LOCATION PREDICTION ON TWITTER USING MACHINE LEARNING

[1]*Krishna Reddy Seelam,* [2]*V.Anusha Reddy,* [3]*N.Arpitha*
[123]*Assistant Proffessor*
Department of CSE
Sreedattha institute of engineering and science

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.

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 toentries in a geographical database.

The use of social media is being explored as a tool for disaster management by developers, researchers, government agencies and businesses. The disaster-affected area requires both, cautionary and disciplinary measures (Sushil 2017). Dai et al. (1994) first suggested the need for a computerized decision-making system during emergencies. Nowadays, information and communication technology (ICT) is being used widely during different phases of disaster for relief activities (Kabra and Ramesh 2015). Twitter plays a major role in informing people, acquiring their status information, and also gathering information on different rescue activities taking place during both, natural disasters (tsunamis/floods) and man-made disasters (terrorist attack/food contamination) (Al-Saggaf and Simmons 2015; Gaspar et al. 2016; Heverin and Zach 2012; Oh et al. 2013).

Social media platforms can be efficiently used for supply chain management by professionals, organizations, and retailers for their operations (Chae 2015; Mishra and Singh 2016; Papadopoulos et al. 2017). Social networks like Twitter and Facebook allow users to update information on social activities that they undertake (Mishra et al. 2016). Twitter provides the space where both official and common people can post their experiences and advice regarding disasters (Macias et al. 2009; Neubaum et al. 2014; Palen et al. 2010), which makes it a popular choice for disaster management. A lot of research work is going on to make this platform more suitable for disaster management. However, as suggested by Comfort et al. (2012), a more systematic study of social media is needed to improve public response. Turoff et al. (2013) is also of the same view, and have appealed to the research community to devise methods to improve citizen-engagement during emergencies. Quick and accurate responses from the leaders during disaster may boost their personal political standing (Ulku et al. 2015). Several agencies such as BMKG in Indonesia are actively engaged in providing updates and warnings to public through Twitter. Social media is also used by various agencies to coordinate rescue efforts and help victims.

Twitter is a micro blog where users send brief text messages, photographs and audio clips. Since users write small messages, they regularly send it and check for updates from others. Twitter updates include social events such as parties, cricket match, political campaigns, and disastrous events such as storms, heavy rainfall, earthquakes, traffic jams etc. A lot of work (Atefeh and Khreich 2015) has been done to detect events, both social as well as disastrous from Twitter messages. Most disastrous event detection systems are confined to detect whether a tweet is related to the disaster or not, based on textual content. The related tweets are further used to warn and inform people about precautionary measures (Sakaki et al. 2010, 2013). These tweets are also used to study the tweeting behavior of users during disasters. We view Twitter not only as an awareness platform, but a place where people can ask for help during disaster. The tweets asking for help need to be separated from other tweets related to the

disaster. These tweets then can be used to guide the rescue personnel.

To help victims in need, one needs to have his/her exact location in their tweet, which is another important issue in emergency situations. Distribution centers play a big role in helping victims. Burkart et al. (2016) proposes a multi-objective location routing-model to minimize the cost of opening a distribution center for relief routing. The real time location estimation plays a big role in logistics, stockpiles, and medical supply planning (Duhamel et al. 2016; Lei et al. 2015; Paul and Hariharan 2012; Ozdamar et al. 2004). The growing number of location-based Social Networks provide the spatiotemporal data that has substantial potential to increase situational awareness and enhance, both planning and investigation (Chae et al. 2014). The analysis by Cheng et al. (2010) shows that only 26% users mention their location at a city level or below, and the remaining are mostly a country name, or even words with not much meaning, such as Wonderland. According to Cheng et al. (2010), only 0.42% tweets have geo-tags, but Morstatter et al. (2013) found that about 3.17% tweets are geo-tagged. These analyses reveal that Twitter has limited applicability as a location-based sensing system.

The rise of mobile Internet users in the last couple of years has significantly increased the number of mobile twitter users. According to a report by IAMAI (2016), the mobile Internet users in India will be 371 million by the end of 2016. The same report also highlights the fact that in rural areas, 39% of users are using social media, whereas in urban areas, this percentage is much higher. Mobile Twitter users can switch on and off their geo-tagging, as and when preferred. The battery power of smart phones plays a significant role here, as the global positioning system (GPS) consumes significant amount of battery power. Users prefer switching off their GPS to save power. On the other hand, applications such as taxi hiring services and e-commerce sites such as flipkart.com require GPS to work properly. The analysis of mobile Twitter users thus shows some tweets with geo-tagging, and others without geo-tagging. During emergencies, people want to preserve the battery power of their phones; hence, tweets with geo-tags will be very few on such occasions.

India is a multilingual country, where English is used as the main language for communicating on social media websites. However, users of these sites also use their regional languages. Hence, event detection in the Indian context also needs to identify variations in the language used.

The major contribution of this paper is a tweet classification system to classify tweets into high and low priority. High Priority tweets are those, which ask for help, such as food, shelter, medicine etc. during a disaster. Two sample tweets of high priority. Tweet is in the English script, but the words used here are in the Hindi language. The translation of the tweet is, “Mr. @narendramodi, heavy floods in Chhapra Bihar, please arrange for administrative help, people here are very worried.” Low priority tweets convey information related to a disaster, such as “Rescue team has done a good job.” An example is where a user thanks Twitter for its help during a disaster. The other contribution of this paper is location prediction of high priority tweets, if geo-tagging information is missing in a tweet. To predict location, we use historical geo-tagged tweets of the specific users and build a Markov chain. The low priority tweets are analyzed to find the spread of the disaster. These may also be used to evaluate the performance of different agencies during a disaster.

II. SYSTEM ANALYSIS

EXISTING SYSTEM:

In the Existing system to the problem of finding location from social media content. The Social Networks from and 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. 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.

DISADVANTAGES OF EXISTING SYSTEM:

- The issue of location prediction related named as geolocation prediction is examined for Wikipedia and web page documents.
- Entity recognition from these formal documents has been researched for years.
- The location prediction problem from twitter depends highly on tweet content.
- **Algorithm**: Term Frequency (TF) and Inverse Document Frequency (IDF)

PROPOSED SYSTEM:

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. Live tweet stream from twitter for keyword "apple" is collected and stored in Tweettable. 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 Indian city hashtag names. You can also search tweets based on hashtags.

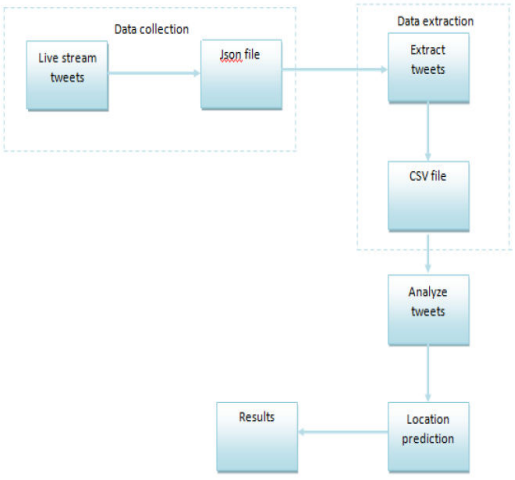
ADVANTAGES OF PROPOSED SYSTEM:

- The information extracted from live includes tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation.
- Tweet text is compared with natural language tool kit package available in python to extract data from Cursor to Pandas Dataframe.
- Python programming, with few libraries used are scikit learn, numpy, pandas and geography.

Algorithm: Naive Bayes, Support Vector Machine, Decision Tree

SYSTEM DESIGN

SYSTEM ARCHITECTURE:



III. IMPLEMENTATION

MODULES DESCRIPTION:

User:

The User can register the first. While registering he required a valid user email and mobile for further communications. Once the user register then admin can activate the customer. Once admin activated the customer then user can login into our system. User can search tweets based on hashtag. The first 100 tweets will get from twitter database and displayed to the user. At this time we are using geo code to identify the user location and tweet location. Most of the time user will not provide coordinates of his identity in the twitter account. So we are taking that as label class. This all tweets and geo code will stored in the database. Later we can apply the machine learning algorithms to test prediction result. The y_pred and y_test will displayed on the console. By help of sklearn.model_selection we can split the data into trainandtest. here we taken 80% of data for training and remiaing 20% for the testing.

. **Admin:**

Admin can login with his credentials. Once he login he can activate the users. The activated user only login in our applications. The admin can set the training and testing data for the project dynamically to the code. After user operated the algorithms on provded dataset. The admin can view the results of naivebayes, svm and Decision tree results on his screens.

Data Preprocess:

Extra characters are removed from tweet text. Capitalize all words to find for geo location. Here we are using geography python library to get the exact latitude and longitude points of the users. Remove the tweet if user home location not mentioned. Mention home location in tweet location, if user tweet location is null

Removes tweets if no location is mentioned in tweet text. Final extract geodata from tweet text. Last step is to assign float value to the locations by its latitude and longitude values.

Machine Learning:

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.

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.

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.

IV. 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.