Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)
IEEE Xplore Part Number: CFP19J32-ART; ISBN: 978-1-5386-9439-8

Location prediction on Twitter using machine learning Techniques

1st Indira K
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
indira.it@sathyabama.ac.in

2nd Brumancia E
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
brumancia.it@sathyabama.ac.in

3rd Siva Kumar P
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
kumar.siva805@gmail.com

4th Shyamala Pavan Teja Reddy
School of Computing
Sathyabama Institute of Science and Technology
Chennai, India
Pavantejareddy1997@gmail.com

***Abstract*—Location prediction of users from online social media brings considerable research these days. Automatic recognition of location related with or referenced in records has been investigated for decades. As a standout amongst the online social network organization, Twitter has pulled in an extensive number of users who send a millions of tweets on regular schedule. Because of the worldwide inclusion of its users and continuous tweets, location prediction on Twitter has increased noteworthy consideration in these days. Tweets, the short and noisy and rich natured texts bring many challenges in research area for researchers. In proposed framework, a general picture of location prediction using tweets is studied. In particular, tweet location is predicted from tweet contents. By outlining tweet content and contexts, it is fundamentally featured that how the issues rely upon these text inputs. In this work, we predict the location of user from the tweet text exploiting machine learning techniques namely naïve bayes, Support Vector Machine and Decision Tree.**

Keywords—Social media, Twitter, Tweets, location prediction, Naive Bayes, Support Vector Machine, Decision Tree, Machine Learning

I. INTRODUCTION

Users may post explicitly their location on the tweet text they post, whereas in certain cases the location may be available implicitly by including certain relevant criteria. Tweets are not a strongly typed language, in which users may post casual with emotion images. Abbreviated form of text, misspellings, and extra characters of emotional words makes tweet texts noisy. The techniques applied for normal documents are not suited for analysing tweets. The character limitations of tweets about 140 characters may make the tweet uneasy to understand, if the tweet context is not studied.

The issue of location prediction related named as geolocation precition is examined for Wikipedia and web page documents. Entity recognition from these formal documents has been researched for years. 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. Users living in specific regions, locations may examine neighborhood tourist spots, landmarks and buildings and related events.

Home Location: User's residential address given by user or location given by user on account creation is considered as home location. Home location prediction can be used in various application namely recommendation systems, location based advertisements, health monitoring, and polling etc. Home location can be specified as administrative location, geographical location or co-ordinates.

Tweet Location: Tweet location refers to the region from where the tweet is posted by user. By construing tweet location, one can get tweet person's mobility. Usually home location collected from user profile, whereas tweet location can be arrived from user's geo tag. Because of the first perspectives on tweet location, POIs are comprehensively received as representation of tweet regions.

Mentioned Location: When composing tweets, user may make reference to the names of a few locations in tweet texts. Referenced location prediction may encourage better understanding of tweet content, and advantage applications like recommendation systems, location based advertisements, health monitoring, and polling etc. In this study, we include two sub-modules of mentioned location:
First one is recognizing the mentioned location in tweet text, which can be achieved by extracting text content from a tweet that refers to geography names. Second one is identifying the location from tweet text by solving them to entries in a geographical database.

II. RELATED WORK

Many existing techniques have been studied by the researchers on location prediction problem from tweet content and social media content, few of them are discussed below.

In [1], the author refers to the problem of finding location from social media content. The author from [1] and [2] motivated by Term frequency (TF) and inverse document frequency (IDF), they arrived Inverse City Frequency (ICF) and Inverse Location Frequency (ILF) respectively. They raked the features by using these frequency values and TF then by TF values. From this they arrived that local words spread in document in few places and have high ICF and ILF values.

Han et al [3] in their work, they approached model for identifying local words indicative or used in certain

locations only. They aimed to identify automatically by ranking the local words by their location, and they find their degree of association of location words associated to particular location or cities.

Li et al. [4] proposed multiple locations profiling (MLP) model to arrive user location accurately by finding the probability based on Bernoulli distribution. Their work represents that users home location can be predicted accurately using this model. The author used multinomial distribution to estimate probability of tweet versus the venue name from each location.

Mahmud et al. proposed classification model for predicting location, they improved the accuracy of prediction by first predicting regions and then city. They registered the movement of users using classifier models, if the user travels for a certain period, then they are registered to improve the accuracy of prediction. The authors considered the person is travelling when the location distance for two tweets is more than 100 miles.

Most of the techniques used in existing works are machine learning, whereas few works in deep learning also proposed. Miura et al. [6] on his work used neural network is implemented for twitter location prediction. The author classified tweet or user using neural networks and they integrated metadata with tweet texts and trained the model. Their model achieved around 41 percentage of accuracy on predictions.

III. Proposed Methodology

Live stream of twitter data is collected as dataset using authentication keys. The aim of proposed system is to predict the user location from twitter content considering user home location, tweet location and tweet content. To handle this we used three machine learning approaches to make prediction easier and finding the best model amongst them. Fig. 1, represents the overall architecture of the proposed system with methodology modules represented.

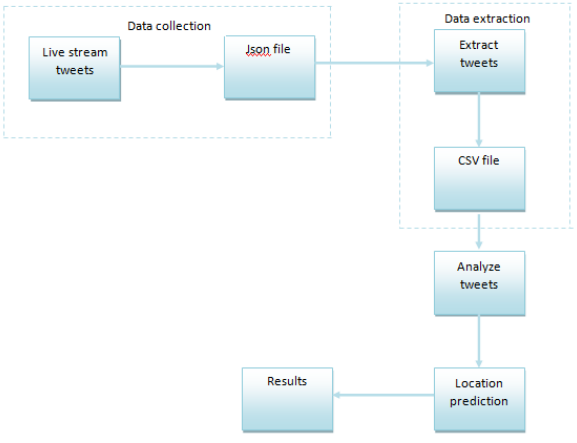

Fig. 1. System Architecture

Live tweet stream from twitter for keyword “apple” is collected and stored in 'twitter.json’ file. Live twitter data can be collected by registering a consumer_key, consumer_secret, access_token, access_token_secret for authentication and collecting live stream of tweets. We have collected more than 1000 tweets of particular keywords such as ‘Chennai, Mumbai, Kerala’.

The information extracted from live includes tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation, Lvalue.

Primary analysis was a basic processing of the text of the tweets. This was done by merging the collected tweets for a given user into a single “document” and analysing that.

A	B	C	D	E	F	G
tweet_id	Name	screen_name	tweet_text	Home Location	Tweet Location	Mention Loca Lva
7.89858E+17	Savishkar Live	SavishkarLive	RT Kerala Govt invites applications from SE	Bhopal India	Bhopal India	Kerala
1.03437E+18	cheeks	uniklln	ito pa	puso mo	puso mo	Nil
2895688958	A Masked Error	BumchikSeenu	RT Smt Vyjayanthimala age 86 who was the	Chennai	Chennai	Nil
169426623	Jai Hind	arbind1982	RT Railway 7 2 5	Lagos Nigeria	Lagos Nigeria	Nil
8.18519E+17	Vijith	vijithfilmlover	Smt Vyjayantimala Age 86	India	India	Nil
7.43735E+17	Johns	CricCrazyJohns	RT Melbourne or Mumbai MCG crowd abo	Kerala India	Kerala India	Nil
108272890	Mahesh Veeramal	MaheshMaddy	No words	Bengaluru India	Bengaluru India	Nil
3032998642	M La	ItzMiLu	RT rt Bumping into you made my day Live /	Bharat	Bharat	Nil
131520960	Ashish Chandorkar	AshishChandMT	Dear Sir This is MahishmatiThali in Pune pr	Pune	Pune	Nil
173786873	Jai Italy Jai Jai Italy	ravi enigma	some serious mental issues out there in K	Uttara Prachand	Uttara Prachand	Kerala
8.39501E+17	Austinne	Austianoo7	RT Telugu Sarkar gross gt Gang S3 Tamil Sar	Kerala India	Kerala India	Kerala
9.25043E+17	Arul Vignesh	ArulVignesh7	RT Adopted Son Of Kerala Suriya Fan Girl o	Chennai India	Chennai India	Kerala
298368845	Nelson Ji	Nelson Ji	RT FC Trade Updates Viswasam Chennai Ci	Chennai India	Chennai India	Chennai

Fig. 2. Extract Live Location Live Twitter

A. Data Collection and Extraction

Live tweet stream from twitter for keyword “apple” is collected and stored in 'twitter.json’ file. Live twitter data can be collected by registering a consumer_key, consumer_secret, access_token, access_token_secret for authentication and collecting live stream of tweets. We have collected more than 1000 tweets of particular keyword such as ‘Chennai, Mumbai and Kerala’. The information extracted from live includes tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation, Lvalue.

Data from 'twitter.json’ file is read and extracted tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation are extracted. Tweet text is compared with natural language tool kit package available in python to extract data from json file to csv is done here.

B. Data Preprocessing

Data pre-processing include the following steps,

1. Extra characters are removed from tweet text.

2. Capitalize all words to find for geo location

3. Remove the tweet if user home location not mentioned

4. Mention home location in tweet location, if user tweet location is null

5. Removes tweets if no location is mentioned in tweet text.

Final extract geodata from tweet text. Last step is to assign integer value to the locations, for example Chennai—1, Mumbai—2, Kerala—3. Lcoder is used to assign location as integer value.

The work is implemented using Python programming, with few libraries used are scikit learn, numpy, pandas, matplotlib, geography.

C. Naive Bayes Classification

Naive Bayes classifier is the most popular and simple classifier model used commonly. This model finds the posterior probability based on word distribution in the document. Naïve Bayes classifier work with Bag Of Words (BOW) feature extraction model, which do not consider the position of word inside the document. This model used Bayes Theorem for prediction of particular label from the given feature set. The dataset is split into trainset and test set. Upon test set, NB_model is applied to find the location prediction.

D. Support Vector Machine

Support vector machine is one of most common used supervised learning techniques, which is commonly used for both classification and regression problems. The algorithm works in such a way that each data is plotted as point in n-dimensional space with the feature values represents the values of each co-ordinate.

E. Decision Tree

Decision tree is the learning model, which utilizes classifications problem. Decision tree module works by splitting the dataset into minimum of two sets. Decision tree's internal nodes indicates a test on the features, branch depicts the result and leafs are decisions made after succeeding process on training.

Decision Tree works as follows

- Decision tree starts with all training instances linked with the root node
- It splits the dataset into train set and test set.
- It uses information to gain and chooses attributes to label the each node. Subsets made contain information with a similar feature attribute.
- Above process is repeated till in all subset until leafs get generated in tree.

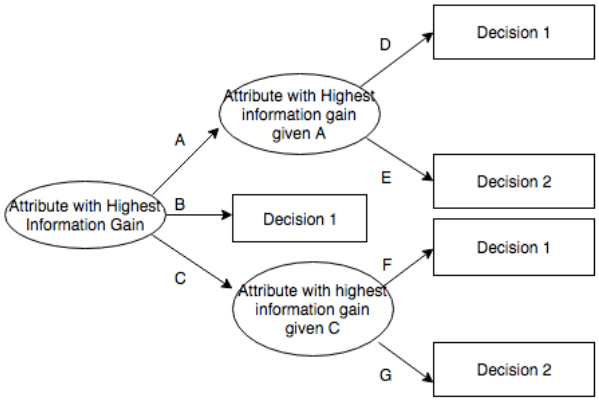

Fig. 3. Decision Tree Model

The tree is constructed in such a way that no root to leaf node path contains same attribute twice. This is done repeatedly to construct every sub tree on the training instances, which is classified down through the path in the tree. For every record in the dataset, class label prediction problem starts with root of the tree. The root attributes are checked for the given record and then it checks next record attributes. This process continues till the value next node to go. The sample decision tree applied is depicted in Fig.3.

Implementation done as represented in the use case diagram given the fig.4.

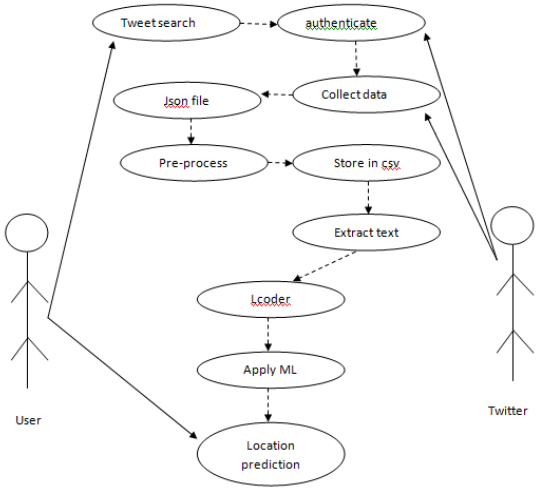

Fig. 4. Decision Tree Model

The extracted features from the tweet are mentioned below code snippet.

(user["features"]["id"],user["features"]["name"],user["features"]["screen_name"],user["features"]["tweetstext"],user["features"]["HomeLocation"],user["features"]["TweetLocation"]).

Instead of attaching the geo-tags to tweets, user may sometimes reveal the relevant location by specifying their name or landmarks in the tweets. During pre-processing the location names are important, thus we capitalize every words of tweet text to identify the geo-locations. Geo location can be processed in two ways, one is through recognition, label the text and if recognized then they are converted to location. Next is through disambiguation, which makes the entries as identified location.

TABLE I. PREDICTION RESULTS

ID	Decision Tree	SVM	Naive Bayes
1	1	1	1
2	2	2	1
3	0	0	0
4	2	2	2
5	1	1	1
6	0	0	0
7	0	0	0
8	2	2	2
9	1	1	1
10	1	1	2

IV. Results and Discussions

The pre-processed dataset are taken for machine learning process, we applied Naïve Bayes, SVM algorithm and Decision Tree on the dataset. The dataset is given 80% as training set and 20% as test set, we predicted the location and compared accuracy under following chart, Figure 4.

The following table shows the performance evaluation of three machine learning algorithm namely Naive Bayes, Support Vector machine (SVM) and Decision Tree. The evaluation parameters showed in the table are Accuracy of prediction. The table clearly depicts that decision tree outperforms the other algorithms in terms of efficiency in accuracy.

TABLE II. Accuracy Comparison

Algorithm	Accuracy
Naive Bayes	43.67
SVM	86.78
Decision Tree	99.96

The following table shows the error rates in prediction. There are four error types calculated are Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and R-squared.

TABLE III. Error Rate

Error Types	Naive Bayes	SVM	Decision Tree
MAE	1.06	0.13	0.02
MSE	2.31	0.13	0.02
RMSE	1.52	0.36	0.04
R-Squared	0.01	0.88	1.00

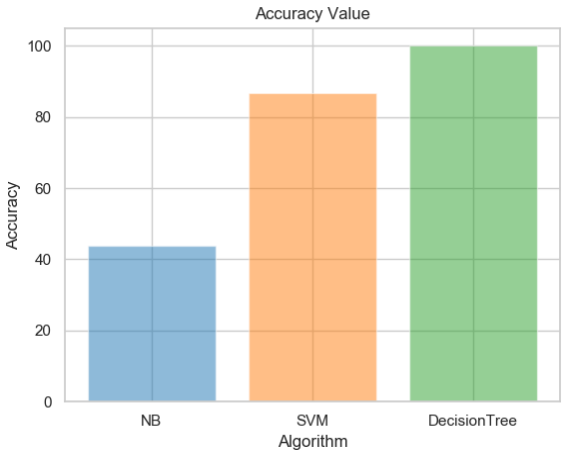

Fig. 5. Performance Comparison

The above figure, Fig. 5 shows the experimental results achieved using three machine learning algorithms. Naive bayes achieves around 40% of accuracy, SVM algorithm achieves around 85% of accuracy and Decision Tree achieves around 99% accuracy. Thus from this work, we can conclude that Decision Tree is the suitable algorithm for location prediction problem in tweet texts

V. Conclusion

Three locations are considered from twitter data, namely home location, mentioned location and tweet location. When the twitter data is considered, geolocation prediction becomes a challenging problem. The tweet text nature and number of characters limitation make it hard to understand and analyze. In this work, we have predicted the geolocation of user from their tweet text using machine learning algorithms. We have implemented three algorithms to show the better performed one, which is suitable for geolocation prediction problem. Our experiment analysis concluded that decision tree is suitable for tweet text analysis and location prediction problem.

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

[6] Miura, Yasuhide, Motoki Taniguchi, Tomoki Taniguchi and Tomoko Ohkuma. “A Simple Scalable Neural Networks based Model for Geolocation Prediction in Twitter.” NUT@COLING (2016).

[7] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. M¨ uhlh¨auser, “A multi-indicator approach for geolocalization of tweets,” in Proc. 7th Int. Conf. on Weblogs and Social Media, 2013.

[8] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, “Towards social user profiling: unified and discriminative influence model for inferring home locations,” in Proc. 18th ACM Int. Conf. on Knowledge Discovery and Data Mining, 2012, pp. 1023–1031.

[9] B. Han, P. Cook, and T. Baldwin, “A stacking-based approach to twitter user geolocation prediction,” in Proc. 51st Annual Meeting of the Association for Computational Linguistics System Demonstrations, 2013, pp. 7–12.

[10] D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza, “On the accuracy of hyper-local geotagging of social media content,” in Proc. 8th ACM Int. Conf. on Web Search and Data Mining, 2015, pp. 127–136.

[11] 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.

[12] J. Mahmud, J. Nichols, and C. Drews, “Where is this tweet from? inferring home locations of twitter users,” in Proc. 6th Int. Conf. on Weblogs and Social Media, 2012.