

Location Prediction On Twitter Using Machine Learning Techniques

¹Prof.Swapnil Wani, ²Mr.Shashank Barve, ³Mr.Prathamesh Khedekar, ⁴Mr.Ankit Yadav

¹Asst.Professor, ^{2,3,4}UG Student, ^{1,2,3,4}Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India. ¹swapnilwani27@gmail.com, ²shshankbarve223@gmail.com, ³prathamkhedekar28@gmail.com, ⁴yankit203@gmail.com

Abstract- This project predicts the position of a user from the text content of a tweet by applying machine learning methods like Support Vector Machine, Naive Bayes, & Decision Tree. These days, predicting location of a user from various social media sites requires extensive research. For decades, researchers have been researching the automatic detection of places associated with or most relevant to records. As a most visited social networking site, Twitter has attracted many users who send several tweets on a regular basis. Prediction of user's position via tweet has drawn a great deal of interest lately. Given the global participation of its regular users as well as continual tweets, the proposed framework investigates the general concept of predicting location via tweets. A tweet's location can be predicted based on its content. It's been highlighted how these challenges are essentially dependent on these text inputs by explaining the content and contexts of tweet.

Keywords— social media, location prediction, Twitter, Machine Learning, Tweets, support- vector-machine, naive bayes, decision tree.

I. INTRODUCTION

Users can provide their location explicitly in their tweet content however, in rare instances, certain things are made implicitly public by providing specific relevant conditions. Tweets aren't a strongly typed language, so users can use them to share casual or emotional images. Tweet texts becomes noisy because of abbreviated language, misspellings, and additional characters for strong emotional terms. Most approaches used to analyze regular papers are unsuitable for studying tweets. If an aspect of posted tweet is not properly examined, Twitter's character constraint of 140 characters may make it difficult to read. The main concern of prediction of location, often known as geolocation precision, is being investigated over web page documents. The problem with predicting location on twitter is that it relies heavily on tweet data.

Home Location: The users residence address or the location specified by the end user when the account was created is considered the users home location. Administrative, geographical, or co-ordinates can be utilized to specify the home location.

Tweet Location: Tweet location is a term for the region from which tweet is posted. The flexibility of a tweet person can be determined by interpreting the tweet location. The location is obtained from user's profile, whereas the tweet location is taken from user's profile geotag.

Mentioned Location: When tweeting, users may mention the names of a few locales in the tweet text.

Prediction of referenced location may help with tweet comprehension and benefits applications like location-based advertising and recommendation system.

II. AIMS AND OBJECTIVE

a) Aim

The project's goal is, to predict the location of Twitter's client by taking users' tweet content. Different machine learning approaches, like support vector machines, naive Bayes, and decision trees, are accustomed to inferring the user's position on Twitter.

b) Objective

The project's goal is to differentiate between the accuracy of various machine learning approaches like naive bayes, & support- vector-machine, & decision tree to anticipate the location and determine which is best among them in agreement with the result. To assist victims in distress, one must include their precise location in their tweet, which is another critical issue in emergency situations. The function of distribution centers in assisting victims cannot be overstated.

III. LITERATURE SURVEY

For predicting twitter users' location by machine learning is

largely depends on the data's availability, which is tweet content. To validate our findings on our research we took help of certain research papers to comprehend the level of finding and maintain accuracy the current system of finding location based on tweet content i.e., based on text language and particular word that use in certain part of a world. They focused on identifying automatically by ranking local words in accordance with their location, and they discovered the degree of correlation of location words connected with specific occasions or cities.

Paper 1: Geolocation Prediction in social media Data by Finding Location Indicative Words: This paper focuses on finding the (LIWS) Location Indicative words through feature selection & determining whether a smaller feature set improves geolocation accuracy.

Paper 2: Where Are You Settling Down: Geolocating Twitter Users Based on Tweets and Social Networks:

This paper develops models for estimating where people will settle based on data from tweets in two dimensions as well as their social contacts when their profiles reveal city & data at town level.

Paper 3: Multiple Location Profiling for Users & Relationships from Social Network and Content:

This paper suggests a training sample expansion method for facial features, in addition with a parallel neural network face detection algorithm.

Paper 4: A multi indicator approach for Geolocalization of tweets:

This is the first paper to present a multi-indicator strategy for determining the user's position and the location of a tweet.

IV. EXISTING SYSTEM

The current system does not address the issue of determining the location on social media material. Inverse document frequency (IDF) and Term frequency (TF) drive and stimulate social networks. The frequencies they arrived at were "Inverse Location Frequency" (ILF) and "Inverse City Frequency" (ICF). These frequency values and TF values were applied to rake the features, which were subsequently raked by TF values. As an outcome, they suggested the concept of the local terms that are spread across the document and have high ICF and ILF values. They approached the model with the goal of detecting local words that are only used in specific areas. They focused on identifying automatically by ranking local words on their basis of location.

V. COMPARATIVE STUDY

SR NO.	PAPER TITLE	AUTHOR NAME	METHOD	ADVANTAGE	DISADVANTAGE
1.	Geolocation Prediction in Social Media Data by Finding Location Indicative Words.	Bo, Cook, & Han Baldwin & Paul Timothy	(LIWs) Location Indicative Words	Using feature selection to find location-indicative words. This feature improves geolocation accuracy.	location indicative words (LIWs) are not very accurate
2.	Where Are You Settling Down: Geo-locating Twitter User Based on Tweets and Social Networks	Shaowu Zhang Hongfei Lin	Inverse Location Frequency (ILF) & Remote Words (RW)	Use name identity recognition to identify placename.	Accuracy is quite less. Combine both techniques give just 56.6% accuracy
3.	Multiple Location Profiling for Users & Relationships from Social Network and Content	Li, Rui & Wang, Shengjie & Chen-Chuan Chang, Kevin	multiple location profiling model (MLP)	MLP Predict 62% User location successfully. It has accuracy of 57%.	Only suitable for locating permanent resident location of user
4.	A multi-indicator approach for geolocalization of tweets	A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. M'uhl	weighted indicators	Capable of detecting 92% of all tweets.	Only 1% of the no. of tweets have geotagged so not suitable for all kinds of tweet

Table 1: Comparative Study

VI. PROBLEM STATEMENT

To create a project using ML methods for predicting the position of users on Twitter based upon a live data stream or dataset. Twitter's location prediction problem is extremely reliant on tweet content. Users that live in specific regions or locations can look up nearby tourist attractions, landmarks, and structures, in addition to connected events. Based upon these findings, it's possible to infer that the Naive Bayes algorithm is implemented for many prediction models, including location- and context-based ones. location prediction system needs to find a solution following problems: identifying terms that are specific to a location or are commonly used in that location within in contents, data extraction, and data pre-processing and analysis of data

VII. PROPOSED SYSTEM

Using authentication keys, live Twitter data is captured as a dataset. The proposed system's purpose is to predict the user's location from tweet content by considering the user's tweet location, and tweet content. In order to deal with this, three ML algorithms must be improved prediction and choose the best model among them. By registering a consumer key(CK), a consumer secret(CS), a consumer token, & a consumer token secret for authenticating and collecting a live feed of tweets, twitter statistics in real time may be obtained. Over 1000 tweets containing specified terms, such as Indian city hashtag names, were gathered. Also, you can search for tweets using hashtags.

VIII. ALGORITHM

The general idea of working of proposed system algorithm is given as follow:

Step 1: Start

Step 2: Collect live feed from twitter to form a dataset

Collect Live tweets are collected as json file

class GetTweetLocatin:

```
def getLocations(self, tweet):
    search_words → "#" + tweet
    date_since → ""
    consumer_key → "
    consumer_secret → "
    access_token_secret → ""

    auth → tw.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)

    api → tw.API(auth, wait_on_rate_limit → True) # Collect tweets

    tweets → tw.Cursor(api.search, q → search_words, lang → "en",
    since → date_since).items(50) print("====>", tweets, dict )

    users_locs → [[tweet.id, tweet.user.name, tweet.created_at,
    tweet.user.screen_name, tweet.text,
    tweet.user.location, tweet.coordinates] for
    tweet within tweets]
Step 3: After collecting data, perform Data
    Pre-processing on dataset. Tweet content is pre-processed:

    #gn → geocoders.GeoNames() return users_locs, dataframe

    # return users_locs

def getLatitudeLongitude(self, cityname):
    geolocator →
```

Nominatim(user_agent → "datapointprojects13@gmail.com")

location → geolocator.geocode(cityname) try:

return location. latitude, location.longitude, location.address
except Exception as ex:

return 0,0, None

Step 4: Split Training Set & test Set. $P \rightarrow df[['latitude', 'longitude']]$

$R \rightarrow df[['userloc']]$

$P_train, P_test, R_train, R_test \rightarrow train_test_split(P, R,$
 $test_size \rightarrow 1 / 3, random_state \rightarrow 0)$

Step 5: Apply Naive Bayes, SVM & Dalgorithm

class UserNaiveBayesClass:

def getNaiveResults(self, df):

$df \rightarrow df[['latitude', 'longitude', 'userloc']]$
 $P \rightarrow df[['latitude', 'longitude']]$

$y \rightarrow df[['userloc']]$

$P_train, P_test, R_train, R_test \rightarrow train_test_split(P, R,$
 $test_size \rightarrow 1 / 3, random_state \rightarrow 0)$

model → GaussianNB ()
model.fit (P_train, R_train)
ypred → model.predict(P_test)

accuracy → accuracy_score (R_test, R_pred) mae
→ mean_absolute_error(R_pred, R_test) mse →
mean_squared_error (R_pred, R_test) rmse → math. sqrt(mse)

r_squared → r2_score (R_pred, R_test) #return round(accuracy,2),
round(mae,2), round(mse,2), round(rmse,2), round(r_squared,2)

return accuracy, mae, mse, rmse, r_squared
Step 6: Calculate and compare accuracy & error values for all above algorithm.

Step 7: Stop.

IX. MATHEMATICAL MODEL

1. NAÏVE BAYES

Naive Bayes is based on Bayes theorem and is a supervised classification algorithm method. It is a classification based on Bayes Theorem and the assumption of predictor independence. The Naive Bayes assumption states that the existence of one attribute in a class has no influence on the existence of additional features.

The following equation expresses the Bayes Theorem:

$$P(B|A)P(A)$$

$$P(A|B) =$$

$$P(B)$$

SUPPORT VECTOR MACHINE

It's an ML algorithm that is supervised & is applied to regression and classification for plotting data points in n-dimensional space, where n denotes various features. The classification is then completed by choosing a suitable hyper-plane that distinguishes two classes. The equation for

SVM is as follows:

$$Q_0 + Q_1X_1 + Q_2X_2 + \dots + Q_nX_n = Q_0 + \sum_{i=1}^n Q_iX_i$$

2. DECISION TREE

It is constructed using the top down approach, beginning with a root node, and includes splitting the data into subsets containing instances with comparable values (homogenous). To calculate a sample's homogeneity, the ID3 method employs entropy.

A. Entropy

The entropy shown in the sample is 0 if it's totally homogenous, and one if it's evenly distributed.

$$E(s) = \sum_{i=1}^C -P_i \log_2 P_i$$

B. Information Gain (IG)

The decrease of entropy after dividing a dataset on an attribute to calculate information gain. It's all about figuring out which attribute delivers the maximum information gain while making a DT (Decision tree).

$$IG(T, A) = Entropy(T) - \sum_{v \in A} \frac{|T_v|}{|T|} Entropy(T_v)$$

3. EVALUATION METRICS FOR PREDICTING LOCATION

A. Distance-Based Metrics

For predicting location in a tweet or metrics In-home location, we aim at making predictions for each user or tweet.

The Euclidean distance between ground-truth and predicted coordinates is therefore defined as the distance (ED):

$$ED(s) = dist(l(s), l^*(s)).$$

B. Median Error-Distance

Given a predefined threshold d of error distance, any prediction whose error distance does not exceed d is regarded as "tolerably correct". The fraction of tolerably correcting predictions is thus defined as the Acc@d measure over the corpus:

$$MeanED = \frac{1}{|S|} \sum_{Xs \in S} dist(l(s), l^*(s))$$

$$MedianED = medians \in S \{ dist(l(s), l^*(s)) \}$$

$$Accuracy = \frac{|\{s \in S : ED(s) \leq d\}|}{|S|}$$

C. Mean Absolute Error:

The average of all absolute errors is the Mean Absolute Error. The Equation is:

$$MAE = \frac{1}{n} \sum_{i=1}^n |X_i - \hat{X}_i|$$

D. Root-Mean Square Error:

This approach is often used to calculate the difference between two numbers projected by a mode or an estimator. The Equation is:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Predicted_i - Actual_i)^2}$$

E. Mean Squared Error:

It indicates the proximity of a regression line to a collection of points. It achieves it by squaring the distances between the points and the regression line.

The formula for this approach is given as follows:

$$MSE = (1/n) * \sum (actual - forecast)^2$$

SYSTEM ARCHITECTURE

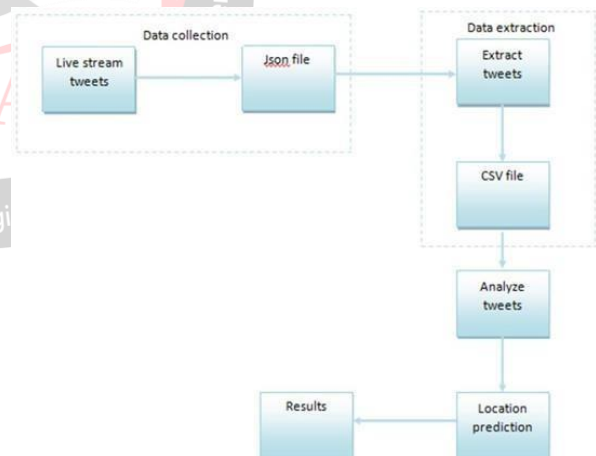


Fig 1: System Architecture which shows steps used for location prediction on twitter.

DISCRIPTION:

Step1. Collecting data from tweet data is taken through live stream tweets & then processed into json file

Step2. After obtaining data it's then extracted into a CSV file.

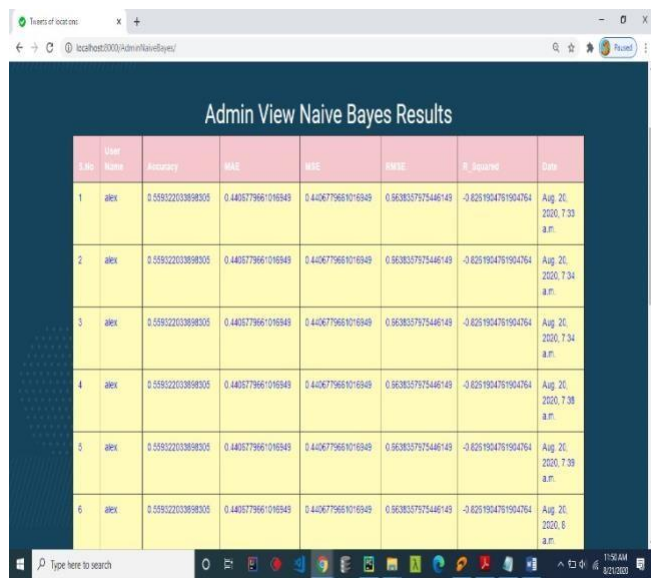
Step3. Evaluation of data later the extracted tweet data is evaluated. Through various tools and technique, position

prediction is done based on the findings.

X. ADVANTAGES

1. To predict the location from a post such as tweet could be significantly useful in the present moment.
2. This data is utilized for damage control or emergency assistance.
3. The data produced through tweet content could help in tracing the origin of tweet and could help in taking down the fake news.

XI. DESIGN DETAILS



S.No	User Name	Accuracy	NAID	NID	NUID	R_Location	Date
1	alex	0.559220338989005	0.4403779661016949	0.4406779661016949	0.9638537975446149	-0.82519047919504764	Aug 20, 2020, 7:33 a.m.
2	alex	0.559220338989005	0.4403779661016949	0.4406779661016949	0.9638537975446149	-0.82519047919504764	Aug 20, 2020, 7:34 a.m.
3	alex	0.559220338989005	0.4403779661016949	0.4406779661016949	0.9638537975446149	-0.82519047919504764	Aug 20, 2020, 7:34 a.m.
4	alex	0.559220338989005	0.4403779661016949	0.4406779661016949	0.9638537975446149	-0.82519047919504764	Aug 20, 2020, 7:38 a.m.
5	alex	0.559220338989005	0.4403779661016949	0.4406779661016949	0.9638537975446149	-0.82519047919504764	Aug 20, 2020, 7:39 a.m.
6	alex	0.559220338989005	0.4403779661016949	0.4406779661016949	0.9638537975446149	-0.82519047919504764	Aug 20, 2020, 6 a.m.

Fig 2. Results provided in the admin page shows accuracy for a particular user location through tweets.

XII. CONCLUSION

Thus, we have tried to implement the paper K. Indira, E. Brumancia, P. S. Kumar, and S.

P. T. Reddy, "Location prediction on Twitter using machine learning Techniques", ICOEI 2019 and according to the implementation, the conclusion is for analyzing of the location indicative words. In this study, several ways for estimating, depending on where the Twitter data is stored, origin, and provided location, are presented. The limited length of twitter material can be a hindrance, Nonetheless, the volume of tweet material enables for more in-depth investigation. The technique for finding the location is heavily influenced by Tweet content. The tweet data is rich content that carries valuable information. We have applied different ML methods were utilized to estimate the location and train the algorithm to produce the best result.

REFERENCE

[1] "K. Indira", "E. Brumancia", " P. S. Kumar" and " S. P. T. Reddy", "Location prediction on Twitter using machine learning Techniques", ICOEI 2019.

[2] "Han", "Bo" & "Cook", "Paul & Baldwin", "Timothy". (2012). "Geolocation Prediction in social media, Data by Finding Location Indicative Words" Coling 2012

[3] "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). AIRS 2012.

[4] "Han", "Bo" & "Cook", "Paul" & "Baldwin", "Timothy". (2014). "Text Based twitter User Geolocation Prediction". The Journal of Artificial Intelligence Research (JAIR).

[5] "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.

[6] "Jalal Mahmud", "Jeffrey Nichols", and "Clemens Drews". 2014. "Home Location Identification of Twitter Users". ACM Trans.

[7] "O.V.Laere", "J.A.Quinn", "S.Schockaer", and "B. Dhoedt", "Spatially aware term selection for geotagging," IEEE 2014.

[8] "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.