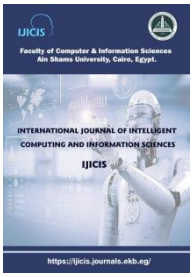
IJICIS, Vol.21, No.3, 119-133
DOI: 10.21608/ijicis.2021.84159.1111

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



A Location Prediction Methods: state of art

Aml Mostafa*
Information system Dept.
Faculty of Computer and Information Sciences
Ainshams University
Cairo, Egypt
Aml.mostafa@cis.asu.edu.eg

Walaa Gad
Information system Dept.
Faculty of Computer and Information Sciences
Ainshams University
Cairo, Egypt
Walaagad@cis.asu.edu.eg

Tamer Abdelkader
Information system Dept.
Faculty of Computer and Information Sciences
Ainshams University
Cairo, Egypt
Tammabde@cis.asu.edu.eg

Nagwa Badr
Information system Dept.
Faculty of Computer and Information Sciences
Ainshams University
Cairo, Egypt
Nagwabadr@cis.asu.edu.eg

Received 2021- 7-5; Revised 2021-9-18; Accepted 2021-10-7

Abstract: The rapid use of social media made location prediction the key to research studies based on-location services like; advertising, recommendations, climatological forecast, and security system. Locations are the center of information for these applications. According to millions of users who post tweets every day, Twitter is known as one of the most important and familiar social media blogs. Depending on the importance of catching the location of the users and the rapid usage of Twitter, Location prediction on Twitter has been a point of research in many studies. This survey provides a comprehensive overview picture of the prediction of the user's location on Twitter. that focuses on the home location prediction and tweet location prediction. This occurs by; first, defining these two kinds of research and the inputs of these research views that are content, network, and context. Then, reviewing existing location-prediction techniques and the latent challenges. Finally, the conclusion of the survey and a list of the future research directions.

Keywords: Twitter, Location Prediction, , Home Location prediction, Tweet Location Prediction, Social media.

* Corresponding author: Aml Mostafa
Information system Dept., Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
E-mail address: Aml.mostafa@cis.asu.edu.eg

1. Introduction

In recent years, microblogging services platforms Like; Facebook and Twitter have been used on daily basis and spread all over the world. Especially Twitter which has more than 300 million users who are always active and post more than 500 million tweets daily1]. On the other side, Twitter limits the tweet posts' text to be maximum 280 characters, which makes users write tweets constantly. Users on Twitter can share anything they are thinking of, such as, opinions, feelings, and activities. Nowadays Twitter has become a virtual world that people live in with their friends or followers and they can make new friendships or relations too. This virtual online world has intersections with the actual world, where locations link these two worlds. According to the rapid usage of social media platforms, researchers and organizations are interested in the updated data that has been provided by users on these platforms. The general profile of each Twitter user has the address of the user, that is called home location. Their home locations push them to write tweets about their surrounding events, share news and activities in their surrounding area.
like clubhouses or restaurants. In this survey, we focus on these two types of Twitter locations, particularly home location and tweet location.
Many studies have been produced and applications were developed to get the benefit of these social human-powered networks; such as, advertisements [2], crime prediction [3], restaurant detection, recommendation [4], public health [5], pandemic inspiration, tourism [6], locations of emergency [7], etc. Despite this, the quantity of geotagged tweets is declared to be only around 1–3% of the total number of tweets [8,9]. The features of the home location and time zone can be inaccurate because users may write any home location or time zone without any protocol. These features have been used by some researchers to estimate the accuracy, but others only use the content of the tweet to evaluate the range that can be mapped using a gazetteer [10]. Although an accurate location does not mean the real location of the user, only 66% of users insert their exact location information [11]. In 2009, the Twitter platform maintained an additional property which is tracking users' location linking their latitude and longitude, this feature is called per-tweet geo-tagging.
Researchers used text processing methods to analyze tweet text on Twitter to solve location prediction problems like ambiguation. When the user stays in a specific area, he talks about this area and shares news or events. The user can use GPS explicitly for sharing his location or implicitly including the text with specific related words. Moreover, researchers investigated the problem of sharing location explicitly or implicitly including the text. As mentioned before the tweets are limited to be ~~a~~ maximum 280 characters as its Twitter's policy that makes the tweets too short to be clearly understood, especially for the ones who are uninformed of the context of the tweets. Acronyms, ambiguations, and misspellings are general problems in text processing as the users write their tweets in a casual way or slang. On the other side, Twitter is an abundance of information about users, taking into consideration the relationships of the users on Twitter, they may add geo-location explicitly or implicitly in their text that helps in this research point.
Three kinds of information are used as input to solve location prediction problems on Twitter even home location or tweet location. The content of the tweet, the network of Twitter, and the context of the tweet.
the Content of the tweet is defined as the body or the text of the tweet, This text has a limited length of up to 280 characters. The user can tweet anything happening, or what he is thinking of. There is another type of tweet called retweet which the user can share others tweets on his own profile. Both tweets and retweets will appear on the user's profile.

The network of Twitter is defined as friendship, friendship on Twitter does not necessarily mean that they have the same friendship in real life. Two strangers in real life may become friends on Twitter by luck or because they are sharing the same interests or ideas. Moreover, friends in real life often mention each other on Twitter [12], [13].
The Context of the tweet According to the rapid spread of tablets and smartphones that may enable GPS automatically, a tweet may be attached to the geotagged location of the user and the timestamp of the posted tweet, also users can complete the information on their profile through this. It also supports the researchers to understand the users' behavior through; geotagged tweets, timestamps, and the profile of the user belonging to the context of the tweet.
Our study aims to review the location prediction problems on Twitter including home location prediction problems and tweet location prediction problems. This paper is organized as follows: Section 2 explains the prediction of home location on Twitter. Section 3 explains the prediction of the tweet location problem on Twitter. Section 4 explains the discussion and the overview of the research papers. Finally, the conclusion is in the last section which discusses the future work.

2. The prediction of home location

In this section, the problem of Home prediction on Twitter is defined. Each user on Twitter has a long-term household address on his profile. This location is called the home location of the user. The awareness of this location can be used in several applications the most important of them is recommendation, advertisement, and public health.
The applications that are interested in home location can obtain this information from the user's profile. but there are a lot of profiles that are incomplete, and this data are not available. There are studies that depend on users' tweets to predict the home location. Some studies observes the most common city tagged in the tweets of the user, others depends on the first geotagged available location, and others calculate the median of the geotagged location of the user.
Researchers seek this data as it is very beneficial. The challenge of knowing this data is that many users did not write this data in their profiles on Twitter as it is optional according to Twitter's policy. In this section, we categorize the studies of home location prediction based on the inputs i.e., the content of the tweets, the network of Twitter, the context of the tweet, and hybrid. Hybrid means that the study has more than one input.

2.1. Prediction based on the content of the tweet.

A tweet Content is defined as the body or the text of the tweet. In this section, the studies inferring home location based on the content of the tweet will be described. The home location of the user can be shown in specific words in the tweet. For example, people in New York would mention red bulls in their tweets more than people who live in another state. In [11] they depend on classification for solving this problem, they use Naive Bayes for training the dataset and predict the home location. In [14] they use sparse coding and the techniques of lexicon learning to extract the features of the word.

In [15] use a hierarchical classification, they report the geo-coordinates in adaptive grids to find the centroid of the locations of the user. They found that the centroid of the location is better than the mid-points of the reporting grid. Authors in [5] apply the model from this study [16] which is the fitted spatial variation model to get the smoothed distribution, they enhance the work in [8] by applying the one-peak model from [16] by a wave which can get the distribution of the words by allowing multi-peaks. The authors in [17] applied Inverse Location Frequency (ILF), they also applied Inverse City Frequency (ICF) to determine the area spot of the words in the tweet. They assume that the distribution of the words in fewer locations has more Inverse Location Frequency (ILF), and Inverse City Frequency (ICF) values. based on Information Retrieval (IR) measures.
In [18] the authors assume that the distribution of the local words in the tweet content can be biased than the normal words in the content of the tweet using information theoretical studies like K-L divergence. Table 1 summarize the previous works based on the tweet content as an input for the Home Location Prediction (HLP).

Table 1 A summary showing previous Home Location Prediction (HLP) based on the content of the tweet

Work Reference	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	State, Country	Tweets	Accuracy
Error! Reference source not found.	Grid	Geo Text data	Mean, Median
Error! Reference source not found.	Grid	Data from [17] [19]	Accuracy, Mean, Median
Error! Reference source not found.	City	Geo-tagged tweets	Accuracy, Mean
Error! Reference source not found.	City	Geo-tagged tweets	Accuracy, Mean
Error! Reference source not found.	City	Data from [19], geotagged tweets	Accuracy, Mean, Median
Error! Reference source not found.	Grid	Tweets	Recall, Precision, Median

Table 1 shows a summary of the previous works based on the tweet content as an input for the Home Location Prediction (HLP). The comparison is done between the granularity level, dataset used for training and testing, and the performance measures.
The granularity level is categorized into three categories of granularity:

- Administrative level like country, state or city where the users live in.
- Geographical grids level is the ground which is divided into cells and the cell that the user stays in is called home location cell.
- Geographical coordinates level like locations is represented by the longitudes and latitudes.

3.2. Prediction based on the network of Twitter.

The network of Twitter can be defined as friendship, the user on Twitter can follow others. In [20] assume that the user lives in the city that almost his friends lived in. In [21] suppose that

the higher rate of the location of the user's friends where they lived in, it means that the location has more probability to be a user's location. They apply this model in only mutual friends. In [22] observe that the protected account means the two users' locations are approximately close, they use a decision tree to attach the users.
In [13] utilize the mentions of the user with his friends, they make a graph of mention friends and suppose the user's location is very close to the location of the mentioned friends. In [23] also depends on the mentioned relationships. In [24] suppose a landmark to predict the home location of the user, they consider the landmark is a region that a user and a lot of his friends live in. Table 2 summarize the previous works based on the friendship network.

Table 2 A summary showing previous Home Location Prediction (HLP) based on the network of Twitter

Work Reference	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	City	Tweets	Accuracy, Mean
Error! Reference source not found.	City	Tweets	Recall, Precision
Error! Reference source not found.	Coordinates	Geo-tagged tweets	Accuracy, Mean
Error! Reference source not found.	Coordinates	Geo-tagged tweets	Recall, Mean, Median
Error! Reference source not found.	Coordinates	Geo-tagged tweets	Median
Error! Reference source not found.	Coordinates	Data from [29]	Accuracy, Median, Recall, F1

Table 2 shows a summary of the previous works based on the friendship network as an input for the Home Location Prediction (HLP). The comparison is done between the granularity level, dataset used for training and testing and the performance measures.

3.3. Prediction based on the context of the tweet.

In this section, the research papers that are based on the context of the tweet as input for Home Location Prediction (HLP) will be mentioned. In [25] the authors determine the home location and work area of the user by using a probabilistic model that links the distribution of the geo-tags tweet of the user to the user's activities. The used methodology in this paper depends on the user's posting time, they suppose that the tweets in the working hours in the morning tend to be posted from the work area, and the tweets during sleep time at night tend to be posted from home location of the user**.** In [26] the authors suppose that the user has several activities during the day, they follow the cluster geo-tags location of the user and conclude that the cluster that has the highest number of tweets is the home location of the user. They notice this method instead of detecting home location by calculating the median of the clusters. Moreover, they assume the home location coordinates of the user are the median of the home cluster. Table 3 summarize the previous works based on the context of the tweet.

Table 3 A summary showing previous Home Location Prediction (HLP) based on the context of the tweet.

Work Reference	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	Grid	Data from [27], geotagged tweets	Accuracy, Mean
Error! Reference source not found.	Coordinates	geotagged tweets	Accuracy, Mean

Table 3 shows a summary of the previous works based on the context of the tweet as an input for the Home Location Prediction (HLP). The comparison is done between the granularity level, dataset used for the training and testing process, and the performance measures.

3.4. Prediction based on hybrid methodology.

The hybrid methodology means that the researchers depend on more than one input in their studies, the input may be (content, and network) or (content, and context), or (content, network, and context) the most common input is the content of the user's tweet as the researchers can predict the user's home location. Firstly, we mention the studies that depend on content and network. The authors in [28] applied Inverse Location Frequency (ILF), they also applied Inverse City Frequency (ICF) to determine the area spot of the words in the tweet. They assume that the distribution of the words in fewer locations has more Inverse Location Frequency (ILF), and Inverse City Frequency (ICF) values based on Information Retrieval (IR) measures. For the network input they suppose that the higher rate of the location of the user's friends is where they live in, it means that the location has more probability to be a user home location. The authors in [29] apply the probabilistic models to solve the Home Location Prediction (HPL) problem. They apply Gaussian mixture models to perform the distribution of the words, . In addition, they calculate the probability of the user x following user y.
In [30] the authors also apply the probabilistic model, they divide the work into two dimensions, the first dimension is assuming the probability of the venue names of the tweets to location and the second dimension is assuming the probability to random (not location-based). Furthermore, they apply the Bernoulli distribution to determine the tweet based on which dimensions are location-based or posted randomly. They enhance their work in [29] by assuming that several people have more than one home city, as long as the working city may not link to their home city. In [31] applying Term Frequency–Inverse Document Frequency (TF-IDF) vectors, they use unidirectional mention instead of bidirectional mention, as they find the bidirectional mention is less useful than unidirectional mention and rare.
Secondly, we mention that the studies depend on content and context. The authors in [32] utilize hierarchical classification rules to decide the local words to groups of state-city or timezone-city. They enhance their work in [33] by deleting the traveling people from the dataset to enhance the results for detecting the user's home location. They consider people who post two or more tweets with distance more than 100 miles are travelling people.
Finally, mention that content, network, and context as inputs for home location prediction. The authors in [34] apply a neural network to solve the home location problem, they encode the tweets' content, context, and the friendship network information to the Recurrent Neural Network (RNN) model. They are different from other works is the separation to the users in the linked network and their similar cities. Table summarize the previous works based on the hybrid methodology

Table 4 A summary showing previous Home Location Prediction (HLP) based on hybrid methodology.

Work Reference	Input Approach	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	Content, network	City	Data from [5], [35]	Accuracy, Mean
Error! Reference source not found.	Content, network	City	Tweets	Accuracy, Mean
Error! Reference source not found.	Content, network	City	Tweets	Accuracy
Error! Reference source not found.	Content, network	Coordinates	Data from [19], [17], [27]	Accuracy, Mean, Median
Error! Reference source not found.	Content, context	City, State	Geo-tagged tweets	Accuracy. Recall
Error! Reference source not found.	Content, context	City	Geo-tagged tweets	Accuracy. Recall
Error! Reference source not found.	Content, network, context	City	Data from [19]	Accuracy, Mean, Median

Table 4 shows a summary of the previous works based on the hybrid methodology for the Home Location Prediction (HLP). The hybrid methodology means that the researchers depend on more than one input in their studies. The comparison is done between the input, the granularity level, dataset used for the training and testing process, and the performance measures. We find the tweet content input is common in the hybrid methodology, and it is rare to depend on network and context only.

3. The prediction of tweet location

Tweet location is different from home location. Tweet location means the area or region where the tweet is written, and where the user was when he wrote this tweet. As mentioned before home location can be obtained from users' profiles or geotagged tweets of users or both. However, the tweet location can be obtained from geotagged tweets only.

Every day there are more than 500 million tweets written by more than 300 million users [36]. Users send the tweets for sharing with their friends. any information, feelings, opinions, recommendations or asking questions. For example, a user may send a tweet to recommend a restaurant where he ate delicious food. If the name of the restaurant is written clearly or mentioned as a tag in the tweet, this will support the restaurant, it is like an advertisement. Unfortunately, the geotagged tweets quantity is stated to be around 1–3% only of the total number of tweets [8] and [9]. Moreover, the tweet location prediction gives the motivation to understand the user's mobility and define his location. The tweet location of the user means the place where the tweet has been sent out by the user. It differs from the home location prediction problem in the input as the home location prediction problem using all the tweets of the user as an input, but in tweet location prediction the input is only one tweet. However, in this section, we mention the studies of tweet location prediction based on the inputs i.e., the content of the tweets, the network of Twitter, the context of the tweet and hybrid. which means that the study has more than one input.

3.1. Prediction based on the content of the tweet

In this section, the previous studies for tweet location prediction based on the content of the tweet as input will be described. The authors in [37] use Information Retrieval (IR) techniques to predict the tweet location, they assume the location by ranking functions by using IR models like KL-divergence between the local words that indicate location and normal tweets. The authors in [38] also applied Information Retrieval (IR) techniques using KL-divergence as the Retrieval method. In this research paper, they use the data from the Foursquare microblog. (Foursquare is a trusted location site to understand the user's mobility in the real world). The idea is to use the data from Foursquare to predict the tweet location using Laplace smoothing and Jelinek-Mercer smoothing to distribute the words, but no enhancement in the performance.

The authors in [39] apply Hidden-Markovbased, they apply language model on geotagged tweets to predict the tweet location on city level granularity. In [40] the authors treat this problem as a classification problem; they extract the features from the words in the content of the tweet and classify them into cell grids. However, they found a problem when the cell grid size is too small, they couldn’t complete the prediction process in the right way, they handle this problem by applying Gaussian kernel to calculate the probability of each cell in the grid, then each word in the tweet is presented as a cell in the grid. In this work [41] the goal is to infer the Point Of Interest (POI) of the user like a restaurant, cinema, or club instead of the tweet's exact location. They investigate the relationship between the content of the tweet and feature classes as it differs from the home location prediction in the class numbers. The problem is that the number of classes is huge, there may be thousands of Points Of Interest (POI) for the user in one city. Moreover, in [42] they apply Neural Network (NN) to solve the tweet location prediction problem. They use the tweet content to apply a convolutional mixture density network, then applying the Gaussian mixture model to predict the coordinates of the tweets. Table 5 summarize the previous works based on the tweet content as an input for the Tweet Location Prediction (TLP).

Table 5 A summary showing previous Tweet Location Prediction (TLP) based on the content of the tweet.

Work Reference	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	State, Country	Geo-tagged tweets	Accuracy
Error! Reference source not found.	POI	Foursquare data, geotagged tweets	Recall, Precision
Error! Reference source not found.	City	Geo-tagged tweets	Accuracy, Mean, Median
Error! Reference source not found.	Coordinates	Data from [27], geotagged tweets	Mean, Median
Error! Reference source not found.	POI	Geo-tagged tweets	Accuracy, Recall, Precision
Error! Reference source not found.	Coordinates	Geo-tagged tweets	Mean, Median

Table 5 shows a summary of the previous works based on the tweet content as an input for the Tweet Location Prediction (TLP). The comparison is done between the granularity level, dataset used for training and testing, and the performance measures.
The granularity level is categorized into three categories of granularity:

- Administrative level like country, state, or city where the users live in.
- Geographical grids level is the ground, which is divided into cells, the cell that the user stays in is called home location cell.
- Geographical coordinates level like locations are represented by the longitudes and latitudes or Point Of Interest (POI).

3.2. Prediction based on the Network of Twitter.

Tweet Location Prediction (TLP) differs from Home location Prediction (HLP) in granularity level, we find that tweet location expressed more granular than home location. Coordinates or Points Of Interest (POI) are being used instead of the administrative level like a city. One of the challenges in predicting tweet location is the tweet, which is almost very short, so we find researchers try to use the friendship network of the user to solve this problem.

Authors in [43] taking the friendship network of the user into their consideration, they utilize the friend's location of the user as an input besides his historical location data. They use Dynamic Bayesian Network (DBN) in the training step for each user in the dataset, the data is trained on over ten thousand users. Each user may have over hundred geotagged tweets. Their methodology is applied also to the non-linear model. For example, if two users are working in the same workshop and the system in this workshop is night and day shift, the model can use the historical data for those users and predict if one has a shift in the workshop, the other is at home at this time. Table 6 summarize the previous work based on the friendship network as an input for the Tweet Location Prediction (TLP).

Table 6 A summary showing previous Tweet Location Prediction (TLP) based on the Netword of Twitter

Work Reference	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	Coordinates	Geo-tagged tweets	Accuracy, mean

Table 6 shows a summary of the previous work based on the friendship network as an input for the Tweet Location Prediction (TLP). We did not find much more research papers which depend on the friendship network only, the most common are the content and the friendship network together.

3.3. Prediction based on the context of the tweet.

The time of the tweet is a characteristic input for the prediction process. For home location prediction, the researchers depend on the distribution of the tweet time [32, 33]. However, for the tweet location prediction, the access would be at the time of the tweet not to the time distribution of the user in general. Moreover, a value of time stamp information would be very useful if enough user's data are known. For example, historical data of the user can suggest that the club leads to posting more tweets at night, but the park leads to posting more tweets on weekends. Unlike Home Location Prediction (HLP) it is very difficult to depend on only the context of the tweet in the prediction process. The most common are the content and the context together as a hybrid input.

3.4. Prediction based on hybrid methodology.

The hybrid methodology means that the researchers depend on more than one input in their studies. Finding the tweet content input is common in the hybrid methodology and it is rare to utilize network and context only without the content feature. Authors in [44] use a Naive Bayes Model to link the words with a venue, if the tweet text is too short or did not has enough information, they take words from users' tweets to complete the prediction process. The assumption depends on the same user visits and the same place more than one time. In the network part, they assume a different theory. They do not depend on the user's followers as mentioned in other papers, they assume the users have the same history content, they have the same history for visiting the venue. Authors in [45] treat with prediction problem as a classification problem. They classify the content of the tweet to geo-location. Authors in [46] use Information Retrieval (IR) techniques to predict the tweet location, assuming the location of the user by ranking functions and using IR models like KL-divergence between the local words that indicate location and normal tweets. They divide the timestamp into three main classes: day, week, and month. In [4] they assume that each user has regions where he most visits every day, like region (home) or region (university) or region (work). Using Gaussian distributions to center the user between these two regions (home), and (work). Taking the timestamp into consideration, they divide the time stamp into two parts weekday or weekend for each user. and from the history of the user, they conclude the user's preference visiting places. Table 7 summarize the previous works based on the hybrid methodology for the Tweet Location Prediction (TLP).

Table 7 A summary showing previous Tweet Location Prediction (TLP) based on hybrid methodology

Work Reference	Input Approach	Granularity Level	Dataset	Performance Measures
Error! Reference source not found.	Content, network	POI	tweets, foursquare data,	MRR
Error! Reference source not found.	Content, network	POI	Geo-tagged tweets, foursquare data,	Accuracy, Mean
Error! Reference source not found.	Content, context	POI	Geo-tagged tweets	Accuracy
Error! Reference source not found.	Content, context	Coordinates	Data from **[27]**, geotagged. tweets	Accuracy, Mean

Table 7 shows a summary of the previous works based on the hybrid methodology for the Tweet Location Prediction (TLP). The hybrid methodology means that the researchers use more than one input in their studies. The comparison is done between the input, the granularity level, dataset used for the training and testing process, and the performance measures.

4. Discussion

The problem of location prediction on Twitter is defined. Moreover, the input of this problem that helps the researchers to solve and infer the location. there are four types of inputs as mentioned before: the tweet content, the network of Twitter, the context of the tweet, and the hybrid methodology. the tweet content means that the text body of the tweet, the network of Twitter means that the friendship of the user (followers and followees), we can define the context of the tweet as the timestamp of the tweet, and the hybrid methodology means that the researchers use more than one input in their studies for location prediction. We observe that the researchers usually use the user content input to resolve this problem, and the content of the user is a very common input in most of the previously mention studios. Regardless of the input of the research, we will focus in this section on the methodology of the research we found the papers [37,38,46] they treat the prediction problem as an information retrieval (IR) problem. they use Information Retrieval (IR) techniques for solving location prediction problems, they all assuming the location of the user by ranking functions and using IR models like KL-divergence between the local words that indicate the location and normal tweets. in[37,38] they use the content only as input but in [46] they use the hybrid methodology they use the content and the context as an input, they divide the timestamp into three main classes: day, week, and month. we found the adding timestamp achieves a better result than uses the content only in these techniques.

On the other hand, the most technique that has been used to solve the location prediction problem is classification, we found the most of the researchers used the classification for training the dataset and predict the location, in [39] the authors use a hidden Markov based model, the authors in [40, 43] uses a Naive Bayes classifier Model to link the words with a venue, we found an advantage in [43] is better than [4040] as they treating with the short text, they take words from users' tweets to complete the prediction process. The assumption depends on the same user visits and the same place more than one time. besides that, they are taking the relationship of the user into the consideration, they assume the users have the same history content when they have the same history for visiting the venue. plus, the authors in [40] use the content of the user only but the authors in [4443] use hybrid methodology, they used the content of the user and his relationship between him and his friends. we observe the authors in [4141, 42] also going to classification models for solving this problem and they both use the tweet content. in 4141] the goal is to infer the Point of Interest (POI) of the user like a restaurant, cinema, or club instead of the tweet's exact location. They investigate the relationship between the content of the tweet and feature classes, but the disadvantage of this method is the number of the classes. if the number of classes is huge, there may be thousands of Points of Interest (POI) for the user in one city that makes the prediction problem is very difficult. the authors in [4242, 45] use classification models to infer the location. in [4242] apply Neural Network (NN) to solve the tweet location prediction depending on the tweet content as an input.

Conclusion

According to the rapid usage of social media platforms, researchers and organizations are interested in the updated data that has been provided by users on these platforms. Recently researchers are very interested in location prediction, as it can be used in several applications like recommendation, and advertisements. In this paper, we present the research papers that discussing the location prediction problem using Twitter data, the Twitter platform maintained an additional property which is tracking users' location linking their latitude and longitude, this feature is called per-tweet geo-tagging. a Twitter microblog has millions of users who post tweets every day. One of the challenges of solving this problem is an abundance of information about users, and the user may write an inaccurate or wrong information, but there are features helps in the research point like the relationships of the users on Twitter, also they may add geo-location explicitly or implicitly in their text that helps in this research point. We review Tweet Location Prediction (TLP). Reviewing the four inputs for the prediction process; tweet content, the network of Twitter, the context of the tweet, and the hybrid methodology. the content of the tweet is the text of the posted tweets, the network of Twitter which is the relationship network or the user's followers and following, the context of the tweet which is the timestamp of the sending tweet, and the hybrid methodology means that the researchers use more than one input in their studies for location prediction. Reviewing each study use which of these inputs.

References

1. Oluwaseun Ajao, Jun Hong, and Weiru Liu, A survey of location inference techniques on Twitter, Journal of Information Science 41, 6 (2015), 855–864.
2. Shen-Shyang Ho, Mike Lieberman, Pu Wang, and Hanan Samet, Mining future spatiotemporal events and their sentiment from online news articles for location-aware recommendation system, In Proceedings of the First ACM SIGSPATIAL International

Workshop on Mobile Geographic Information Systems. Redondo Beach, California, pages 25–32.
3. Mingjun Wang and Matthew S. Gerber, Using Twitter for Next-Place Prediction, with an Application to Crime Prediction. IEEE Symposium Series on Computational Intelligence.
4. Q. Yuan, G. Cong, Z. Ma, A. Sun, and N. M. Thalmann, "Who, where, when and what: discover spatio-temporal topics for twitter users," in Proc. 19th ACM Int. Conf. on Knowledge Discovery and Data Mining, 2013, pp. 605–613.
5. Z. Cheng, J. Caverlee, and K. Lee, "You are where you tweet: a content-based approach to geo-locating twitter users," in Proc. 19th ACM Conf. on Information and Knowledge Management, 2010, pp. 759–768.
6. A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, "Mining user mobility features for next place prediction in location-based services," in Proc. 12th IEEE Int. Conf. on Data Mining, 2012, pp.1038–1043.
7. J. Ao, P. Zhang, and Y. Cao, "Estimating the locations of emergency events from twitter streams," in Proc. 2nd Int. Conf. on Information Technology and Quantitative Management, 2014, pp. 731–739.
8. Cheng, Z.Caverlee, J., & Lee, K, A content-driven framework for geolocating microblog users, ACM Transactions on Intelligent Systems and Technology, 4(1) 2:1–2:27.
9. Graham, M., Hale, S. A., & Gaffney, D, Where in the world are you? geolocation and language identification in twitter, The Professional Geographer, 66(4), 568–578.
10. Satyen Abrol and Latifur Khan, Tweethood: Agglomerative clustering on fuzzy k-closest friends with variable depth for location mining, In Proceedings of the IEEE 2nd International Conference on Social Computing (SocialCom'10). 153–160.
11. Brent Hecht, Lichan Hong, Bongwon Suh, and Ed H. Chi, Tweets from Justin Bieber's heart: The dynamics of the location field in user profiles, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 237–246.
12. J. McGee, J. A. Caverlee, and Z. Cheng, "A geographic study of tie strength in social media," in Proc. 20th ACM Conf. on Information and Knowledge Management, 2011, pp. 2333–2336
13. R. Compton, D. Jurgens, and D. Allen, "Geotagging one hundred million twitter accounts with total variation minimization," in Proc. IEEE Int. Conf. on Big Data, 2014, pp. 393–401.
14. M. Cha, Y. Gwon, and H. T. Kung, "Twitter geolocation andregional classification via sparse coding," in Proc. 9th Int. Conf.on Web and Social Media, 2015, pp. 582–585.
15. B. Wing and J. Baldridge, "Hierarchical discriminative classificationfor text-based geolocation," in Proc. Conf. on EmpiricalMethods in Natural Language Processing, 2014, pp. 336–348.
16. L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatialvariation in search engine queries," in Proc. 17th Int. Conf. onWorld Wide Web, 2008.
17. B. Han, P. Cook, and T. Baldwin, "Geolocation prediction in socialmedia data by finding location indicative words," in Proc. 24thInt. Conf. on Computational Linguistics: Technical Papers, 2012, pp.1045–1062.
18. Y. Yamaguchi, T. Amagasa, H. Kitagawa, and Y. Ikawa, "Onlineuser location inference exploiting spatiotemporal correlations insocial streams," in Proc. 23rd ACM Int. Conf. on Information andKnowledge Management, 2014, pp. 1139–1148.
19. S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge,"Supervised text-based geolocation using language models on anadaptive grid," in Proc. Joint Conf. on Empirical Methods in NaturalLanguage Processing and Computational Natural Language Learning,2012, pp. 1500–1510.

20. D. Rout, K. Bontcheva, D. Preot¸iuc-Pietro, and T. Cohn, “Where’s@wally?: a classification approach to geolocating users based ontheir social ties,” in Proc. 24th ACM Conf. on Hypertext and SocialMedia, 2013, pp. 11–20.
21. C. A. Davis Jr, G. L. Pappa, D. R. R. de Oliveira, and F. de L Arcanjo,“Inferring the location of twitter messages based on userrelationships,” TGIS, vol. 15, no. 6, pp. 735–751, 2011.
22. J. McGee, J. Caverlee, and Z. Cheng, “Location prediction insocial media based on tie strength,” in Proc. 22nd ACM Int. Conf.on Information and Knowledge Management, 2013, pp. 459–468.
23. D. Jurgens, “That’s what friends are for: Inferring location inonline social media platforms based on social relationships,” in Proc. 7th Int. Conf. on Weblogs and Social Media, 2013.
24. Y. Yamaguchi, T. Amagasa, and H. Kitagawa, “Landmark-baseduser location inference in social media,” in Proc. Conf. on Online Social Networks, 2013, pp. 223–234.
25. H. Efstathiades, D. Antoniades, G. Pallis, and M. D. Dikaiakos,“Identification of key locations based on online social networkactivity,” in Proc. of IEEE/ACM Int. Conf. on Advances in SocialNetworks Analysis and Mining, 2015, pp. 218–225.
26. A. Poulston, M. Stevenson, and K. Bontcheva, “Hyperlocal homelocation identification of twitter profiles,” in Proc. 28th ACM Conf.on Hypertext and Social Media, 2017, pp. 45–54.
27. J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing, “A latentvariable model for geographic lexical variation,” in Proc. Conf. onEmpirical Methods in Natural Language Processing, 2010, pp. 1277–1287.
28. K. Ren, S. Zhang, and H. Lin, “Where are you settling down:Geo-locating twitter users based on tweets and social networks,”in Proc. 8th Asia Information Retrieval Symposium, 2012, pp. 150–161.
29. R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, “Towardssocial user profiling: unified and discriminative influence modelfor inferring home locations,” in Proc. 18th ACM Int. Conf. onKnowledge Discovery and Data Mining, 2012, pp. 1023–1031.
30. R. Li, S. Wang, and K. C.-C. Chang, “Multiple location profilingfor users and relationships from social network and content,” PVLDB, vol. 5, no. 11, pp. 1603–1614, 2012.
31. A. Rahimi, D. Vu, T. Cohn, and T. Baldwin, “Exploiting text andnetwork context for geolocation of social media users,” in Proc.of the Conf. of the North American Chapter of the Association forComputational Linguistics: Human Language Technologies, 2015, pp.1362–1367.
32. 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.
33. J. Mahmud, J. Nichols, and C. Drews, “Home location identificationof twitter users,” ACM Transactions on Intelligent Systems andTechnology, vol. 5, no. 3, pp. 47:1–47:21, 2014.
34. Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma, “Unifyingtext, metadata, and user network representations with a neuralnetwork for geolocation prediction,” in Proc. 55th Annual Meetingof the Association for Computational Linguistics, 2017, pp. 1260–1272.4
35. H. Kwak, C. Lee, H. Park, and S. Moon, “What is twitter, a socialnetwork or a news media?” in Proc. 19th Int. Conf. on World WideWeb, 2010, pp. 591–600.
36. Oluwaseun Ajao, Jun Hong, and Weiru Liu. 2015. A survey of location inference techniques on Twitter. *Journal of Information Science* 41, 6 (2015), 855–864.

37. S. Kinsella, V. Murdock, and N. O'Hare, "I'm eating a sandwichin glasgow: modeling locations with tweets," in Proc. 3rd Int.CIKM Workshop on Search and Mining User-Generated Contents,2011, pp. 61–68.
38. K. Lee, R. K. Ganti, M. Srivatsa, and L. Liu, "When twittermeets foursquare: tweet location prediction using foursquare," inProc. 11th Int. Conf. on Mobile and Ubiquitous Systems: Computing,Networking and Services, 2014, pp. 198–207.
39. Z. Liu and Y. Huang, "Where are you tweeting?: A context anduser movement based approach," in Proc. 25th ACM Int. Conf. onInformation and Knowledge Management, 2016, pp. 1949–1952.
40. M. Hulden, M. Silfverberg, and J. Francom, "Kernel densityestimation for text-based geolocation," in Proc. 29th AAAI Conf.on Artificial Intelligence, 2015, pp. 145–150.
41. S. Hahmann, R. S. Purves, and D. Burghardt, "Twitter location(sometimes) matters: Exploring the relationship between georeferencedtweet content and nearby feature classes," Journal ofSpatial Information Science, vol. 2014, no. 9, pp. 1–36, 2014.
42. H. Iso, S. Wakamiya, and E. Aramaki, "Density estimation forgeolocation via convolutional mixture density network," CoRR, vol. abs/1705.02750, 2017.
43. A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends andfollowing them to where you are," in Proc. 5th Int. Conf. on Web Search and Web Data Mining, 2012, pp. 723–732.
44. W. Chong and E. Lim, "Tweet geolocation: Leveraging location,user and peer signals," in Proc. 26th ACM Conf. on Information andKnowledge Management, 2017, pp. 1279–1288.
45. B. Cao, F. Chen, D. Joshi, and P. S. Yu, "Inferring crowd-sourcedvenues for tweets," in 2015 IEEE Int. Conf. on Big Data, 2015, pp.639–648.
46. W. Li, P. Serdyukov, A. P. de Vries, C. Eickhoff, and M. Larson,"The where in the tweet," in Proc. 20th ACM Conf. on Information and Knowledge Management, 2011, pp. 2473–2476.