User Location Prediction in Social Network

Rohini R[1]
Student
Department of Information Science and Engineering
RV College of Engineering
Bengaluru, India

Prof. Smitha G R[2]
Assistant Professor
Department of Information Science and Engineering
RV College of Engineering
Bengaluru, India

Abstract- Location prediction of users from online social media brings enormous research in recent times. Automatic identification of locations related with or mentioned in records has been investigated for decades. As one of the famous online social network platforms, Twitter has attracted a massive number of users who send millions of tweets on each day basis. As Global inclusion of its users and continuous tweets, location prediction on Twitter has extended noteworthy attention in these days. In proposed framework, a standard user location prediction in online platform using tweets is studied. In precise, tweet location is predicted from tweet contents. By outlining tweet content and contexts, it fundamentally that how the problems depend upon the ones text inputs. Ensemble model is introduced to combine all multiple machine learning technique like Support Vector Machine, Decision Tree, Random forest and Logistic Regression to predict user location model.

Keywords- Online Platform, Twitter, Machine Learning

1. INTRODUCTION

Internet social media services, such as social networking and microblogging which are offered by social platforms like Twitter and Facebook, location-based ones like Foursqure [5] and Gowalla, photo sharing sites like Flickr and Pinterest, as well as other domain-specific platforms such as LinkedIn have seen phenomenal growth in their user bases. On these social platforms, users may additionally establish online friendship with others sharing similar interests. Users may proportion with online buddies, each day lives in varieties of texts, pics, videos, or check-ins.

Among all online social network, Twitter is portrayed by its special method for following companions and sending posts. From one viewpoint, Twitter companionships are not really shared. For instance, users may "follow" superstars without expecting them to follow back. Tweets are not a specifically language, in which user may post with emotion pictures. Abridged type of content, incorrect spellings, and additional characters of enthusiastic words makes tweet writings uproarious. The methods applied for typical archives are not appropriate for analyzing tweets. The character confinements of tweets around 140 characters may make the tweet uncomfortable to understand, if the tweet setting is not contemplated.

Users, online friendships and tweets make Twitter a virtual online world. This virtual world meets with this present reality, where locations acting as intermediate connections. Twitter users have long term private locations. Their home areas cause them to see, get intrigued and tweet news or occasions around their day by day movement locales. With expanding notoriety of GPS-empowered gadgets for example-cell phones and tablets, users may casually attach real-time locations when sending out tweets [7]. Users may likewise make reference to Mentioned locations in their tweets, e.g., urban areas they recently lived in, or eateries they need to attempt.

The issue of location prediction related named as geolocation prediction is inspected for Wikipedia [25][26][27] and web page documents [28]. Entity recognition from these formal documents has been looked into for quite a long time. Different types of content and context handling on these documents are also studied extensively. However, the location prediction problem from twitter depends highly on tweet content

1.1 Tweet Content

A tweet is a bit of user produced content with its length up to 140 characters. It might depict anything a user needs to post, e.g.-her state of mind or occasions occurring around her. Besides unique posts, a user may likewise retweet others' tweets she peruses. Tweet and retweets from a user will be pushed to her followers. Twitter interface for them to peruse. When composing tweet contents, a user may incorporate hashtags, which are words or unspaced expressions beginning with "#". At last, one can likewise make reference to another user's name by a first "@" in tweet content. A mentioned user will be notified and may begin a discussion with the referencing user through resulting makes reference to.

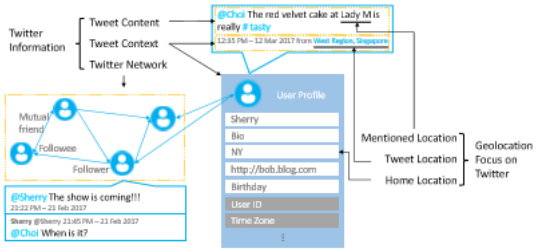

Fig. 1 - An illustration of tweet content, tweet context, and Twitter network, and the three types of locations: home location, tweet location, and mentioned location in Twitter

To predict the user location in twitter there are three types of Twitter related locations, namely home location, tweet location, and mentioned location. For each different type of location, we give its definition and show how it is been represent. We likewise briefly examine how to set up ground truth for each assignment.

2.1 Home Location Prediction

Home locations allude to Twitter user long-term residential address locations. The prediction of home locations empowers different applications, e.g., neighborhood content suggestion, area-based ad, public opinion polling estimation and public health monitoring. Home Location might be spoken to at various degrees of granularity.

2.2 Tweet Location Prediction

Tweet location implies where a tweet is posted. By interfacing tweet areas, we may draw a progressively complete image of a user's portability. Not the same as home locations, which are gathered from both user profiles and geo-tags, tweet locations are commonly founded on geo-tags of tweets. Due to the first perspectives on tweet locations, purpose of-interests (POIs) or coordinates are broadly adopted as representations of tweet locations, rather than managerial districts or grids.

2.3 Mentioned Location Prediction

When composing tweets, users may make reference to the names of certain areas in tweet contents. Mentioned location prediction may encourage better comprehension of tweet content and benefit applications like area suggestion and fiasco and illness the executives. Mention location divide into two sub class:

- **Mentioned location recognition**- separate content sections in a tweet that allude to location names.[7]
- **Mentioned location disambiguation**- distinguish what areas those sections allude to by settling them to passages in a location database.[7]

Fig. 2 – User Mentioned location in the tweet content

2. LITERATURE SURVEY

In reference [1] This paper centers around exploring geolocation forecast approaches dependent on content examination in social media life information. The audit result shows that geolocation forecast approaches can be classified into two classifications called Content-based Geolocation Prediction and User-profiling-based Geolocation Prediction. This audit further presumes that Content-based Geolocation Prediction is appropriate for tending to geotagged information constraints in Location-explicit Analysis in light of the fact that the area forecast results are explicit to put level.

In reference [2] In this paper it analyzes the study of the capacity of system-based home area estimation with cycle while utilizing the informal organization dependent on following connections on Twitter. The outcomes show that the capacity that chooses the most regular area among the companions' area has the best precision. Our investigation additionally shows that the 88% of users, who are in the interpersonal organization dependent on following connections, has in any event one right home location inside one-bounce.

In reference [3] In this paper, a model of an Android cell phone application named "T-support" is introduced in this examination. This application empowers bolstered users who needs every day backing to share their area organizes by means of Twitter. Supporting users would then be able to check the area directions of the bolstered clients when required.

In reference [4] In this paper, the ways of finding citizen problems with their locations by using tweet data is discussed. Tweets in Turkish language from the Aegean Region of Turkey were used for the study. It is aimed to form a smart system, which detects problems of citizens and extracts the problems' exact locations from tweet texts. Firstly, the collected data was analyzed to get information of any city event, citizen's complaint or requests about a problem. After the possibility of detecting tweets, which have any city problem and was ensured to two datasets were created

In reference [5] In this paper, figure a pattern likelihood gauge of the dissemination of words utilized by a user. This conveyance is shaped by utilizing the way that terms utilized in the tweets of a specific conversation might be identified with the area data of the client starting the conversation. Additionally, gauge the top K likely urban cities for a given client and measure the exactness. Find the pattern estimation yields a precision higher that the 10% exactness of the present cutting-edge estimation.

3. METHODOLOGY

In this section, we explain methodologies of the project. In Fig-3 shows the overall architecture of the system. The aim of proposed framework is to predict the user location from tweet content, considering users-Home location, Tweet location and Mentioned location. To deal with this we utilize ensemble learner in machine learning technique which take helps of several base model and combine their output to produce an optimized model.

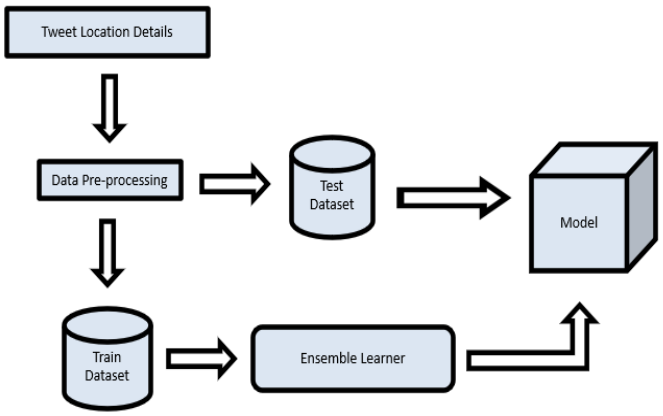

Fig. 3 – System Architecture

Twitter data is been downloaded from live stream to extract the location; we will utilize a powerful python library called tweepy to access tweets from the web in real-time. Live stream twitter helps to generate twitter credentials like consumer_key, consumer_secret, access_token, acess_token_secret to extract the user API from twitter account. Create the Listener class that will acquire from the Stream Listener object in tweepy. Later make the audience class that will acquire from the Stream Listener object in tweepy. Then make a wrapper and then define methods that will be activated depending what the listening is hearing. We will build the on_data and on_error method inside the StdOutListener class.

The on_data method is actuated at whatever point a tweet has been heard. Its input is the variable status, which is the genuine tweet it heard in addition to the metadata. Here, data can be viewed as an object with various parameters. The method on_error fills in as an error handler for our listener. Now and then, error 420 are being sent in our listener due to twitter's rate limit approach. At whatever point this sort of blunder shows up, it will provoke our listener to disconnect.

Regularly, it's a decent practice that you store your twitter credentials or anything that is private in a different record. Then add filters in the way we stream using the stream.filter() method.Stream.filter() is a track parameter is an array of keywords, such as Mumbai, Chennai, Karnataka and Kerala that will be listened in a listener class.

Data pre-processing should be done before converting the json file into .csv file. The data should be cleaned as twitter have huge data it contains unwanted data or noisy data, any special characters like symbols @, #, $, &, * should be removed, should capitalized all words to find the geo-location. If the user has not mentioned the home location while tweet it should be removed. If the Tweet Location is null then mention the Home Location of the user. The keywords should be assigned with the integers value as machine learning approach support only integer value i.e. Lvalue such as Nil-0, Chennai-1, Kerala-2, Mumbai-3 and Karnataka-4 for mentioned location. Later save tweets into dataset.

The language used for implementing the work is python, where it includes libraries like NumPy, scikit learn, pandas, Tweepy, matpolitlib, seaborn, geography. Geography is used to find the user geo-location in tweet text.

	tweet_id	Name	screen_name	tweet_text	HL	TL	ML	Lvalue
0	1072875969055121408	NamoFans4India	NamoFans4India	RT Birthday greetings to Karnataka s hardworki...	India	India	Karnataka	4
1	2835658765	Yash	YashDarira	RT Birthday greetings to Karnataka s hardworki...	New Delhi	New Delhi	Karnataka	4
2	956577567087218690	Ansh	Anshuman230	RT Birthday greetings to Karnataka s hardworki...	New Delhi India	New Delhi India	Karnataka	4
3	1139410664697368576	Madhu Bala	HrithikianMadhu	RT Haaan Hello Duggu App boliye mujhe sunai de...	India	India	Nil	0
4	491986148	LoksattaLive	LoksattaLive	Maharashtra Vidhan Bhavan	Mumbai	Mumbai	Nil	0
5	549543111	Azeem	azeemjamshed	RT Indian Muslims are being killed in the stre...	Bengaluru New Delhi UP	Bengaluru New Delhi UP	Nil	0
6	222835287	DarkKnight	iamshinerk	RT an elephant called padmanabhan who served t...	India	India	Kerala	2
8	2794801401	SumitOfficial	3sumitofficial	Mumbai Maharashtra	Mumbai India	Mumbai India	Mumbai	3
9	944616889480757249	Satheesh kumar	satheesh 1997	RT Today Darbar50Days Huge blockbuster of the ...	India	India	Nil	0
10	2357685042	People Truth	Bindas Bol	RT Birthday greetings and warm wishes to Karna...	India	India	Karnataka	4
11	3243753291	DIL SE SRKian	SRKian satish	RT Before coming to Mumbai Dil hai mera deewan...	India	India	Mumbai	3

Fig. 4 – Extract Live Location in Twitter

The above figure is the dataset of live stream data of twitter parameters like- the user tweet ID, name, screen_name, tweet_text, Home Location, Tweet Location, Mentioned Location and Lvalue.

Mentioned Location of Lvalues in percentage is been calculated in pie chart where Karnataka-4(35.59%), Mumbai-3(11.86%), Kerala-2(13.56%), Chennai-1(32.20%) and Nil-0(6.78%).

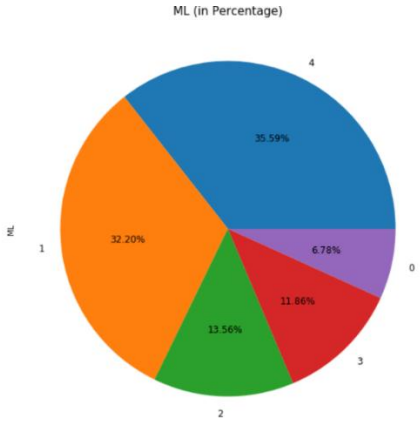

Fig. 5 – Pie Chart of Mentioned Location

Ensemble method help to combine multiple classifier model to form hybrid predict model for user location taking same input as machine learning algorithm and give effective accuracy by clubbing all classification algorithm.

Ensemble learner are combination of multiple classifiers, different classification algorithms used are:

- Logistic Regression
- Support Vector Machine
- Decision Tree
- Random Forest

By combining all classifiers in one model by taking same input data how the accuracy and performance of model is effective is shown by ensemble method.

ENSEMBLE METHOD

In machine learning, ensemble method uses multiple algorithms to obtain optimized predictive model that could be obtained from any individual machine learning algorithm. Ensemble method work with unstable classifier. A model selection should be done first, then the twitter dataset is read which is same input data taken for other classification algorithms. Separate data frame pixels and labels as df_x and df_y respectively. Split the dataset, to measure the accuracy train the data.

Decision Tree is selected as random forest is ensemble of Decision Tree to measure the accuracy. Boosting classifier can well boost the accuracy of the model by the great factor and Bagging helps to reduce the variance and overfitting of the model. Estimator Voting classifier is done by combining three algorithms.

In my ensemble model, implementation is done through voting classifier. Voting classifier is created by three different base classifiers, which are Logistic Regression, Support Vector Machine and Decision Tree. The relevant libraries are import in voting ensemble classifier and make the voting classifier to take specific feature as input and give effective output when compared to single base model which increase the performance of the model, this is done by kflod model cross validation where pre-trained model is not used. The new parameter is trained after receiving from training set, test with model which return k different values based on the number of splits done in dataset. Analyze the values generated for each dataset and consider the average as accuracy score of the model. Kflod cross validation overcomes some of data which was untouched while training and small chunked of data which may overfit model.

4. RESULTS

Voting classifier use multiple models like Logistic Regression, Support Vector Machine and Decision Tree and give the accuracy of 91.66% to the model. Random Forest ensemble with Decision Tree gives accuracy of 81.66%. Boosting and bagging classifier gives 100% in Decision tree for both training and testing data.

Create ensemble model by kflod cross validating as it returns k different values for accuracy score, based on the kth test data set. Fig.6 and Fig.7 shows the average or mean result to analyze the model

```
[0.83333333 1.         0.83333333 0.66666667 1.         0.83333333
 1.         1.         0.8        0.8        ]
0.8766666666666667
```

Fig.6 Mean Accuracy of ensemble Model

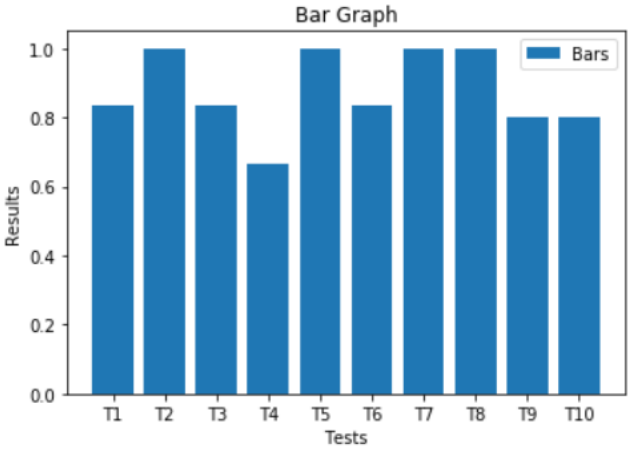

Fig.7 Results of different accuracy score

Ensemble classifier give better prediction than single base classifier so this model predicts the user location in tweet content. Based on the same user the count increases, in machine learning the fit transform function is used so the location name take automatically in background from dataset as integer and predict the location of user in integer as seen in fig.8. The Lcoder gives value for the locations. As the location is taken automatically it predict the location of user in which

home and tweet location, they have tweet. In my work for mentioned location have taken 4places, if the location is not from those 4 places than it is shown as nil.

	Home Location	Tweet Location	Mention Location
0	6	37	4
1	23	24	3
2	12	11	4
3	8	7	3
4	11	10	4
5	15	16	2
6	0	0	3
7	25	27	1
8	20	21	3
9	9	8	3
10	21	22	2
11	13	13	3
12	31	35	3
13	23	24	3
14	20	21	4
15	12	11	4
16	12	11	3
17	22	23	3
18	30	34	4
19	19	20	4
20	28	30	4

Fig.8 – Predicted Result

5. CONCLUSION

Three geolocation problems on Twitter is summarized as home location, tweet location, and mentioned location. When twitter data is considered, geolocation prediction becomes a challenging problem. It is hard to understand and analyze the tweet text nature and number of tweet characters limitation. In this work, user geolocation is been predicted by tweet content using machine learning techniques. I have implemented my project using ensemble classifier to show all combined base model give better performance and show prediction result which is suitable for location prediction problem and tweet text analysis.

ACKNOWLEDGMENT

This study and research are carried out at R.V. College of engineering, Bangalore, Department of Information Science and Engineering, under the support and guidance of prof. Smitha G R. I also show gratitude to our Head of the Department Dr. Sagar B M and the principal of the institution, Dr. K N Subramanya for giving us all the required facility and suitable environment to successfully complete this research.

REFERENCES

[1] Muhammad Nur Yasir Utomo, Teguh Bharata Adji, Igi Ardiyanto, "Geolocation Prediction in Social Media Data Using Text Analysis", IEEE Conference on Information and Communications Technology Vol.4, No.4, May 2018.

[2] Shiori Hironaka, Mitsuo Yoshida, Kyoji Umemura, "Analysis of Home Location Estimation with Iteration on Twitter Following Relationship", International Conference on Computer Science and Engineering, June, 2016.

[3] Yuuto Ohtsuka, Hiroshi Ishii, Keisuke Utsu,"A Smartphone Application for Location Recording and Rescue Request Using Twitter", IEEE Transactions On Knowledge And Data Engineering July 2017.

[4] Gizem Abalı, Enis Karaarslan, Ali Hürriyetoglu, "Detecting Citizen Problems and Their Locations Using Twitter Data",September 2018.

[5] Swarup Chandra, Latifur Khan, Fahad Bin Muhaya," Estimating Twitter User Location Using Social Interactions – A Content Based Approach", IEEE International Conference on Social Computing, August 2011.

[6] Linyuan Xia, Qiumei Huang and Dongjin Wu, "Decision Tree-Based Contextual Location Prediction from Mobile Device Logs", in School of Geography and Planning, Sun Yat-Sen University, Guangzhou, China, vol 618, April 2018.

[7] Xin Zheng, Jialong Han, and Aixin Sun," A Survey of Location Prediction on Twitter", in IEEE Transaction, vol 16, 2018.

[8] O. V. Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt, "Spatially aware term selection for geotagging," IEEE Trans. Knowl. Data Eng., vol. 26, no. 1, pp. 221–234, 2014

[9] Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in Proc. 19th ACM Conf. on Information and Knowledge Management, pp. 759–768,2012.

[10] Yuuto Ohtsuka, Hiroshi Ishii, Keisuke Utsu, "A Smartphone Application for Location Recording and Rescue Request Using Twitter", International Journal of Scientific & Engineering Research, Volume 4, Issue 5, April-2017.

[11] Y. Qian, J. Tang, Z. Yang, B. Huang, W. Wei and K. M. Carley, "A probabilistic framework for location inference from social media," arXiv:1702.07281, 2017.

[12] Z. Ji, A. Sun, G. Cong and J. Han, "Joint recognition and linking of fine-grained locations from tweets," in Proc. 25th Int. Conf. on World Wide Web, 2016, pp. 1271–1281.

[13] S. Zhao, I. King and M. R. Lyu, "A survey of point-of-interest recommendation in location-based social networks," vol, 2016.

[14] Y. Liu, T. Pham, G. Cong and Q. Yuan, "An experimental evaluation of point-of-interest recommendation in location-based social networks," PVLDB, vol. 10, no. 10, pp. 1010–1021, 2017.

[15] L. Kong, Z. Liu, and Y. Huang, "SPOT: Locating Social Media Users Based on Social Network Context," Proceedings of the VLDB Endowment, vol. 7, no. 13, pp.

1681–1684, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.

[17] Pengfei Li, Hua Lu, "Location Inference for Non-geotagged Tweets in User Timelines", IEEE Transactions On Knowledge And Data Engineering Vol.4, No.4, May 2019.

[18] M. Hulden, M. Silfverberg and J. Francom, "Kernel density estimation for text-based geolocation," in Proc. 29th AAAI Conf. on Artificial Intelligence, 2016, pp. 145–150.

[19] M. Dredze, M. Osborne and P. Kambadur, "Geolocation for twitter: Timing matters," in Proc. 14th Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 1064–1069

[20] W. Chong and E. Lim, "Exploiting contextual information for fine-grained tweet geolocation," in Proc. 11th Int. Conf. on Web and Social Media, 2017, pp. 488–49

[21] W. Chong and E. Lim, "Tweet geolocation: Leveraging location, user and peer signals," in Proc.26thACMConf.onInformationand Knowledge Management, 2017, pp. 1279–1288

[22] Z. Liu and Y. Huang, "Where are you tweeting? A context and user movement-based approach," in Proc. 25th ACM Int. Conf. on Information and Knowledge Management, 2016, pp. 1949–1952

[23] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, and D. Ruths, "Geolocation prediction in twitter using social networks: A critical analysis and review of current practice," in Proc. 9th Int. Conf. on Web and Social Media, pp. 188–197,2015.

[24] P. Anantharam, P. Barnaghi, K. Thirunarayan, and A. Sheth, "Extracting city traffic events from social streams," ACM Transactions on Intelligent Systems and Technology, vol. 6(4), pp. 43:1–27, July 2015.

[25] B. P. Wing and J. Baldridge, "Simple supervised document geolocation with geodesic grids," in Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, 2011, pp. 955–964.

[26] B. Wing and J. Baldridge, "Hierarchical discriminative classification for text-based geolocation," in Proc. Conf. on Empirical Methods in Natural Language Processing,pp. 336–348,2014.

[27] E. Amitay, N. Har'El, R. Sivan, and A. Soffer, "Web-a-where: geotagging web content," in Proc. 27th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 273– 280,2004.