Applied GIS

ISSN: 1832-5505
Vol-9Issue-03 july 2021

LOCATION PREDICTION ON TWITTER USING MACHINE LEARNING TECHNIQUES

JANGILI RAVI KISHORE, ASSISTANT PROFESSOR, jangiliravi.kishore1@gmail.com

VIJAYA BHASKAR MADGULA, ASSISTANT PROFESSOR, vijaya.bhaskar2010@gmail.com

BALAJI SUNIL CHANDRA, ASSISTANT PROFESSOR , hod.cse@svitatp.ac.in

Department Of CSE, Sri Venkateswara Institute of Technology, N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

ABSTRACT

These days, a lot of people are researching how to leverage online social media to anticipate where consumers will go. Researchers have been looking on automatic location recognition for decades, particularly as it relates to or is mentioned in documents. Twitter has attracted a large user base that routinely posts millions of tweets, making it stand out among online social network organisations. These days, location prediction on Twitter is getting a lot of attention because of its global user base and constant posts. Tweets, being brief, loud, and rich-natured communications, present several difficulties for academics in the field. 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. It is shown that the problems depend on these text inputs by describing the content and circumstances of tweets. Using machine learning methods such as naïve bayes, Support Vector Machine, and Decision Tree, we attempt to deduce the user's location from the content of their tweets in this study.

I. INTRODUCTION

Depending on the context, users may choose to make their location openly visible in tweets or make it accessible implicitly by specifying certain criteria. Users of Twitter are free to share photographs that convey a wide range of emotions as the language is not tightly written. Tweets take on an emotive tone when they include abbreviations, misspellings, and messages may be rather distracting. Twitter analysis is not a good fit for the methods used for traditional texts. Without studying the tweet's context, the character constraint of 140 characters could make the tweet difficult to interpret. For Wikipedia and other online content types, we take a look at the geolocation prediction problem. Researchers have spent years studying how to recognise entities in 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. Visitors from certain areas may research local attractions, historical sites, and upcoming events. The user's home location is their stated residence address or the place they provided when they created their account. Several applications may benefit from home location prediction, including recommendation systems, polling, health monitoring, location-based ads, and more. Administrative, geographical, or coordinates may all be used to describe a person's home. Location of the Tweet: The location of a tweet is defined as the geographic area from where the tweet was sent. 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. portrayal of areas inside tweets. Twitter Users may Mention Specific Places: Users may mention specific

ISSN: 1832-5505
Vol-9Issue-03 july 2021

places in tweets. Applications like as recommendation systems, location-based ads, health monitoring, polls, etc., may benefit from referenced location prediction, which might lead to a better comprehension of tweet content. Two location-related sub-modules are included in this study: One is to identify the place stated in the tweet itself. This may be done by gleaning any mentions of geographic names from the tweet's text content. The second is finding the exact location using the tweet's content and matching it to records in a database of geographical information. 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). Al-Saggaf and Simmons (2015), Gaspar et al. (2016), Heverin and Zach (2012), and Oh et al. (2013) all agree that Twitter is an important tool for keeping people informed, getting their status updates, and tracking rescue efforts in the aftermath of both natural and man-made disasters, such as terrorist attacks and food contamination. 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. Platforms such as Twitter and Facebook enable users to provide updates on their social activities become involved with (Mishra et al. 2016). In terms of crisis management, Twitter is a popular option since it gives a platform where both official and ordinary people may share their experiences and advise (Macias et al. 2009; Neubaum et al. 2014; Palen et al. 2010). A great deal of effort is being put into improving this platform so that it can better handle catastrophe management. But, to enhance public reaction, it is necessary to do a more thorough investigation of social media, as proposed by Comfort et al. (2012). Even Turoff et al. (2013) shares this opinion and has urged academics to find ways to get more people involved in disaster relief. Leaders' personal political status may improve if they respond quickly and accurately during disasters (Ulku et al. 2015). In Indonesia, for example, BMKG is one of the authorities that uses Twitter to keep the public informed and provide warnings. Multiple government organisations also make use of social media to assist victims and organise rescue operations. Users of the microblogging service Twitter may post short messages, photos, and audio samples. Users often post and read updates from others because they compose brief messages. 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 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. In addition, the linked 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. Not only is Twitter a great platform for raising awareness, but it also provides a space for individuals to seek for aid when they need it. disaster. It is necessary to distinguish between tweets discussing the crisis and those requesting assistance. Rescue workers will be able to use these tweets as a guide. 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. When it comes to aiding victims, distribution centres are crucial. For relief routing, Burkart et al. (2016) suggests a multi-objective location routing-model to cut down on startup costs. Several studies (Duhamel et al., 2016;

ISSN: 1832-5505
Vol-9Issue-03 july 2021

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). Only 26% of users specify their location as a city or smaller; the rest either use a nation name or use meaningless terms like "Wonderland," according to research by Cheng et al. (2010). While only 0.42 percent of tweets have geotags, 3.17% do, according to research by Morstatter et al. (2013) and Cheng et al. (2010). According to these studies, Twitter isn't very useful for location-based sensing. There has been a meteoric surge in the number of people using Twitter from their mobile devices, thanks to the proliferation of mobile Internet access in recent years. from the end of 2016, 371 million people in India would be using mobile Internet, according to an estimate 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. This is where smartphone batteries come into play, as the GPS system uses quite a lot of power. mAh capacity of the battery. When not in use, users may conserve electricity by turning off their GPS. Alternatively, GPS is essential for the correct operation of apps like taxi hiring services and e-commerce sites like flipkart.com. Results from the study of Twitter users on mobile devices reveal both geotagged and non-tagged messages. Tweets including geotags will be in short supply during crises because people are concerned about draining their phone batteries. Even though English is spoken in many parts of India, most people utilise it while interacting with friends and family online. 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. A mechanism to categorise tweets as either high or low importance is the main contribution of this article. In times of crisis, tweets requesting necessities like food, shelter, medication, etc. are of high priority. Here are two urgent tweet samples. Though Tweet is written in English, the terms used here are from the Hindi language. The tweet reads as follows: "Mr. @narendramodi, people here are very worried about the heavy floods in Chhapra Bihar. Please arrange for administrative help." "Rescue team has done a good job." is an example of a low-priority tweet that conveys disaster-related information. For instance, a user may express gratitude to Twitter for its assistance in the aftermath of a tragedy. If a tweet does not include geo-tagging information, the paper also contributes by predicting the location of high-priority tweets. We construct a Markov chain using the past geo-tagged tweets of the individual users in order to forecast their position. The disaster's spread is determined by analysing the low-priority tweets. Additionally, they may be used to assess how various agencies fared in the event of a crisis.

II. SYSTEM ANALYSIS

EXISTING SYSTEM:

Finding a user's location in social media posts is an issue with the current system. Inverse Location Frequency (ILF) and Inverse City Frequency (ICF) are the social network metrics that stem from and are driven by inverse document frequency (IDF) and term frequency (TF), respectively. After calculating the TF values, they used the frequency values to rake the features. They deduced that local terms had high ICF and ILF values and were scattered out throughout the paper in a few locations. They utilised a model to find terms that were exclusive to a single area or that were suggestive of that area. In an effort to automatically identify, they ranked local terms according to their location and determined the degree of link between words and certain cities.

DISADVANTAGES OF EXISTING

ISSN: 1832-5505
Vol-9Issue-03 july 2021

SYSTEM:

- This article examines the geolocation prediction problem as it pertains to Wikipedia and web page content.
- For years, researchers have studied how to identify entities from these official papers.
- Twitter's location prediction issue is heavily reliant on the content of tweets.
- Method: Inverse Document Frequency (IDF) and Term Frequency (TF)

PROPOSED SYSTEM:

Authentication keys are used to get data from Twitter streams in real-time. The goal of the suggested system is to utilise the user's home location, the location of their tweets, and the content of those tweets to make a location prediction. With the help of three different machine learning techniques, we were able to simplify the prediction process and identify the most effective model. Tweettable receives and stores the live stream of tweets mentioning the term "apple" on Twitter. Tweets in real time can be gathered during the authentication process and the collection of a live stream of tweets using a consumer key, consumer secret, access token, and access token secret. For certain topics, such the names of Indian cities, we have gathered over a thousand tweets. Hashtags are an additional search parameter for tweets.

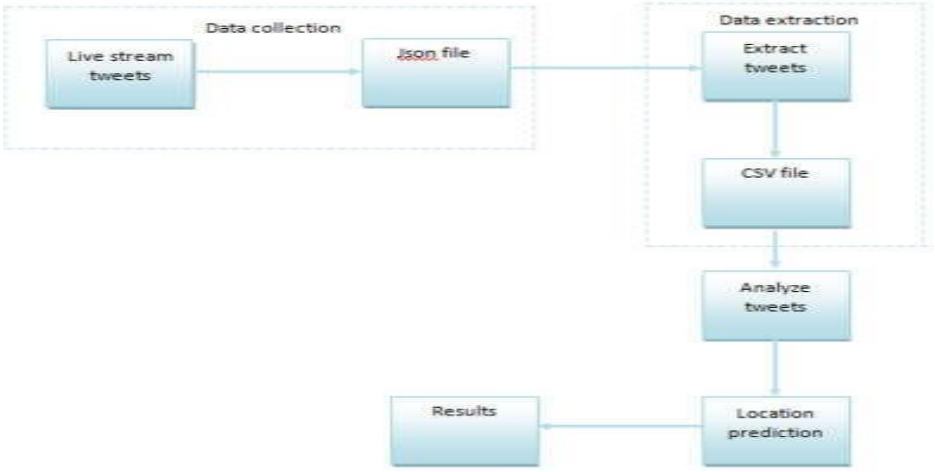
ADVANTAGES OF PROPOSED SYSTEM:

- The data retrieved from real-time sources include tweetid, name, screen_name, tweet_text, HomeLocation, TweetLocation, MentionedLocation, and many more.
- In order to transfer data from Cursor to Pandas Dataframe, the text of tweets is contrasted using the natural language toolkit package that is accessible in Python.
- Scikit-Learn, Numpy, Pandas, and Geography are among of the Python libraries often used in programming.

Algorithm: Naive Bayes, Support Vector Machine, Decision Tree

SYSTEM DESIGN

SYSTEM ARCHITECTURE:



III. IMPLEMENTATION

MODULES DESCRIPTION:

ISSN: 1832-5505
Vol-9Issue-03 july 2021

User:

The first person to register is the user. He needed to provide a working email and phone number when he signed up so he could get future messages. The customer may be activated by the admin after they have registered. User access to our system is granted after the customer has been activated by the admin. Hashtags may be used to search tweets. From Twitter's database, the user will see the first 100 tweets. The current method for determining the user's and tweet's location is based on geocode. The majority of Twitter users would not reveal their exact location. That is why we are considering it a label class. A database will be maintained for all tweets and geocodes. We can use machine learning techniques to test the results of the predictions later on. On the console, you can see the y_pred and y_test. The sklearn.model_selection package allows us to partition the data into learn and test sets. The data was divided as follows: 80% for training and 20% for testing.

. **Admin:**

Login credentials may be used by the administrator. He may activate users once he logs in. Only users who have been activated may log in to our apps. The project's training and testing data may be dynamically set to the code by the admin. The user ran the algorithms on the dataset that was provided. On his displays, the administrator may see the outcomes of naïve bayes, support vector machine, and decision tree analyses.

Data Preprocess:

The text of tweets is stripped of unnecessary characters. Use all capital letters to locate a specific geographic area. Here, we get the precise user locations by use of the geography python package. Tweet should be removed if the user's home address is not included. If a user's tweet location is empty, you should include their home address. Deletes tweets that do not include a location. Retrieve location information from tweets at the end. The last thing to do is use the latitude and longitude information to give the places a float value.

Machine Learning:

Naive Bayes Classification

The Naive Bayes classifier is widely used because it is both popular and simple. By analysing the document's word distribution, this model determines the posterior probability. Word position is not taken into account by the Naïve Bayes classifier when working with the Bag of Words (BOW) feature extraction model. For the purpose of making a label prediction using the provided feature set, this model used Bayes Theorem. Two parts, the "trainset" and the "test set," make up the dataset. The position prediction is obtained by applying NB_model to the test set.

Support Vector Machine

When it comes to classification and regression issues, support vector machine is among the most used supervised learning methods. The method is designed to plot all the data points in an n-dimensional space, where the feature values stand in for the values of each coordinate.

Decision Tree

Using a classifications issue, a decision tree is a learning model. The decision tree module requires at least two pieces of data in order to function. The core nodes of a decision tree represent feature tests, the branches show the results, and the leaves represent the judgements made after training.

IV. 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. In order to demonstrate the best performing algorithm, we have used three different approaches to the geolocation prediction issue. Based on our experimental results, decision trees are the way to go for problems involving tweet text analysis and position prediction.

ISSN: 1832-5505
Vol-9Issue-03 july 2021

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 andSocial 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.