JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANALYSIS AND LOCATION PREDICTIONOF THE DATA FROM THE TWITTER ACCOUNT USING MACHINE LEARNING

Chaithra D
Department of computer science and Engineering
Siddaganga Institute of Technology
Tumakuru, India
chaithradinesh86@gmail.com

Dr. A S Poornima Ph.D
Department of Computer Science and Engineering
Siddaganga Institute of Technology
Tumakuru, India
aspoornima@sit.ac.in

Abstract- This Paper is mainly aiming to find the location from the geo –map which has been predicted from the twitter data. When the user who posted the tweet on twitter for various purposes like territory knowing or an illegal messages, it is necessary to know the location where the data has been posted. This will help to track and reduce the crime activites.For this activity KNN to find the location of the message and SVM to classify the message have been used.Location information will be in the form of country/city names with hashtags.

Keywords—Twitter data , Location Prediction,GPS enable.

I. INTRODUCTION

Even twitter is so popular to connect people all over world with messages they tweted on their account and the thing here less than 1% of the tweet will have the location tagged and posted but most of the posts which are unsual and unwanted which are not good to society these kinds of data will contain the location, this made researchers to to develop the prediction of the location based on twitter data, if the data are harmful. The location prediction consists of location indication words,city/country name, hashtags and tags mentioned.

On organized information investigation activity can be effectively performed and the outcome can be gotten without any problem. However, if there should arise an occurrence of unstructured information from E-mail, Twitter and so forth, it is very hard to finish up the yield due to different issues, for example, virtual commotion impact and vague information. In this paper, we take a gander at one such mainstream miniature blog called Twitter.

Deciding the area of an online media client and where a message is posted from is significant for area based proposal (Ye et al., 2010), emergency identification and the executives (Sakaki et al., 2010), identifying area driven networks (Lim et al., 2015), socioeconomics examination (Sloan et al., 2013) furthermore, directed publicizing (Tuten, 2008). This work plans to allocate a topographical area (generally plausible area from a rundown of pre-characterized areas, like urban communities or nations) to a piece of text. For this text based content, we center around the Twitter interpersonal interaction site, which helps in excess of 500 million tweets.

II. PROBLEM OVERVIEW

The overall survey is based on the location detection of the twitter data , here the overview of the platform as been explained in the view of ordinary users. Basically data will be seen 3 angles i.e.., content,network and context. So here location problem as taken into consideration to avoid the cyber byulling aspects. The below mention fig 1 shows various kinds of location predictions.

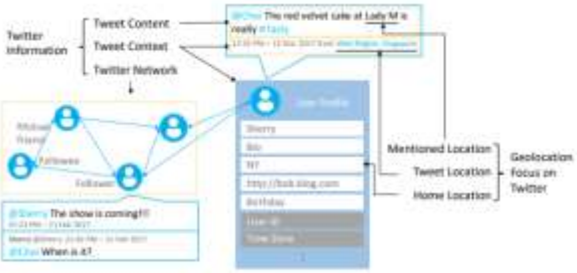

Fig 1. A delineation of tweet content, tweet setting, and Twitter organization, and the three sorts of areas: home area, tweet area, and referenced area in Twitter

On organized information investigation activity can be effectively performed and the outcome can be gotten without any problem. However, if there should arise an occurrence of unstructured information from E-mail, Twitter and so forth, it is very hard to finish up the yield due to different issues, for example, virtual commotion impact and vague information. In this paper, we take a gander at one such mainstream miniature blog called Twitter.

A. MENTIONED LOCATION PREDICTION

AS lot many of users are using tweets and follow their icons each and every one has there own to write the tweet content. Proposed location prediction will also helps to understand and classify the messages whether tweet message is proper or improper. In this method we also involve two sub tasks of mentioned location prediction:

- Recognition of the location mentioned and tagged in tweet.
- Due to Unstructured data in the tweet the location mention is disambiguation, find the location in such tweets label and non-label method I adopted.

III. PROPOSED METHODOLOGY

It is important to examine and classify the tweet data is positive or negative, the user can tweet the data based on any domain such as government, education, social anticipation based issues. The main consideration as to be taken here is what technologies and ways to classify tweet as positive and negative. 1. AI 2.Dictionary based method 3. Dictionary based method is to compare tweet with words present in the dictionary.

In this proposed methodology based on the tweets we are also monitoring the location, basically the tweet once received if it is harmful tweets we will check from where the tweeter has been received for that we are using geolocation tracking technology.

A. Principle Contributions

Our principle commitments as:

1. Here we present geography based location monitoring for twitter data dependent on ML algorithm for classification , content based feature extraction which are picked from twitter data.

2. As it has been discussed in previous section in location extraction based on tweet the fetures we mainly concentrate on words, city/country and hashtags.

3. More than a million tweet datas has been collected from 77k and odd clients and based on these data classification will takes place as tweet is negative or positive and later work will demonstrated.

B. Design and Organization

The rest of the paper is coordinated as follows. Futher Area depicts our proposed approach, including data pre-planning, incorporate set decision, model getting ready, and appraisal. Fragment 3 presents the preliminary outcomes of our proposed approach and various baselines. Fragment 4 summarizes our paper and highlights some future headings for geolocation estimate.

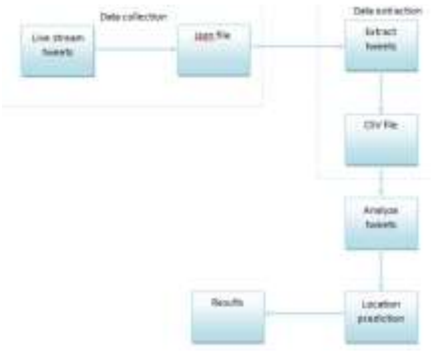

Fig.2: Flow Diagram

IV. PROPOSED APPROACH

In this part, we depict our proposed way to deal with the geolocation expectation of Twitter clients and tweets. Our proposed approach includes three primary stages, in particular:

(I) Information pre-preparing to recognize the arrangement of literary highlights;

(II) Model preparing to prepare our expectation calculation dependent on multinomial Innocent Bayes classifier; and

(III) Assessing our forecast calculation on the turn of events and testing sets.

A. Information Preprocessing and Feature Set Selection

The method of converting data from unstructure to structure we follow: 1.converting data to lowercase, 2. Remove unwanted letters, 3. Tokenizing data to singular words.. These handled tweets are then utilized as contribution to our multinomial Naive Bayes classifier, where the use recurrence tally of a bunch of highlight words is gotten from these prepared tweets. We currently portray the different capabilities utilized in our trials, which are:

- ✓ Location Indicative Words
- ✓ City/Country Names
- ✓ Hashtags
- ✓ Mentions
- ✓ Combination of all the above

B. Information Pre-handling

In the pre-handling step, we eliminated the immaterial Twitter information. To start with, we just kept tweets that were in English in the information stockpiling. After the interpretation of tweets into English, the leftover tweets having a place with some other language were taken out. To eliminate Roman tweets, we looked for strings including ther most normal words utilized Moreover, we eliminated URLs from tweets, since URLs direct to additional data that was not a prerequisite for slant examination in our methodology.

C. Training of multinomial Naive Bayes classifier

The Naïve's Bayes is a social event controlled learning calculation. It depends upon restrictive likelihood hypothesis to pick the class of another part vector. The NB utilizes the arranging dataset to discover the restrictive likelihood respect of vectors for a given class. In the wake dealing with the likelihood restrictive appraisal of every vector, the new vectors class is figured out subject to its chance likelihood. NB is utilized for content concerned issue depiction.

D. Geocoding User Location information:

An interaction called GeoCoding is utilized to change over the got areas into Latitudes and Longitudes.Now you have oneself revealed areas of the record's Twitter supporters, yet you can't plan them since we don't have any advantageous information to disclose to R where/how to plot an passage. Geocoding of areas fixes the issue.

1) ggmap bundle in based on program: ggmap bundle incorporates the geocode() work that permits access the Google Maps API without leaving programming language to utilize this usefulness to get data using API of google and geocode the areas. It is important to eliminate any examples of % since that character doesn't play well with the API.

2) Google designers console: Create the space in developer account , one it is done successfully we get respective keys for authentication. In the event that this progression is skipped, enrolling you can just utilize this API 2,500 times each day.

3) Clean Geocoding results: Cleaning of geocoded results is fundamenta to

a) eliminate the ill-advised areas since certain clients load data in the "area" segment of their Twitter profile that isn't an area

b) Removing the vague areas done since it is not known unequivocally which one of these (assuming any) is the client's genuine area.

c) Remove the misformatted passages (an expected issue when at the same time managing addresses from across the world)

E. K-means Clustering of Locations

1) Cluster the Latitudes and Longitudes: The scopes and longitudes acquired during the time spent geocoding are exposed to k methods bunching where we get the focuses of various bunches. These focuses (Latitudes and Longitudes) acquired will be the Latitudes and Longitudes of the client whose area is obscure.

2) Plotting Follower's Location and User's anticipated Area on Google Maps: To plot the scopes and longitudes on Google Maps, we need an API key of Google Maps. Presently the scopes and longitudes of the client and his supporters can be plotted on Google Maps. It includes the accompanying advances:

a) Add focuses with Latitude/Longitude Coordinates.

b) Plot the areas on Google Maps:

1. Supporters areas :

The Followers areas are plotted on Google Guides by acquiring the supporters' areas of the User and afterward through Geocoding instrument changing them over to Latitudes and Longitudes. The Latitudes and Longitudes got are plotted on Google Maps utilizing Google Maps API.

2. Client's Location :

The Latitude and Longitude secured as a middle utilizing K Means Clustering are taken as the major Latitude and Longitude of the client and plotted on the Google Maps.

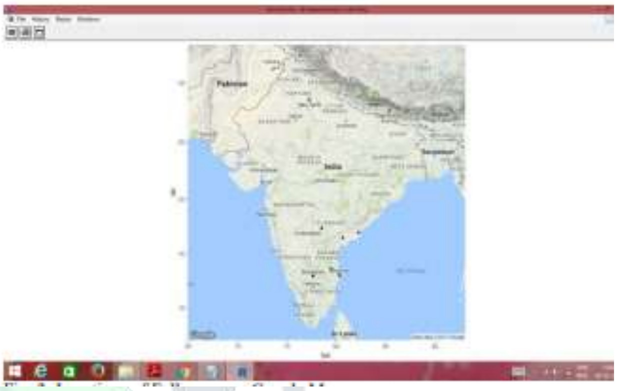

Fig 3(a): Locations of Followers on Google Maps

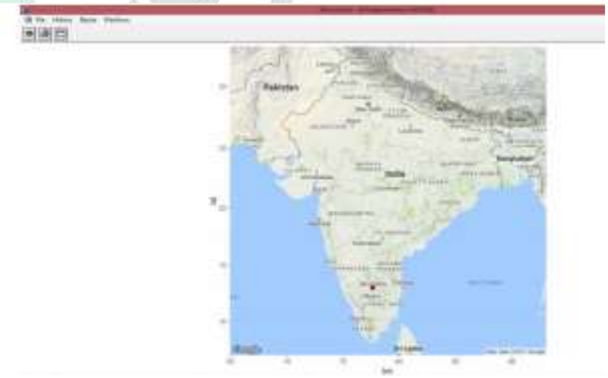

Fig 3(b): Predicted Location of user on Google Map

V. FUTURE SCOPE

Apart from only one algorithm of machine learning we can use different algorithms to preprocess and classification. Here most of the data collected are used to training and left out is used to testing, this implementation gives more accuracy.

VI. CONCLUSION

Firstly the process starts from creating the twitter developer application and the programmer has to collect data from the twitter account and has to be stored in one file. And by utilizing the content of the data one has to find the location of the tweet based what kind of tweet it is which has been done using classification. The got degrees and longitudes are introduced to a cycle called K-mean Clustering to get the central places of the social affairs. These focuses address the degrees and longitudes of the clients. The got degrees and longitudes are plotted on Google Maps. The two basic procedures that were utilized recorded as a printed copy were Network-based assessment and Content-based examination.

REFERENCES

[1] Han, Bo & Cook, Paul & Baldwin, Timothy. (2012). Geolocation Prediction in Social Media Data by Finding Location Indicative Words. 24th International Conference on Computational Linguistics - Proceedings of COLING 2012: Technical Papers. 1045-1062.

[2] Ren K., Zhang S., Lin H. (2012) Where Are You Settling Down: Geo-locating Twitter Users Based on Tweets and Social Networks. In: Hou Y., Nie JY., Sun L., Wang B., Zhang P. (eds) Information Retrieval Technology. AIRS 2012. Lecture Notes in Computer Science, vol 7675. Springer, Berlin, Heidelberg.

[3] Han, Bo & Cook, Paul & Baldwin, Timothy. (2014). Text-Based Twitter User Geolocation Prediction. The Journal of Artificial Intelligence Research (JAIR). 49. 10.1613/jair.4200.

[4] Li, Rui & Wang, Shengjie & Chen-Chuan Chang, Kevin. (2012). Multiple Location Profiling for Users and Relationships from Social Network and Content. Proceedings of the VLDB Endowment. 5. 10.14778/2350229.2350273.

[5] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. 2014. Home Location Identification of Twitter Users. ACM Trans. Intell. Syst. Technol. 5, 3, Article 47 (July 2014), 21 pages. DOI: http://dx.doi.org/10.1145/2528548