Applied GIS

ISSN: 1832-5505
Vol-10 Issue-02 April 2022

PREDICTING YOUR LOCATION ON TWITTER Using ML Techniques

BEDUDHURI HIMAVANI, Assistant Professor, himavanicse.9f@gmail.com
DIGALA RAGHAVARAJU, Assistant Professor, raghava.digala@gmail.com
TALARI SIVALAKSHMI, Assistant Professor, shivalakshmidinesh@gmail.com
Department of CSE, Sri Venkateswara Institute of Technology,
N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

ABSTRACT

There is a lot of current study on the topic of location prediction using online social media users. Researchers have been looking on automatic location recognition for decades, particularly as it relates to or is mentioned in documents. Among the many online social network organisations, Twitter stands out, drawing in a large user base that routinely tweets millions of times. These days, location prediction on Twitter is getting a lot of attention because of its global user base and constant posts. Research in the field is fraught with difficulties because to tweets, which are brief, loud, and rich in nature. A high-level overview of tweet-based location prediction is investigated in the suggested framework. Specifically, the contents of tweets may be used to forecast their locations. The problems depend on these text inputs, and the substance and context of tweets are highlighted by underlining them. The present study utilises machine learning methods, namely naïve bayes, Support Vector Machine, and Decision Tree, to forecast the user's position based on the content of their tweets.

INTRODUCTION

While users have the option to explicitly share their location in tweets, it is also possible to make the location public implicitly by inserting certain criteria. Tweets allow users to submit informal visuals with emotions rather than tightly written text. Tweets take on an emotive tone when they include abbreviations, misspellings, and additional characters. text messages are really loud. Analysis of tweets requires a different set of methods than those used for traditional texts. Without studying the tweet's context, the character constraint of 140 characters could make the tweet difficult to interpret.
Articles on Wikipedia and other websites investigate the problem of location prediction, sometimes known as geolocation prediction. For a long time, researchers have been looking at entity recognition from these official papers. Extensive research is also conducted on various forms of content and context management in these papers. The content of tweets, however, plays a significant role in the Twitter location prediction challenge. Users residing in certain areas may research nearby tourism attractions, buildings, monuments, and events.
Home Location: The home location is the address or location that the user provided when creating their account. Recommendation systems, location-based ads, health monitoring, polling, and other applications may all benefit from home location prediction. Administrative, geographical, or coordinates may all be used to describe a person's home. The geographic area from which a user posts a tweet is known as the tweet's location. You can figure out how mobile a tweeter is by deducing their location. Typically, a user's home location is retrieved from their profile, whereas the location of a tweet may be obtained via their geo tag. Points of interest (POIs) have been widely accepted as a result of the first thoughts on tweet placement.

geographical depiction of tweets. Users may mention specific locations in their tweets by using the names of such places. Better comprehension of tweet content and its potential benefits for applications such as recommendation systems, location-based ads, health monitoring, polling, etc., may be encouraged by referenced location prediction. Our research incorporates two location-related sub-modules: The first is identifying the place in the tweet itself; this is done by sifting through tweets for mentions of geographical terms and extracting their text content. The second one is finding the location in tweets by matching them to entries in a database Businesses, government organisations, academics, and developers are investigating the potential of social media as a disaster management tool. Precautionary and punitive actions are needed in the catastrophe zone (Sushil 2017). It was first proposed by Dai et al. (1994) that an automated system for making decisions in times of crisis be implemented. Various stages of disaster relief operations increasingly make use of information and communication technology (ICT) these days (Kabra and Ramesh 2015). Natural disasters (tsunamis, floods) and man-made disasters (terrorist attacks, food contamination) both rely on Twitter to disseminate information and track people's status updates (Al-Saggaf and Simmons 2015; Gaspar et al. 2016; Heverin and Zach 2012; Oh et al. 2013). According to Chae (2015), Mishra and Singh (2016), and Papadopoulos et al. (2017), professionals, organisations, and merchants may make effective use of social media platforms for supply chain management. Users are able to keep others updated on their social activities using social networks such as Facebook and Twitter.

(Mishra et al. 2016) recommend. According to Macias et al. (2009), Neubaum et al. (2014), and Palen et al. (2010), Twitter is a popular option for disaster management since it offers a venue where both official and ordinary people may submit their experiences and advise about catastrophes. A great deal of effort is being put into improving this platform so that it can better handle catastrophe management. To enhance public reaction, however, more rigorous research into social media is required (Comfort et al., 2012). In a similar vein, Turoff et al. (2013) has urged academics to find ways to increase public participation in crisis response. According to Ulku et al. (2015), leaders' personal political standing may be enhanced if they respond quickly and accurately during disasters. In Indonesia, for example, BMKG is one of the authorities that uses Twitter to keep the public informed and provide warnings. In addition, several organisations use social media to assist victims and organise rescue operations. Twitter allows users to share short messages, photos, and audio snippets in the form of microblogs. Users often send and read short messages because they are interested in others' updates. Events ranging from social gatherings like parties and cricket matches to political campaigns to natural disasters like hurricanes, severe rains, earthquakes, and traffic jams are all part of Twitter's update feed. Identifying social and catastrophic events from Twitter tweets has battery power. When not in use, users may

of locations.

been the subject of many research (Atefeh and Khreich 2015). The majority of systems designed to identify catastrophic events only have the ability to determine, from the content of a tweet, if the tweet is relevant to the catastrophe. Additionally, the associated tweets serve to alert and educate others on safety precautions (Sakaki et al. 2010, 2013). Users' tweeting behaviour during catastrophes may also be studied using these tweets. In addition to raising awareness, we see Twitter as a venue where individuals may request assistance during disaster. Disentangling the tweets pleading for aid from those discussing the tragedy is essential. The rescue workers may then be directed by these tweets. The necessity to provide one's precise location in a tweet in order to assist victims in distress is another critical consideration in times of crisis. The function of distribution centres in aiding victims is significant. To reduce the overhead associated with relief routing distribution centre opening costs, Burkart et al. (2016) suggests a multi-objective location routing-model. Several studies (Duhamel et al., 2016; Lei et al., 2015; Paul and Hariharan, 2012; Ozdamar et al., 2004) have shown the importance of real-time position estimate in logistics, stockpiling, and medical supply planning. The proliferation of location-based social networks is a boon to situational awareness, planning, and research thanks to the spatiotemporal data it provides (Chae et al. 2014). According to research by Cheng et al. (2010), only 26% of users provide their location as a city or smaller. The rest either use a nation name or use meaningless phrases like Wonderland. Morstatter et al. (2013) discovered that around 3.17 percent of tweets include geotags, but Cheng et al. (2010) showed that just 0.42% of tweets have geotags. Twitter is not very useful as a system for location-based sensing, according to their evaluations. With more and more people using their mobile devices to access the Internet, the number of people using Twitter on the go has skyrocketed in recent years. There will be 371 million mobile Internet users in India by the end of 2016, according to a study from IAMAI (2016). Additionally, the survey notes that while the ratio of users in urban regions is much larger, 39% of users in rural areas are also using social media. Twitter users on mobile devices have the option to toggle geo-tagging on and off whenever they choose. Because smartphones' global positioning systems (GPS) use so much power, this is an important consideration for their battery life.

conserve electricity by turning off their GPS.

However, GPS is essential for business apps like rideshares and online marketplaces like flipkart.com. Thus, some tweets with geo-tagging and others without are shown in the study of mobile Twitter users. There will be relatively few tweets using geotags during crises since people are trying to save battery life on their phones. Even though Hindi is spoken in many parts of India, English remains the de facto language of choice for online conversations in India. On the other hand, local languages are also used by users of these platforms. Therefore, event detection in India must also account for linguistic diversity. This paper's main contribution is a mechanism for categorising tweets as either high or low importance. During a catastrophe, tweets asking for food, shelter, medication, and other necessities are considered high priority. Presented below are two outstanding examples of tweets. The terms used here are from the Hindi language, while Tweet is written in English. The tweet reads as follows: "Mr. @narendramodi, people here are very worried about the heavy floods in Chhapra Bihar. Please arrange for administrative help." Tweets of a lower priority provide disaster-related information, such "Rescue team has done a good job." A situation when a user expresses gratitude to Twitter for its assistance follows. If a tweet does not include geo-tagging information, the paper also contributes by predicting the location of high-priority tweets. In order to forecast where a user is, we construct a Markov chain using their geotagged tweets from the past. In order to determine how far the calamity has spread, the low-priority tweets are examined. In the event of a crisis, they might potentially be used to assess how well various authorities handled the situation.

I. 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

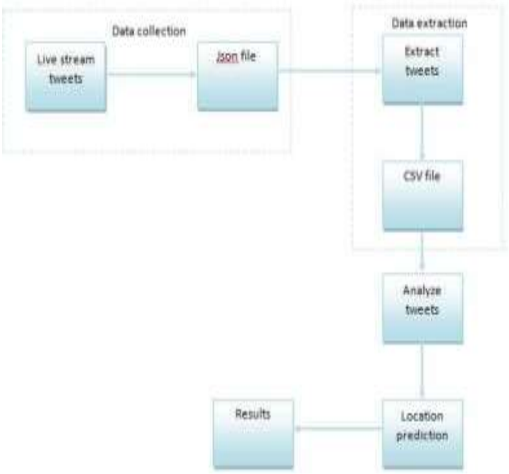
includes tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation.

- 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:



II. 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 naïve bayes, 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.

III. CONCLUSION

Using data from Twitter, three places are taken into account: the user's residence, the location of any mentions, and the location of the tweet itself. Taking into account the data from Twitter makes geolocation prediction a difficult challenge. Understanding and analysing tweets is challenging due to their textual nature and character constraint. Here, we take a user's tweets and utilise machine learning techniques to guess their location. To demonstrate the best solution for the geolocation prediction issue, we have built three different ones. Decision trees worked well for our location prediction and twitter text analysis experiments.

REFERENCES

Han, Bo, Cook, Paul, and Baldwin were the authors of a work that was published in 2012. Making Geolocation Predictions in Social Media Data using Words That Indicate Location. volume 24, issue 6, pages 1045–1062, proceedings of the 2012 COLING conference on computational linguistics.

Researchers Ren K., Zhang S., and Lin H. (2012) geo-located Twitter users using Tweets and social networks. In this special issue of Lecture Notes in Computer Science, number 7675, Information Retrieval Technology, AIRS 2012, the work of editors Hou Y., Nie JY., Sun L., Wang B., and Zhang P. is presented. Springer, Berlin and Heidelberg.

Baldwin, Paul, Han, and Bo Cook all contributed to a 2014 publication. Predicting Users' Locations from Text on Twitter. This article appears in Volume 49 of the Journal of Artificial Intelligence Research (JAIR). This article has the DOI number of 10.1613/jair.4200.

A piece was written by Kevin Chang, Rui Li, and Shengjie Wang and was released in 2012. Location data may be used to create profiles of individuals and their online relationships via social media and other online sources. The Proceedings of the VLDB Endowment 1.9. 10.14778/2350229.2350273.

Jalal Mahmud, Jeffrey Nichols, and Clemens Drews (2014) found the locations of Twitter users' homes. Part 47 of the ACM Transactions on Intelligence Systems Technology, volume 5, number 3, published in July 2014, spans over twenty-one pages. The URL for the article is http://dx.doi.org/10.1145/2528548.

The authors of the citation are Yasuhide Miura, Motoki Taniguchi, Tomoki Taniguchi, and Tomoko Ohkuma [6]. "Predicting Users' Locations on Twitter using a Simple Scalable Neural Networks Model." 2016 was shown on NUT@COLING.

The people responsible for writing this piece are listed as follows: "A multi-indicator approach for geolocalization of

included in the proceedings of the 2013 Seventh International Conference on Weblogs and Social Media.

For instance, a 2012 paper titled "Towards social user profiling: unified and discriminative influence model for inferring home locations" (pp. 1023-1031) was presented by R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang at the 18th ACM International Conference on Knowledge Discovery and Data Mining.

In 2013, B. Han, P. Cook, and T. Baldwin presented an article titled "A stacking-based approach to Twitter user geolocation prediction" in the proceedings of the 51st Annual Meeting of the Association for Computational Linguistics System Demonstrations. The paper can be found on pages 7-12.

"On the accuracy of hyper-local geotagging of social media content" (pp. 127-136) was the title of the paper given by D. Flatow, M. Naaman, K. E. Xie, Y. Volkovich, and Y. Kanza at the 8th ACM International Conference on Web Search and Data Mining (2015).

A 2014 publication from the IEEE Transactions on Knowledge and Data Engineering titled "Spatially aware term selection for geotagging" was penned by O. V. Laere, J. A. Quinn, S. Schockaert, and B. Dhoedt.

"Who sent this tweet?" was questioned. It was C. Drews, J. Mahmud, and J. Nichols who responded with "No idea." "Supporting the Sixth International Conference on Weblogs and Social Media: Deducing the Geographical Location of Twitter Users," 2012.

[1]