LOCATION PREDICTION ON TWITTER USING MACHINE LEARNING TECHNIQUES

A.V.Srinivas
Assistant professor,CSE (Data Science)
Sreyas Institute of Engineering and Technology,Telangana,India
av.srinivas@sreyas.ac.in

Nandyala Bindu
B.Tech Student of CSE,
Sreyas Institute of Engineering and Technology,Telangana,India
nandyalabindu971@gmail.com

Belide Partha
B.Tech Student of CSE,
Sreyas Institute of Engineering and Technology,Telangana,India
parthabelide@gmail.com

Katta Meghana
B.Tech Student of CSE,
Sreyas Institute of Engineering and Technology,Telangana,India
meghanakatta26@gmail.com

Vaddula Anirudh
B.Tech Student of CSE,
Sreyas Institute of Engineering and Technology,Telangana,India
anirudh.vaddula@gmail.com

Abstract- These days, location prediction of users using online social media generates a lot of research. For decades, researchers have looked on automatic location recognition in relation to or referenced in documents. As one of the most popular online social networking sites, Twitter has attracted a large number of users who send millions of tweets on a regular basis. Because of the global reach of its users and the constant flow of messages, location prediction on Twitter has gotten a lot of attention lately. Tweets, brief, noisy, and rich-natured communications, provide numerous study hurdles to researchers. A general picture of location prediction using tweets is investigated in the suggested framework. Tweet location, in particular, is anticipated based on tweet content. It is very important to outline tweet content and situations highlighted how the difficulties are dependent on certain text inputs. Here, we apply machine learning techniques like as naive bayes, Support Vector Machines, and Decision Trees to estimate the user's location from tweet content.

Keywords:- Social media, Twitter, Tweets, location prediction, Naïve Bayes Support Vector Machine, Decision Tree, Machine Learning

I INTRODUCTION

Users can state their location explicitly in their tweet content, but in some circumstances, the location can be implied by providing certain relevant characteristics. Tweets are a loosely typed language in which users can submit casual visuals with intense emotions. Tweet texts are noisy because to their abbreviated format, misspellings, and excess characters of emotional words. The approaches used to analysis conventional documents are ineffective for tweets. If the tweet context is not researched, the character constraints of tweets of roughly 140 characters may make the tweet difficult to interpret.

For Wikipedia and web page documents, the subject of location prediction, also known as geolocation prediction, is investigated. For years, scientists have studied entity recognition from formal documents. On these documents,several sorts of content and context management are also thoroughly investigated. The location prediction problem on Twitter, on the other hand, is largely dependent on tweet content.

Users in certain regions and locales can look up nearby tourist attractions, sites and structures, as well as connected events.

Home Place: The user's residence address or the location specified when the account was created is considered the user's home location. Home location prediction has several applications, including recommendation systems, location-based ads, health monitoring, and polling. Administrative, geographical, or co-ordinates can all be used to specify the home location.

Tweet Location: The region from which a user posts a tweet is referred to as the tweet location. The mobility of a tweet person can be determined by constructing tweet location. Home location is usually obtained from the user's profile, whereas tweet location can be obtained from the user's geo tag. POIs are widely accepted as representations of tweet regions due to the early viewpoints on tweet location.

Mentioned Location: When writing tweets, users may mention the names of a few locations in the text. Referenced location prediction may improve understanding of tweet content and benefit applications like as recommendation systems, location-based ads, health monitoring, and polls, among others. Two sub-modules of the given location are included in this research: The first is identifying the referenced location in tweet text, which can be done by extracting text content from a tweet that contains geographical names. Second, by converting twitter text to entries in a geographical database, the location of the tweet can be determined.

II LITERATURE SURVEY

Researchers have looked into a number of existing techniques for predicting location from tweet and social media content, and a few of them are presented here.

The author of [1] discusses the difficulty of determining location from social media material. [1] and [2] are the authors.

They arrived to Inverse City Frequency (ICF) and Inverse Location Frequency (ILF) correspondingly, motivated by Term frequency (TF) and inverse document frequency (IDF). They scraped the features using frequency values and then TF values. They concluded that local terms are scattered throughout the document in a few locations 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 attempted to automatically detect local words by ranking them according to their location, and they discovered the degree of linkage of location words with certain locations or cities.

Li et al. [4] proposed the multiple locations profiling (MLP) model, which uses the Bernoulli distribution to find the likelihood of a user's location. Their research shows that this model can accurately determine a user's home location. To estimate the chance of a tweet versus the venue name from each location, the author employed a multinomial distribution.

Mahmud et al. created a categorization model for predicting location, and they enhanced prediction accuracy by predicting regions first, then cities. They used classifier models to track user mobility; if a user travels for a specific period of time, they are tracked to increase prediction accuracy. When the distance between two tweets exceeds 100 miles, the writers assumed the user is travelling.

Machine learning is employed in the majority of extant efforts, with deep learning being recommended in a few cases. Miura et al. [6] employed a neural network for twitter location prediction in their research. The author used neural networks to classify tweets

and users, then combined metadata with tweet texts to train the model. On predictions, their model had a 41 percent accuracy rate.

III PROPOSED METHODOLOGY

The relationship between different components might be depicted using a system architecture diagram. Typically, they are made for systems that comprise both hardware and software, which are shown in the diagram to explain how they interact. It can, however, be made for online applications.

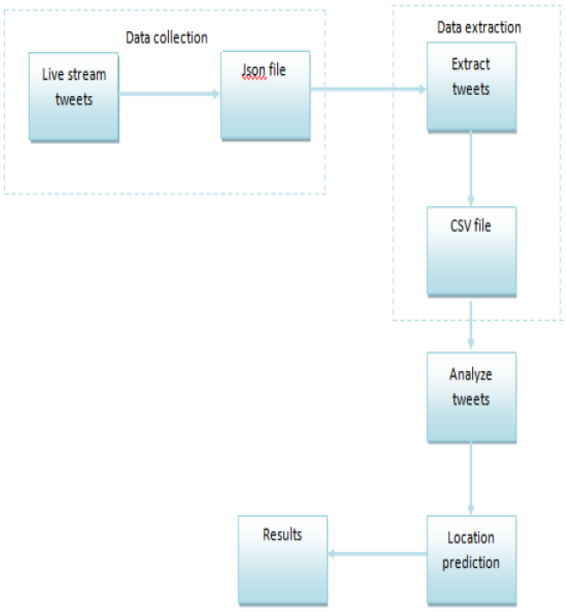

Fig. 1. System Architecture

Authentication keys are used to capture live twitter data as a dataset. The proposed system's goal is to predict a user's location from tweet content by taking into account the user's home location, tweet location, and tweet content. To deal with this, we employed three machine learning algorithms to make prediction easier and determine which model was the best. Tweettable collects and stores a live Twitter stream for the term "apple." By registering a consumer key, consumer secret, access token, access token secret for authentication and collecting a live stream of tweets, live twitter data can be obtained. We gathered over 1000 tweets with specific criteria, such as Indian city hashtag names. You may also use hashtags to find tweets.

A	B	C	D	E	F	G
tweet_id	Name	screen_name	tweet_text	Home Location	Tweet Location	Mention Loca Lval
7.89858E+17	Savishkar Live	SavishkarLive	RT Kerala Govt invites applications from SE	Bhopal India	Bhopal India	Kerala
1.03437E+18	cheeks	unikIln	ito pa	puso mo	puso mo	Nil
2895688956	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 abou	Kerala India	Kerala India	Nil
108272890	Mahesh Veeramalc	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
173706873	Jai Italy Jai Jai Italy	ravi enigma	some serious mental issues out there in Ki	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

The 'twitter.json' file contains a live tweet feed from Twitter for the keyword "apple." By registering a consumer key, consumer secret, access token, access token secret for authentication and collecting a live stream of tweets, live twitter data can be obtained. We gathered over 1000 tweets with specific keywords like 'Chennai, Mumbai, and Kerala.' Tweetid, name, screen name, tweet text, HomeLocation, TweetLocation, MentionedLocation, and Lvalue are among the data extracted from live.

The 'twitter.json' file is read and data is extracted. The following information is extracted: tweetid, name, screen name, tweet text, HomeLocation, TweetLocation, MentionedLocation. To extract data from a json file to a csv file, the tweet text is compared to the natural language tool kit package available in Python.

B. Data preprocessing

1.Extra characters in tweet text are eliminated.

2.To find a geo location, capitalise all terms.

3.If the user's home address is not stated, the tweet will be removed.
4.If the user's tweet location is null, mention their home address.
5.If no location is mentioned in the tweet content, it gets removed.

C. Naive Bayes

The Naive Bayes classifier is the most widely used and basic classifier model. This model calculates the posterior probability based on the document's word distribution. The Bag Of Words (BOW) feature extraction methodology used by the Nave Bayes classifier ignores the position of words within the document. The Bayes Theorem was employed in this model to predict a certain label from a set of features. The dataset is divided into two parts: a trainset and a test set. NB model is used to find the position prediction on the test set.

D. Support Vector Machine

Support vector machines are one of the most widely used supervised learning algorithms, with applications in classification and regression. The algorithm plots each piece of data as a point in n-dimensional space, with feature values representing the values of each co-ordinate.

E. Decision Tree

The learning model that uses categories is the decision tree. The decision tree module divides the dataset into at least two sets. The core nodes of the decision tree represent a feature test, the branch represents the result, and the leafs represent judgments made once the training process is completed.
The Decision Tree functions as follows:

- All training instances are linked to the root node in the decision tree.
- It divides the data into two sets: training and testing.
- It obtains information and selects attributes to label each node. Information with a similar feature property is found in subsets.
- The technique is continued in each subgroup until leafs appear in the tree.

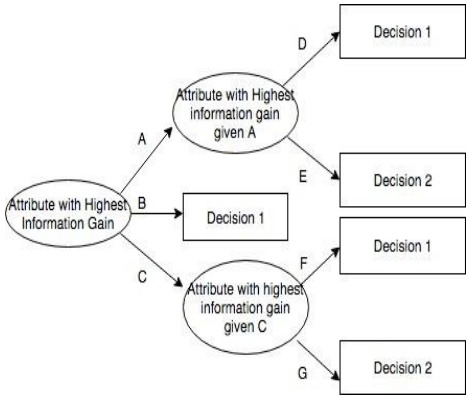

Fig. 3. Decision Tree Model

The tree is built in such a way that no attribute is repeated from root to leaf node. This is done repeatedly to construct every sub tree on the training instances, which is classed down along the route 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 is repeated until the value next node is reached.
Figure 3 shows a sample decision tree that was used.

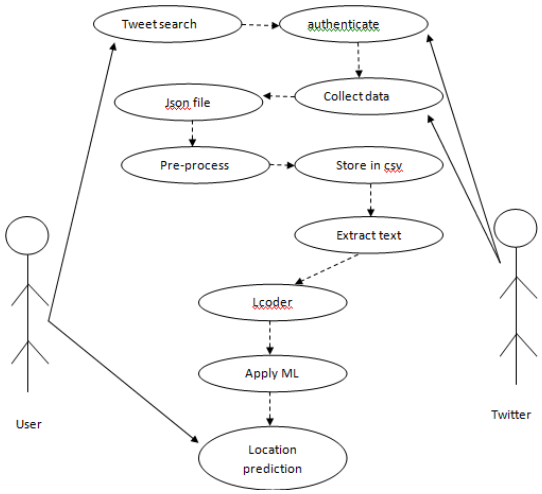

Fig. 4. Decision Tree Model

The features collected from the tweet are listed below the code snippet.

```
(user["features"]["id"],user["featur
es"]["name"],user["features"]["scre
```

```
en_name"],user["features"]["tweet
stext"],user["features"]["HomeLoc
ation"],user["features"]["TweetLoc
ation"]
).
```

Instead of using geo-tags, users can sometimes indicate their location by mentioning their name or landmarks in their tweets. Because place names are vital during pre-processing, we capitalise every word of tweet text to identify geo-locations. Geolocation can be processed in two ways: one is through recognition, which involves labelling the text and then converting it to a location if it is recognised. The entries are then disambiguated, resulting in identified locations.

Prediction Results

	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

TABLE I

IV RESULTS AND DISCUSSIONS

We used the pre-processed dataset for machine learning and applied the Nave Bayes, SVM, and Decision Tree algorithms to it. The dataset is provided 80 percent as training set and 20 percent as test set, we predicted the location and compared accuracy under following chart, Figure 4.
The performance of three machine learning algorithms, namely Naive Bayes, Support Vector Machine (SVM), and Decision Tree, is shown in the table below. Accuracy of prediction is one of the evaluation parameters shown in the table. In terms of efficiency and accuracy, the table clearly shows that decision trees beat the other methods.

Accuracy Comparison

Algorithm	Accuracy
Naive Bayes	43.67
SVM	86.78
Decision Tree	99.96

TABLE II

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

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

TABLE III

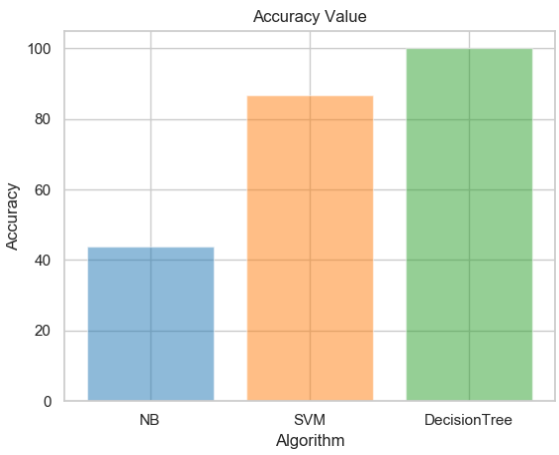

Fig. 5. Performance Comparison

The experimental findings obtained utilising three machine learning techniques are shown in Fig. 5. The accuracy of the Naive Bayes algorithm is around 40%, the SVM algorithm is around 85%, and the Decision Tree algorithm is around 99 percent. As a result of this research, we can

say that Decision Tree is the best algorithm for predicting location in tweet texts.

V CONCLUSION

From Twitter data, three locations are considered: home location, mentioned location, and tweet location. When Twitter data is taken into account, geolocation prediction becomes a difficult task. The nature of twitter language and the limited number of characters make it difficult to comprehend and evaluate. We used machine learning methods to infer a user's geolocation from their tweet text in this study. We constructed three algorithms to demonstrate which one performs the best and is suited for geolocation prediction. According to our findings, decision trees are suitable for tweet text analysis and location prediction problems.

VI REFERNCES

[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.