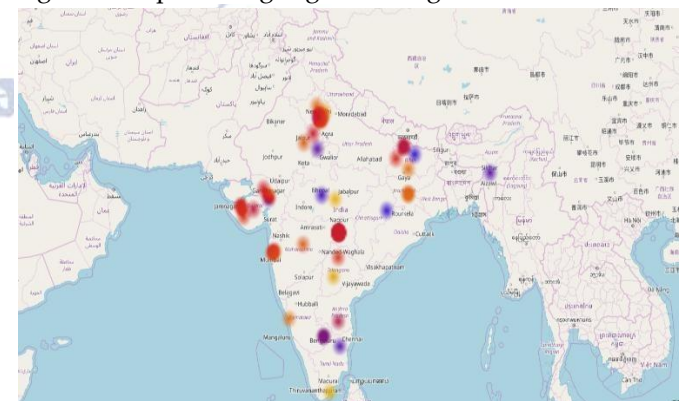


Fig.8. Geo-Spatial Highlights of Positive Tweets

Figures 6-8 illustrates the neutral, negative and positive tweets respectively. In figure 9, the color schemes for showing the density of twitter activity is highlighted.

The most dense area is being shown by violet color, then followed by indigo, then, blue and so on in format.

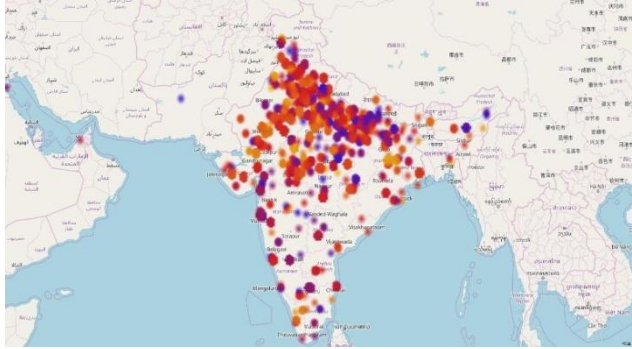


Fig.9. Geo-Spatial Clusters with Chromatic Range Highlighting Density of Twitter Activities

The main objective of the paper was to show the translation of the tweets in the Hindi language and analyze the sarcasms of those tweets. I have used the most optimum clustering techniques from the techniques we have used in here. Clustering is one of the most important steps while data mining. It is the process of grouping data items based on similarity between elements. K-means is a numerical, unsupervised, iterative method. It is simple and very fast, when the data set is small, it is proved to be a very effective way that can produce good clustering results. It cannot produce effective results when data set is large. Algorithm gives good results in small as well as large data sets. Substantially outperformed standard k-means in terms of percentage of correctness.

To find the most optimum we need to assess the accuracy, which is defined as the ratio of correctly clustered tweets to the total tweets collected. Existing research in "Analyzing Twitter Data Using Unsupervised Learning Techniques by Lakshmi Devi and K.S Rividya" shows the correctness of k-Means and with increasing number of records (Table 1). The range of records here is uneven to know the behavior of each clustering technique. It is quite evident that because of less no of tweets, the algorithms are quite accurate while increasing the number of records decreases the accuracy consequently. Hence, the clusters formed are shown in figures 10-15, that represent the clusters formed by -Means and respectively. The k-means clustering is performed with centers. Hence, it is clearly depicted even by the clustering diagrams that DB-Scan provide much more clear clusters with assigning clusters even to the outliers. These clusters are formed by the tweets we

fetches and hence forming clusters according to centers found.

TABLE I. ACCURACY OF CLUSTERING TECHNIQUES

No of records	Percentage of correctness for K-means	Percentage of correctness for DBSCAN
10	100	100
20	98.3	98.9
50	96.5	98.6

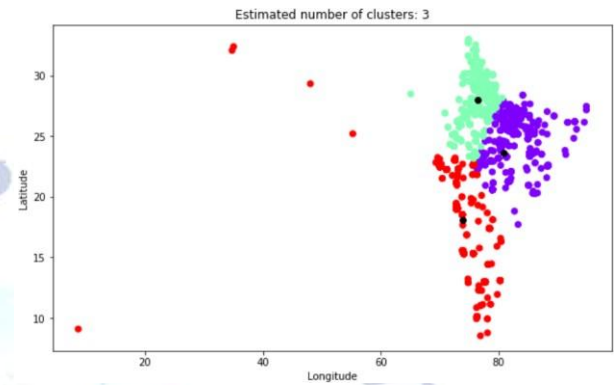


Fig.10. Geo-Spatial Cluster Formation with k-means having 3 Clusters

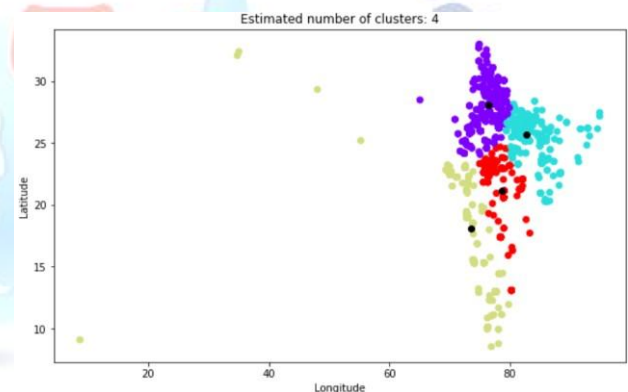


Fig.11. Geo-Spatial Cluster Formation with k-means having 4 Clusters

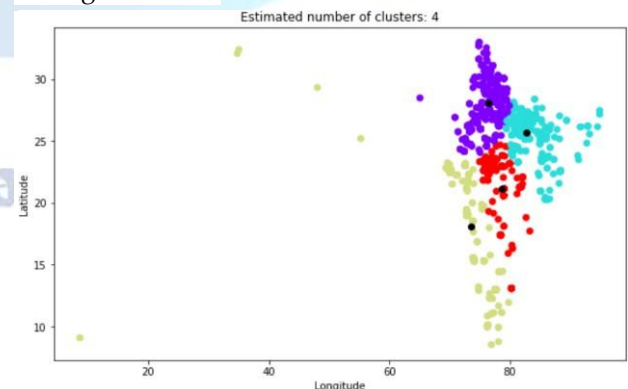


Fig.12. Geo-Spatial Cluster Formation with k-means having 5 Clusters

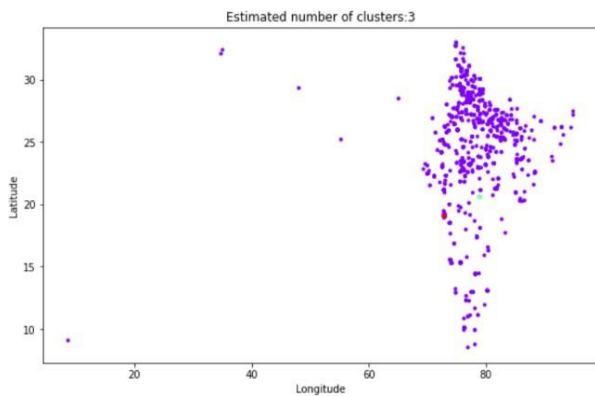


Fig.13. Geo-Spatial Formation with DBSCAN having 3 clusters

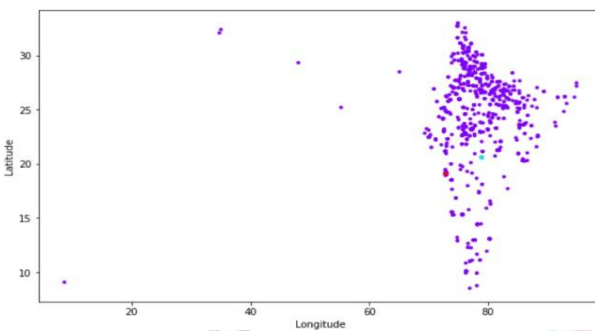


Fig.14. Geo-Spatial Cluster Formation with DBSCAN having 4 Clusters

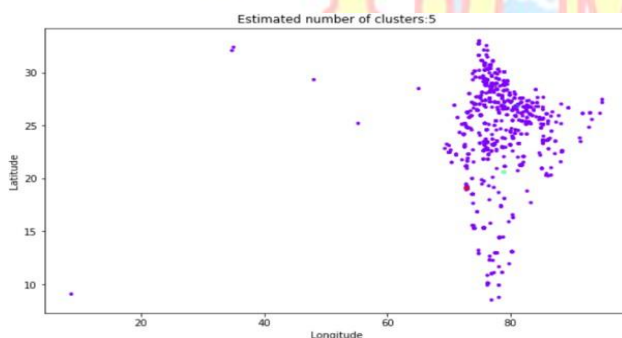


Fig.15. Geo-Spatial Cluster Formation with DBSCAN having 5 clusters

Hence, the observations recorded in table 1 and diagrams in 10-15 jointly interprets that provides a better accuracy in finding the clusters. The results showed that means clustering possess no provision for outlier-separate clusters. However, I have made separate cluster for outliers making it more accurate and convenient to further apply context-based sentimental analysis. 5.

CONCLUSION

Twitter is henceforth the most used social network as well authenticating media platform to share contents online. Analysis of the colossal social data on twitter can be processed and mined to extract relevant information

that helps in commercial field as well as social up fronts for some important future predictions. Also, any social issue can be raise up and by knowing the locations, an area that needs to be circulated more for a topic can be detected easily and worked upon. Thoughts of people are being circulated widely making it easy to know their sentiments over some topic, personality or any social issue In this project, I have performed analysis and classification of sentiments in tweets on the basis of longitudinal information being extracted from each of the tweets collected. In my project, I realized that the model built was not perfect on account of sarcasm as well as the common languages used by most of the people. Therefore, it became necessary to do further research on the sarcasm detection engine that should be built as well as work should be done on the Hindi tweets to make it a perfect model. So, in my major project the tweets in Hindi language are also analyzed. Hence, a sarcasm detection engine for tweets is made with the motivation to detect sarcasm based on textual features and emojis. As future extension of my research, smart GUI are being designed for generation geo-spatial clusters that could dynamically capture the trends and precise locations for high density of tweets from a specific region of interest.

REFERENCES

1. N. Garg and R. Rani, "Analysis and visualization of Twitter data using k-means clustering," 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 670-675.
2. P. Sachar and V. Khullar, "Social media generated big data clustering using genetic algorithm," 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2017, pp. 1-6.
3. E. N. Desokey, A. Badr and A. F. Hegazy, "Enhancing stock prediction clustering using K-means with genetic algorithm," 2017 13th International Computer Engineering Conference (ICENCO), Cairo, 2017, pp. 256-261.
4. A. P. Chunne, U. Chandrasekhar and C. Malhotra, "Real time clustering of tweets using adaptive PSO technique and MapReduce," 2015 Global Conference on Communication Technologies (GCCT), Thuckalay, 2015, pp. 452-457.
5. S. Ahuja and G. Dubey, "Clustering and sentiment analysis on Twitter data," 2017 2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, 2017, pp. 1-5.