IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

User's Location prediction from Tweets using Ensemble machine learning models

Manvik Bhadoria
Pathways School Gurgaon
Faridabad-Gurgaon Road, Baliawas, Gurugram, Haryana 122003

Abstract : On-time and accurate detection of user's home location is a challenging task; however automated detection of location can be used more effectively in applications such as disaster response and public health. Large user-base nature of social media such as twitter makes the application more viable than traditional user-base data like census data. Twitter has been considered as one of the most powerful social media sites due to its worldwide inclusion of users and continuous stream of message from its users. As these tweets are very short text and noisy, identifying the user's location information is quite challenging. To overcome this issue, location information of user's is extracted through geography data. In this proposed framework, the detailed outline of location prediction using tweet text is studied by extraction of location information from user's tweets and user's location and is predicted through their tweet texts. The Natural Language Processing (NLP) technique, Term Frequency-Inverse Document Frequency (TF-IDF) is used for feature extraction from tweets. In this paper, to improve the accuracy of prediction than traditional models, we predict the user's location based on extracted features from tweets using ensemble machine learning models namely Random Forest (RF), Bagging model and Extra Trees model. Experimental results showed that Bagging Classifier (with based model Decision Tree) has achieved the highest accuracy of 99.82% for user's location prediction.

IndexTerms **- Social media, Tweets, location prediction, Machine Learning, Random Forest (RF), Bagging model, Extra Trees (ET), Natural Language Processing (NLP), Term Frequency-Inverse Document Frequency (TF-IDF)**

1.INTRODUCTION

Real time applications such as agriculture, transport and logistics, disaster response, health care and hospitals require distribution of user's locations for bringing up the effective services. Traditional approaches used data from public surveys and information from government portals, however this data is coarse scaled. The recent use of social media among users around the world has increased dramatically over this decade. To overcome the above user-data issues, social media information is serving as a large base of source for user's location information. Twitter has more than 300 million active users worldwide and around 500 million tweets per day as of June 2018 [1]. Twitter offers real-time data sourcing on user's location with high spatial information among a large population.

Despite the large user-based and information on tweets, the real challenges behind the imprecise and missing location information of users. Users post explicitly their location on the tweet text as they check-in several places, whereas in some cases home location of users is available implicitly as they are included in their profile information. In specific, the user's home location is optional mention, thus the missing information can be interpreted from the user's tweet text using geo tags, which has to be processed, and geo information has to be extracted. The user's location information may vary for each tweet as their check-ins also vary, this is more challenging and the extracted location information is more specific, this may lead to highly imbalanced data.

The challenges behind the social media data are that the tweets are not strongly typed, more emoticons are used by the users and shorter text are also used, which makes the text more noisy. This is due to the reason that the number of characters is limited to 140 for each tweet. User's home location is the minor class and the user's tweet location is the major class as the user's check-in multiple points. In this work, the user's home location is predicted using ensemble machine learning models. As this is a classification approach, the tree-based techniques give the best results, the random forest, bagging classifier with base model as Decision Tree and Extra trees models are proposed for the user's home location prediction. In general, a user's tweet information contains three types of location namely

- Home location
- Tweet location and
- Mentioned location

This location information is described further below.

1.1 Home Location:

While creating their accounts, users give the residential address or the location information, which is considered as the user's home location. Predicting the user's home locations accurately helps in many real-world applications including recommendation systems, disaster management, health monitoring, polling etc. Home location is specified in location coordinates or geographical location.

1.2 Tweet Location:

Tweet location is the location as geographical location or co-ordinates where the users post their tweets, it is based on user's check-in locations. These locations can be extracted from user's tweets by geo-tags mentioned in the text. This information helps in identifying user's Point of Interest (POI), which helps in many recommendation systems like restaurants, theaters, etc.

1.3 Mentioned Location:

While posting the tweets, users may mention some specific location on their tweet test, which is referred to as the mention location. User referenced location helps in understanding the user's tweet context and his interests. Understanding this information is helpful for applications such as recommendation systems, location-based advertisements, disaster monitoring, health care etc. This mentioned location information can be extracted from the tweet by using a geographical database, in this work geography package is used.

In this paper, the further chapters discussed below includes Related works in Section 2, Proposed work implementation and algorithms used, and methodology are discussed in Section 3, Experimental results of the work are discussed and analyzed in Section 4 and the work is concluded and discussed the further enhancement opportunities in Section 5.

2. RELATED WORK

Social media content is used to predict user location, this problem is approached by many researchers. As the artificial intelligence techniques machine learning and deep learning are more effectively used in every other industry, the location prediction problem also proposed by ML and DL models. Some of the relevant works are discussed in this section.

User's home location prediction from tweets using deep learning model, Deep Neural Network (DNN) is proposed in the work [2]. As the tweets have noisy content and sparsity in nature, the location prediction is a challenging task. This problem is addressed using the DNN model, the work also handled class imbalance in the location names. The balanced dataset is trained using DNN regression and DNN classification. Random forest model is also used. Experimental results showed that the Random Forest model gained 95.97% recall value. The proposed DNN regression model has gained the highest accuracy of 92.6%.

User next location prediction is proposed in [3], this work used models including dynamic Bayesian network, multi-layer perceptron, Elman net, Markov predictor, and state predictor. In the dynamic Bayesian method apart from predicting the user's next location, time is also predicted, and it was found that time is an independent variable in the prediction. Multi-layer perceptron is used with back propagation model and for the multi-layer perceptron model the optimization of parameters is performed. Elman net was also proposed for location prediction, it is a Neural Network model defined with a multi-layer perceptron with one more hidden layer (which is a context layer). Makov and state predictors are also used for the user's next location prediction. Experimental evaluations showed that accuracy of state predictor model outperformed the other models, the accuracy gained is 81.88%.

User's home location prediction is used in many real-world applications, there are multiple socio-economic attributes computed through user's home location is proposed in [4]. The socio-economic factor like users' income level, occupation, education, are predicted through home location. From home location, the user's personal income, family income, education level, occupation type are predicted with accuracy 45.66%, 50.31%, 52.55%, 54.11% respectively.

User's location prediction as city level is predicted using a named entity using the geographical location mentioned in the tweets in [5]. This work used the Linear Neural Network and Expectation Maximization (EM) algorithm. This work used two type of information local and global distribution. Local information includes frequency feature, user counts, user based average frequency. Global information from the users is also considered in this work. Three types of location information such as location, profile location, and tweet location are considered as single, double and full type model. Experimental results showed that the proposed Named Entity based model has gained the highest accuracy of 71%.

User localization is the relevant area of study that user's location is identified, Multiple-aspect Attentional Graph Neural Networks is proposed in the work [6]. This work take multiple sources of data and the model used has three layers, attention based content learning, attention based network learning and the predictor. In the sentence layer, twitter content are split into sentence tokens, then pre-processed to sentence embedding is given as input. Tweet sentences are converted to lower dimension feature vectors and are generated by word2vec feature extraction technique. Experimental results showed that the accuracy of the model predicted for the distance between predicted cluster center and ground truth is 161km is 67% for twitter dataset.

There are some surveys on the user's next location prediction which also studies its challenges and application, the work [7] addresses the user's next location prediction challenges. In location prediction handling trajectory data is a challenging task, heterogeneous data handling is also complex. The challenges discussed in this survey include handling trajectories, which make it more complex to build clusters of trajectories that best fit with different users' behavior. It was discussed in the survey that developing temporal cyclic patterns helps in accurate location predictions.

Another work [8] proposed the analysis of user reaction for user's location prediction on twitter, which included the news distinctness to predict the location. It was observed in the results that there is significant improvement on rectors from 6.75% to 40% of the user's home location. It was observed from the study that the computing reaction scores of tweets to news articles on similarity computations to the reactions score on their location is high. Feature extraction techniques used in this work includes Word2vec, GLove, Doc2Vec. Experimental results shows that there were highest reactions using Fast Text features for News in Boston of 59% similarity.

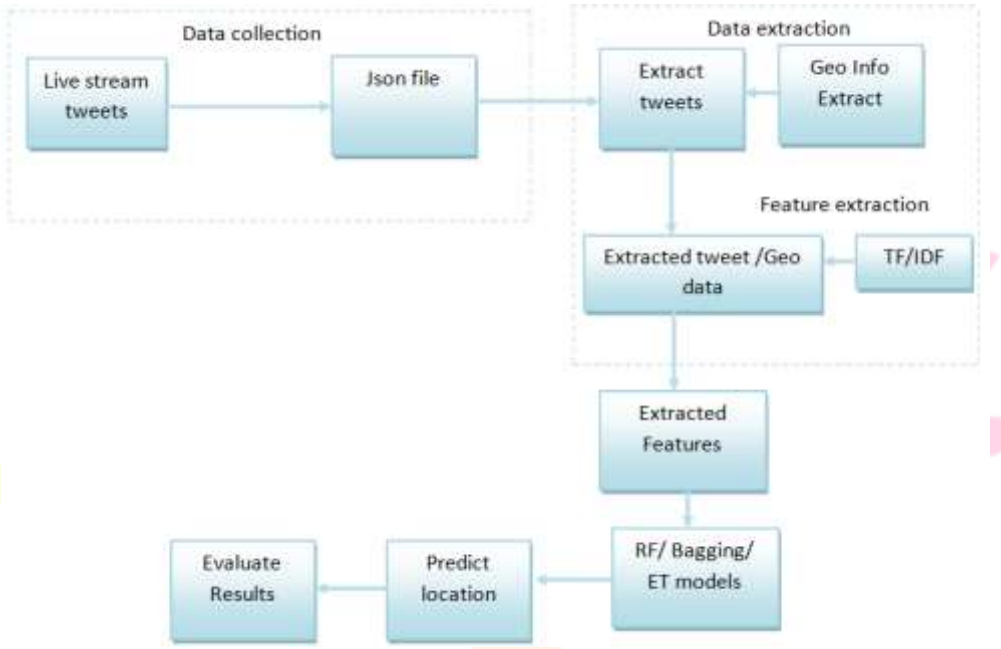
From this literature survey it is inferred that there are many works addressed on user's location prediction using machine learning and deep learning models. However, they achieved good results, the class imbalance problem was not addressed, thus reducing the performance of the model. Thus, in this paper, it is proposed to approach the class imbalance problem as well as the ensemble machine learning models used for the classification of user's location prediction.

3. PROPOSED WORK

The proposed work, user's home location prediction is performed from real time tweets considering their geo-location in tweet text, mentioned location and tweet location. Some of the users may not mention their home location as it is optional, however understanding the user's home location helps in many real-world applications as discussed above. Thus, user's home location is predicted using ensemble machine learning models to achieve higher accuracy. The proposed work has modules data collection, data pre-processing, applying ensemble ML models and evaluating the results are further discussed in this section. Figure 1 represents the overall architecture of the proposed user's location prediction from tweets.

Data Collection

Data collection is performed by streaming live tweets using authentication keys (Consumer Key, Consumer Secret, Access Token, Access Token Secret) are required for API access. The tweets collected are stored in a Json file. The data has been collected with more specific geo tags namely 'Chennai, Mumbai, Kerala'. Extracted data contains user information, tweets, location information and more details are available. However, this work needs limited information like tweet, home location, tweet location thus, this data is extracted from Json file and collected as comma separated values (CSV).



The tweet data extracted are location-based information and tweets, as they are used as the features for learning, home location is the dependent variable, which is going to predict through trained ML models. The data extracted from tweet includes tweet id, name, screen name, tweet text, Home Location, Tweet Location, Mentioned Location. The geography package is used for extracting location information, geo-tags from tweet text and they are stored as the separate feature 'mention location' on the dataset, which refers to the mentioned location in the tweet text.

Table 1: Dataset details of Extracted tweets

Feature	Description
Tweet_id	Tweet id generated for the tweet
Name	User name
Screen_name	User's Screen name
Tweet_text	Tweet text posted by user
Home location	User's home location mentioned by users in their profile
Tweet location	Location from where the tweet is posted
Mention Location	Location mentioned in the tweet text

Table 1 represents the features extracted from the tweet text, there are seven features extracted from the tweet, from which the features used for the learning includes tweet text, home location, tweet location and mention location.

3.1 Data Pre-processing

Data pre-processing involves data cleaning and feature extraction thus to make the data ready for training. As the tweet has noisy as well as short text, it is necessary to clean them to arrive the meaningful information. Data cleaning involves several steps to extract the cleaned test from tweet, they are

1. Special characters are removed from tweet text.

2. Capitalize all words to find for geo location

3. Remove the tweet if user home location not mentioned

4. Mention home location as tweet location, if user tweet location is null

5. Removes tweets if no location is mentioned in tweet text.

The geo-data is converted to numeric values, which represents each location in numeric value using Label encoder. Label encoder assigns each location an integer value.

The tweet text is converted to feature vectors using Term Frequency (TF) and Inverse Document Frequency (IDF), this technique extracts the features from the tweets, the number of features for extraction is given 50 as the cleaned text will be around 50 considering the maximum limitation of the tweet is 140 characters. These feature vectors are concatenated with tweet location and mentioned location, whereas tweet location and mentioned location are the numeric values converted using Label encoders. As the dataset is highly imbalanced, Random Over Sampler technique is applied to make the balanced dataset, this technique generated synthetic samples to make the data balanced. The highest number of class value counts available in dataset is 167 and the least count 1, which makes the data highly imbalanced, which may lead to biased predictions, thus to avoid this problem, the random over sampler techniques re-samples the data to the highest class value count 167 for all classes.

3.2 Random Forest Classifier

Random forest classifier is applied for user's location prediction, the independent feature given are extracted features from tweet, tweet location and mentioned location from tweet, the dependent variable which is going to be predicted is user's home location. The data is split in the ratio 80% training and 20% test dataset. This algorithm follows the multiple decisions given by the Decision Tree making it as an ensemble model. The root node is chosen as the best node based on the Gini impurity function or information gain, thus making it the decision till leaf node as class value. This algorithm splits the data into subsets and each subset is applied to the Decision tree and build the decisions and repeats this process till all the subset is completed. Then make the decision based on the highest voting classes from each tree.

3.3 Bagging Classifier

Bagging (Bootstrap aggregating) is an ensemble learning model, this model reduces the over-fitting and variance of the model thus helps to improve the model performance. The Bagging Classifier works by combining the predictions of multiple instances of the same base model, fore predicting the user's location base model used is Decision Tree Classifier, which is trained on different subsets of the training data. This model works in the following steps.

Steps 1: Initialize the process of Bagging.

Ste 2: Choose a Base Model, for user's location prediction Decision Tree is chosen as base model

Step 3: Randomly generate multiple subsets of the training data by sampling with replacement

Step 4: For generated each subset, train a separate instance of the base (DT) model

Step 5: For a new input, get predictions from each of the trained base models.

Step 6: Bagging model use majority voting to combine the predictions from all base models.

3.4 Extra Trees Classifier

Extra trees (Extremely Randomized Trees) is an ensemble machine learning model, this model improves the prediction performance of Decision tree based on introduction of additional randomness thus avoiding over-fitting. The number of estimators used for this model is 100 and random state 42 is used to generated random subset of training data. The extra tree combines the outcome from several base estimators, generally Decision tree is used as base estimator. Extra Trees selects a random subset of features and random thresholds for splitting dataset as subsets, thus adding more randomness, which helps to decorrelate the trees, making the extra trees more less sensitive to noise and outliers. Extra tree algorithms have the following steps of execution.

Step 1: Initialize the process of building the Extra Trees Classifier.

Step2: Randomly selects the subset of the training data. Each tree will be trained on a different subset.

Step 3: To create extra randomness, for each split in a tree, randomly select a subset of features from the total features available.

Step 4: Creates decision tree by splitting nodes using best random feature selected from the Step 3.

Step5: Repeat steps 2 through 4 to create a specified number of decision trees (n estimator's).

Step 6: For the given new input sample, given to the trained model to get predictions.

Step 7: As the user's location prediction is the classification task, thus it uses majority voting to combine the predictions from all trees.

Step 8: This model aggregates the final output of the Extra Trees Classifier.

4. RESULTS AND DISCUSSIONS

The dataset with feature extracted tweet, mentioned location and tweet location are considered as learning attributes, whereas home location is considered as dependent variable or prediction variable. The feature extracted dataset is applied machine learning models, the dataset is split into 80% training and 20% test dataset. Ensemble models like RF, Bagging and ET are applied. The accuracy of the models is computed and shown in the table 2.

The following table shows the performance of proposed ensemble models Random Forest (RF), Bagging and Extra Trees models. It is observed from the results that the Bagging classifier outperformed other two models in terms of accuracy.

Table 2: Accuracy comparison of Ensemble models

Classification Model	Accuracy (%)
Random Forest	99.68
Bagging Classifier	99.82
Extra Trees	99.74

The following table shows computer error metrics Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE) and R-squared. It is observed from the table that the MAE error is comparatively less for Bagging classifier, whereas Extra trees model also well performed, MAE error for Bagging model is 0.1135 whereas ET model is 0.1590.

Table 3: Error computed for proposed ensemble models

Algorithm	MAE	MSE	RMSE	R2
Random Forest	0.2288	34.4143	5.8663	0.9983
Bagging	0.1135	20.6888	4.5485	0.9990
Extra Trees	0.1590	20.1911	4.4934	0.9990

Figure 5 shows the accuracy computed for proposed ensemble models. It is observed from the figure that the Bagging classifier outperformed the other two models.

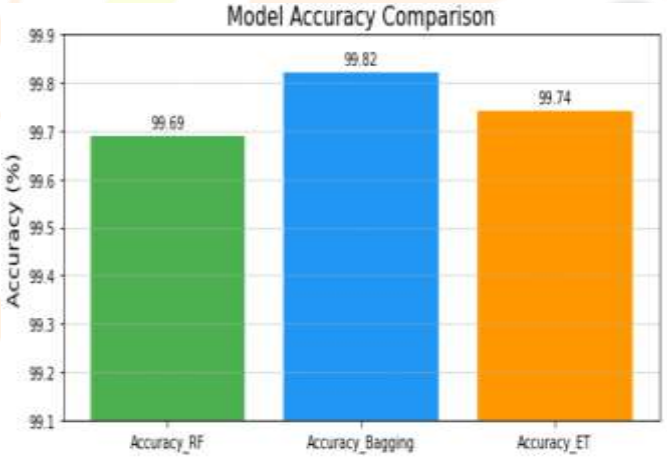


Figure 2: Accuracy computed for Classification models

Figure 6 shows the Mean Absolute Error (MAE) computed for proposed ensemble models. It is observed from the figure that Bagging classifier has less error comparing RF and ET models.

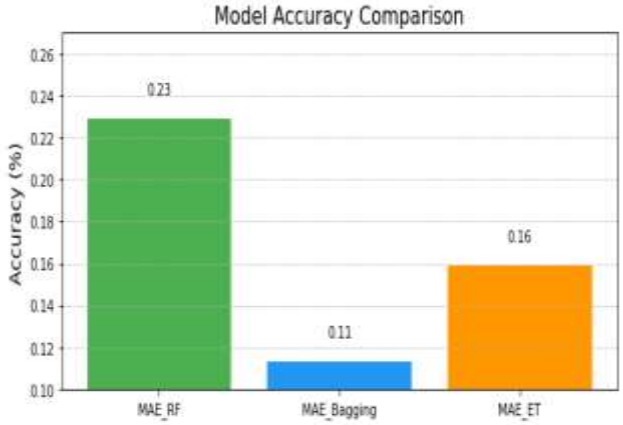


Figure 3: MAE error computed for Classification models

5. CONCLUSIONS

In this paper, we proposed the user's home location prediction from tweets using ensemble machine learning models. Some of the users may not mention the home location, which is more helpful for real world applications, thus identifying the home location is a challenging problem. The mentioned location is extracted from the tweet text using geography package, which take the geo-tag and extracts the places. From the tweet text, tweet location and mentioned location in tweets, it is proposed to identify the home location using ensemble machine learning. The tweet features are extracted using TF-IDF techniques. The ensemble models like Random forest, Bagging classifier, Extra tree classifier are used. Experimental results showed that Bagging classifier with base model Decision Tree has achieved the highest classification accuracy of around 99.82% for user's home location prediction.

As the future enhancement, this study can be further extended to use Deep learning models like Deep Neural Network and optimizing the parameters for DNN model can be proposed.

REFERENCES

[1] Mahmud, Jalal, Jeffrey Nichols, and Clemens Drews. ”Home location identification of twitter users.” ACM Transactions on Intelligent Systems and Technology (TIST) 5.3 (2014): 47.

[2] M. Ghaffari, A. Srinivasan and X. Liu, "High-Resolution Home Location Prediction from Tweets Using Deep Learning with Dynamic Structure," 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Vancouver, BC, Canada, 2019, pp. 540-542, doi: 10.1145/3341161.3342956.

[3] Petzold, J., Bagci, F., Trumler, W., Ungerer, T. (2006). Comparison of Different Methods for Next Location Prediction. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds) Euro-Par 2006 Parallel Processing. Euro-Par 2006. Lecture Notes in Computer Science, vol 4128. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11823285_96

[4] S. Ding, X. Gao, Y. Dong, Y. Tong and X. Fu, "Estimating Multiple Socioeconomic Attributes via Home Location—A Case Study in China," in Journal of Social Computing, vol. 2, no. 1, pp. 71-88, March 2021, doi: 10.23919/JSC.2021.0003.

[5] W. Shen, Y. Liu and J. Wang, "Predicting Named Entity Location Using Twitter," 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 2018, pp. 161-172, doi: 10.1109/ICDE.2018.00024.

[6] T. Zhong, T. Wang, J. Wang, J. Wu and F. Zhou, "Multiple-Aspect Attentional Graph Neural Networks for Online Social Network User Localization," in IEEE Access, vol. 8, pp. 95223-95234, 2020, doi: 10.1109/ACCESS.2020.2993876.

[7] Chekol, A.G., Fufa, M.S. A survey on next location prediction techniques, applications, and challenges. J Wireless Com Network 2022, 29 (2022). https://doi.org/10.1186/s13638-022-02114-6

[8] Jin, Yun-Tae & You, JaeBeom & Wakamiya, Shoko & Kwon, Hyuk-Yoon. (2024). Analyzing user reactions using relevance between location information of tweets and news articles. EPJ Data Science. 13. 10.1140/epjds/s13688-024-00465-2.

[9] I. Hazan and A. Shabtai, "Improving Grid-Based Location Prediction Algorithms by Speed and Direction Based Boosting," in IEEE Access, vol. 7, pp. 21211-21219, 2019, doi: 10.1109/ACCESS.2019.2894809.

[10] C. Gao, Y. Li, J. Yang and Y. Zhang, "A Location Recall Strategy for Improving Efficiency of User-Generated Short Text Geolocalization," in IEEE Transactions on Computational Social Systems, vol. 9, no. 5, pp. 1419-1431, Oct. 2022, doi: 10.1109/TCSS.2021.3116341.

[11] P. Wang, H. Wang, H. Zhang, F. Lu and S. Wu, "A Hybrid Markov and LSTM Model for Indoor Location Prediction," in IEEE Access, vol. 7, pp. 185928-185940, 2019, doi: 10.1109/ACCESS.2019.2961559.

[12] C. Su, Q. Zhou, X. Xie and D. Wu, "Personalized Check-in Prediction Model Based on User’s Dissimilarity and Regression," in IEEE Access, vol. 7, pp. 79418-79432, 2019, doi: 10.1109/ACCESS.2019.2923435.