






# Geolocation Inference Using Twitter Data: A Case Study of COVID-19 in the Contiguous United States

Bingnan Li<sup>(✉)</sup> , Zi Chen , and Samsung Lim 

University of New South Wales, Sydney, NSW 2052, Australia  
{Bingnan.li, zi.chen1, s.lim}@unsw.edu.au

**Abstract.** Under the quarantine for the coronavirus disease 2019 (COVID-19) which has been spreading rapidly across the world since it was first identified in Wuhan City, China, in early December 2019, people are sharing their everyday life via social media more than ever before. Over the last decade, event-related information has been increasingly generated from Twitter by the growing popularity, and it is proved that the emergence and evolution of events can be timely monitored and analyzed on the basis of this platform. Geographic information plays a crucial role in mining social media data, however, only about 2% of tweets hold accurate geographic information due to the operational complexity and privacy concerns. To overcome the geo-tagging restriction, finding effective geolocation inference methods is currently one of the main topics in this research field. Geographic information plays an important role in analyzing and monitoring the spread of an epidemic disease. In this study, we constructed a method of geolocation inference based on the whole potential location-related metadata of tweets. A crude form of geographic coordinate information can be obtained from every tweet's bounding box, while location-related information can be mined from the textual content, user location and place labels via Named Entity Recognition (NER) techniques. Three coordinate datasets of the United States counties are built and used as the coordinate references. Models with different data sources have been employed to predict the geolocations of the tweets related to COVID-19 in the contiguous United States. Results show that the models with four data sources, namely textual content, user location, place labels and bounding box of place, with Digital Boundary's Average (DBA), perform better than other models. When the area threshold of the bounding box is set to 10,000 km<sup>2</sup>, the best model can successfully predict the geolocation of 90.8% of COVID-19 related tweets with the mean error distance of 4.824 km and the median error distance of 3.233 km. It is concluded that the proposed method enhances the granularity of geographic information of tweets and makes the surveillance of COVID-19 effective and efficient.

**Keywords:** COVID-19 · Social media · Geolocation inference · Twitter data · Data mining

## 1 Introduction

In December 2019, the initial cases of pneumonia associated with a novel coronavirus occurred in Wuhan City, China [1]. However, measures to control the spread of the virus

were not implemented effectively to keep its spread within China [2]. Since then, the coronavirus disease 2019 (COVID-19) has been rapidly spreading around the world, causing tens of millions of cases in more than 160 countries [1]. As of August 17<sup>th</sup>, 2020, almost 22 million (21,852,024) cases have been recorded, including 773,586 deaths where 25.48% (5,567,765) of those cases occurred within the United States, including 173,139 deaths according to the worldometer coronavirus pandemic tracker [3]. Therefore, an overarching objective of this study is to contribute to the identification of spatio-temporal patterns of the COVID-19 pandemic with a particular interest in the United States.

Over the past decade, the Internet has helped revolutionize every aspect of people's lives, and it is not only a source to get information, but also a platform to disseminate personal information [4, 5]. In addition, the development of mobile devices made it easier to send digital information (e.g., texts, location labels, and pictures). At the same time, social media platforms have experienced a tremendous and profound reform. Twitter and Facebook mainly provide basic services, but other types of social media are being used to connect online for different reasons, such as location-based services (e.g., Foursquare and Whrrl), media sharing services (e.g., Instagram, Snapchat, and Flickr), as well as other types of services (e.g., Quora, Medium, and LinkedIn). Users can establish online friendships based on mutual interests and share their everyday life with each other.

Supported by previous studies [4, 6–8], Twitter outshines other platforms in regard to social network analysis and event detection because of not only its excellent design, but also its vast user base of different age groups. According to the most up-to-date Twitter statistics for 2020, its monthly active users are around 330 million, which accounts for 23% of the Internet population, and about 500 million tweets are posted every single day [9]. Compared with Instagram and Snapchat regarding the demographics, Twitter is widely used by people of different ages and nearly 63% of them age between 35 and 65 [10]. The large quantity of user-generated contents is employed for data mining in various research areas [4]. Tweets with accurate geographic information can provide significant benefits to event response and monitoring, hence those without geographic information become useless unless geolocation inference is applicable. Accurate prediction of tweets' geolocation can effectively benefit the response and rescue in emergency events [11].

The development of Global Positioning System (GPS)-enabled mobile devices enables users to share and track their locations with accurate geographical coordinates. However, due to the operational complexity and privacy concerns, most users do not turn this function on [12]. As Laylavi et al. [13] illustrated, the percentage of tweets with geo-tags account for only 2%, which severely limits the development of associated applications. Therefore, accurate geolocation inference of tweets has become an urgent problem in this research field.

Nowadays, disease-related information is increasingly shared in real time through Twitter, while timely data with spatial and temporal information plays a significant role in surveillance of an epidemic disease [14, 15]. Every single tweet has its own metadata, which includes its creation time, but under most circumstances, does not contain its created geographical coordinates, hence geolocation inference of tweets is still a critical issue. Real-time data without any geographic information can be almost meaningless

for emergency response and surveillance of an epidemic disease. Thus, this study aims to develop novel methods to predict geolocation of tweets based on their own metadata.

In this study, models based on multiple attributes of the tweet's metadata are built to predict the non-geotagged tweets' geolocation. Attributes of textual content, user location, place labels, and bounding box are fully used during the modelling process. The dataset used in this study was collected between the 10<sup>th</sup> and 30<sup>th</sup> of June 2020. During this time, the United States (US) was suffering a severe effect of the COVID-19 pandemic. The development of technologies, including Natural Language Processing (NLP) and Named Entity Recognition (NER) make it easier to extract location entities from textual data.

The main contributions of this paper are summed up as the following two points: (1) Exploring potential location-related attributes of the tweet's metadata and extracting location entities via NER techniques; (2) Three geographic coordinate datasets of counties are used to predict geolocation and the proposed models are built according to different priorities of location-related attributes.

The rest of this paper is structured as follows. Section 2 describes a literature review of relevant research. Section 3 presents a brief introduction of Twitter data's structure. Detailed explanation of the proposed models is described in Sect. 4. A case study of the COVID-19 in the contiguous US based on the models mentioned in Sect. 4 is illustrated in Sect. 5. The paper finally concludes in Sect. 6.

## 2 Related Works

Users sometimes add geo-information in their tweets, but in most cases, it is still not that complete or accurate. Therefore, various methods and algorithms from other fields are being used in the field of geolocation inference. With the development of technologies such as machine learning, deep learning, NLP as well as Geographical Information Systems (GIS), much more methods have made breakthroughs in this research field [16]. However, different from formal articles which are well written and grammatically correct, social media messages always contain informal elements, e.g., acronyms, emojis, hashtags and even typos, which is often attributed to the limit of character count and the use of mobile devices.

In the past few years, many studies of geolocation inference based on Twitter data have been published [16]. Ajao et al. [15] reviewed previous research related to geolocation inference of tweets, and summarized relevant methods and evaluation metrics. In the work of Cheng et al. [17], they discovered merely 20% of Twitter users in the US prefer to show cities where they live in their user profiles, and only 5% of them provide geographical coordinates information. The study of Hecht et al. [18] illustrated that even though self-described addresses are shown in their profiles, some of them are not accurate or valid, and geo-tagged tweets account for merely 0.77% of the whole. From the study of Ryoo et al. [19], the percentage of tweets with geographic information is only about 0.4%. Bartosz et al. [20] as well as Priedhorsky et al. [21] showed the similar percentages in their studies. More importantly, geolocation inference of social media data is the basis of other relevant studies. Consequently, further research in this area is needed.

When tweets are posted, some places information in the textual content enables us to understand them better. Textual content is used to predict the geolocation of tweets in the studies of Cheng et al. [22], Chandra et al. [23] as well as Chang et al. [24]. However, Ikawa et al. [25] described that some users always mention places that are not exactly where they are. In the study of Abrol et al. [26], they researched the social network relationships among their online friends. Backstrom et al. [27] and Bouillot et al. [28] described that geolocation inference of tweets can be achieved by the user profile in their studies.

NLP techniques enable various methods and algorithms of this field to be used in information extraction and geolocation inference. Techniques of NER and part-of-speech tagging (POS) have been introduced in the research of Lingad et al. [29]. Li et al. [30] introduced methods of machine learning and probabilistic to geolocation inference. Takhteyev et al. [31] used gazetteers and location databases in their research. In the study of Huang et al. [12], deep learning models are used to predict geolocation of Twitter data. Previous studies have obtained a great achievement in this field and have the potential to pursue more accurate results of geolocation prediction [32].

Most studies conducted on geolocation inference of tweets focus on either textual content or other location-related attributes. However, this research aims to implement all feasible combinations of potential attributes related to location to predict the geolocation of tweets.

### 3 Structure of Twitter Data

Twitter was released in March 2006 and now has about 330 million active users per month. Tweets can be posted by users via this platform. In its early days, every tweet can contain up to 140 characters, but the length of it was doubled in 2017 [33]. This increase provided users more space to express their ideas and saved more time of text compression than before. Every tweet's metadata contains a wealth of information about itself, while it is only visible to developers, not common users. Twitter data can be collected based on Twitter application programming interfaces (APIs) and stored with the format of JavaScript Object Notation (JSON). JSON format is lightweight and easy for both human beings and machines to understand and use. A JSON object contains a key/value pair and is normally enclosed in a pair of curly braces [34]. The structure of Twitter data consists of several objects, including tweet object, user object, coordinates object, place object, and bounding box object, which are all encoded in JSON format. For every tweet, the metadata can tell us its username, textual content, unique identification (ID), created time, and occasionally geographic details of where it was posted. In general, every tweet's metadata contains more than 150 attributes, while only spatio-temporal information related attributes (shown in Fig. 1) are taken into consideration in our research.

Figure 1 shows the spatio-temporal information related attributes in a tweet's metadata. The attribute of "location" is an element of the user object and is defined by user himself/herself, therefore, it can be a location that does not exist in the real world or cannot be recognized by computers. Another one is "geo\_enabled", which means if the current user can attach geographic data or not. This attribute is very important for location-related studies, although it does not contain any essential geographic information.

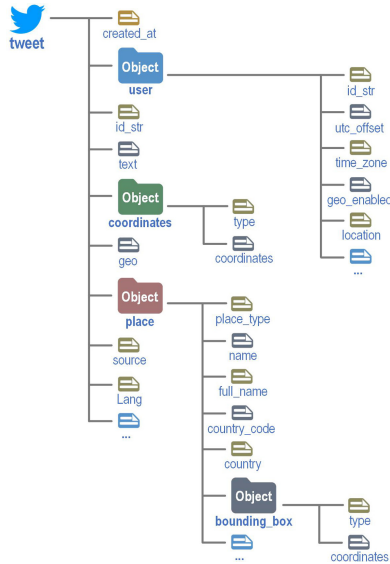


Fig. 1. Spatio-temporal attributes of a tweet’s metadata.

Both attributes of “coordinates” and “geo” represent the specific longitude and latitude of the tweet’s location, as a collection in the form [longitude, latitude]. However, “geo” has been deprecated according to the twitter official document, hence we used the attribute of “coordinates” to acquire accurate geo coordinates of tweets [35].

Place object contains various location-related attributes. The attribute “place\_type” represents the type of location of this place and it has five values to choose from. Table 1 shows five values of attribute of “place\_type” and statistics of our research dataset. For POI, it represents the specific location of a place, e.g., Washington Square Park, while the other four values stand for a certain area. Due to the large regional extent of city, admin, and country, we used data from only POI and neighborhood. Attributes of “name” and “full\_name” are two ways to describe the place’s names. While “country\_code” and “country” provide the short code and exact name of the country of the place. The attribute of “bounding\_box” is four lon/lat pairs of each corner of a box that contains the place [35].

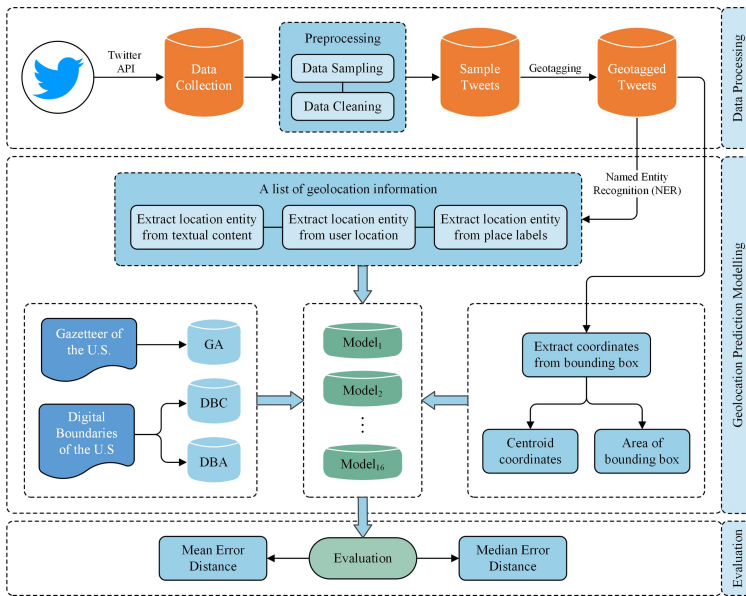
## 4 Proposed Method

Figure 2 plots the workflow to illustrate the architecture of the proposed method of this research. This method is generally divided into three modules. In the first module, real time tweets within a bounding box are collected. Tweets data are initially stored into text files and then read based on JSON format. Then the data enters the preprocessing and geotagging stage, after which a dataset with geo-tagged tweets is created. In the second module, location entities are extracted from textual content, user location and place labels via NER techniques. Combining geometric properties of the place’s bounding box, as well as coordinate datasets of gazetteers and digital boundaries of the US, all

**Table 1.** Typical values and statistics of “place\_type” attribute.

Category	Amount	Percentage	Example
POI	119,655	0.96%	Washington Square Park
Neighborhood	25,183	0.20%	Downtown Jacksonville, FL
City	10,301,683	82.98%	Los Angeles, CA
Admin	1,942,596	15.65%	California, USA
Country	26,105	0.21%	Canada

these data are fed into 16 models to predict tweets’ geolocation. Finally, predicted results are evaluated by mean error distance (MED) and median error distance (MEDD).



**Fig. 2.** Workflow of geolocation inference of tweets [36].

### 4.1 Data Collection

Twitter data can be gathered from both business companies and Twitter API which is available free of charge. As for commercial purchases, the companies can provide both historical and real time tweets from all over the world, but the price is very high. Twitter API can help collect tweets freely, but only real time tweets within the specific bounding box can be collected. Therefore, it normally takes several months to collect the whole research data using Twitter API. In this study, data collection was done via Twitter API,

and it was implemented by the *tweepy* library of python [13, 37]. The data were collected from June 10<sup>th</sup> to June 30<sup>th</sup>, 2020 in the contiguous US during the COVID-19 pandemic. During this period, 12,408,538 unduplicated tweets were collected and stored into local text files. Only tweets located in the area of longitudes from 66°W to 125°W and latitudes from 24°N to 49°N are collected, as shown in Fig. 3. While within the bounding box, some tweets from Canada, Mexico, and the Bahamas were also included, but excluded in this research.

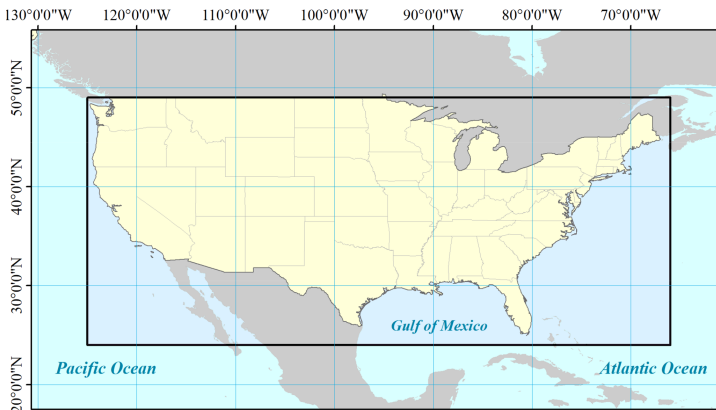


Fig. 3. Area of data collection.

## 4.2 Data Preprocessing

**Data Cleaning.** In the textual content of every tweet, it often contains noises, including hashtags, mentions, emojis and Uniform Resource Locator (URL) links, hence preprocessing operation is necessary. In this step, we used regular expressions to process textual data. A regular expression is a pattern that attempts to match with input text and can be implemented by python *re* library [38]. URL links started with “https://” and “http://” were removed from the textual content since they do not contain any location related information. We replaced unnecessary punctuation marks into a space, and consecutive spaces into one. Marks of user mentions, hashtags, non-English letters as well as stop words were all deleted [37]. As for the user location, it can be modified by users at will, thus the information was processed in the same manner.

**Data Sampling.** A workflow was plotted to illustrate how useless tweets are filtered out and generated a new dataset. The dataset was mainly processed via the python *pandas* library. Firstly, the method of “drop\_duplicate” is employed to delete duplicated tweets from the dataset. The attribute of “lang” indicates the language used by every tweet, and only English tweets are kept in our study. As noted above, tweets posted outside the contiguous US are also removed from the dataset.

Another problem is that many tweets are meaningless to this study, such as those posted by advertisers or spambots. This kind of tweets is mainly posted by computers, therefore, only tweets posted by mobile devices (e.g., iPhone, Android, iPad, and Instagram) are kept, and the attribute of “source” was used to implement this function [13, 37]. Then tweets without geo-tags were filtered out and implemented by the “coordinates” attribute. Finally, the COVID-19 related tweets were extracted by using the keywords to match the “text” attribute of every tweet. We introduced Term Frequency-Inverse Document Frequency (TF-IDF) to get keywords from news articles about the COVID-19 pandemic in the US, and TF-IDF score helped us extract keywords from the related articles [39].

Supported by recent studies [1, 40, 41] and TF-IDF techniques, we used the following keywords: “corona”, “coronavirus”, “covid”, “covid-19”, “ncov”, “sarscov2”, “ncov2019” and “2019ncov” to extract COVID-19 related tweets. Through data sampling, 3,600 corresponding tweets were retrieved from the Twitter dataset. Figure 4 shows the whole data sampling process.

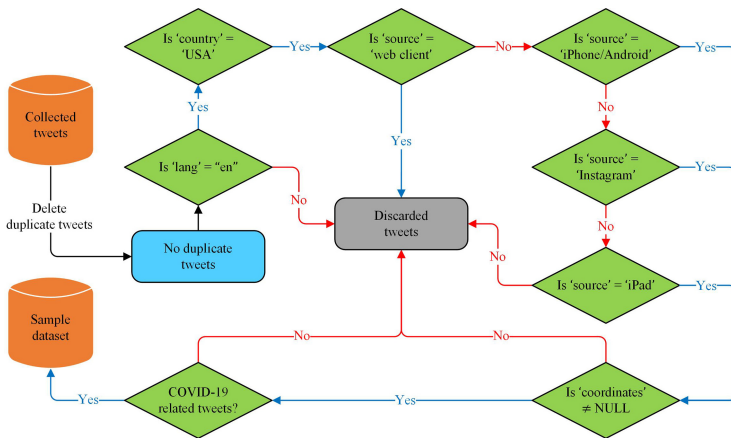


Fig. 4. Flowchart of data sampling [36].

### 4.3 Location Information Extraction

**Named Entity Recognition.** NER can be used to recognize and classify different types of entities (e.g., location names, person names, and organizations) from unstructured texts. It has been extensively studied in the last few years in machine learning and NLP. While it does not work well on informal texts like tweets since it is usually built on the basis of formal texts [29]. As for this technique, it can help to answer many real-world questions, such as: does a tweet contain the name of a person or does the tweet provide a person’s current location? In this study, we use NER to identify locations from textual content, user location, and place labels of the tweet based on Stanford NER, spaCy, and Natural Language Toolkit (NLTK). After testing all tools in real tweet dataset, spaCy



showed a much better performance than the other two tools, therefore, spaCy is used to identify location-related information from tweets in this research.

**Bounding Box.** The bounding box is a specified 4-sided geographic area and matching the tweet’s location falling into the area. Unlike other location related geographical metadata, the bounding box contains the accurate lon-lat coordinates of the four points enclosing the place. Due to different types of places, bounding box has different areas. For instance, four points of a bounding box are  $Point_1 = (\lambda_1, \varphi_1)$ ,  $Point_2 = (\lambda_2, \varphi_1)$ ,  $Point_3 = (\lambda_2, \varphi_2)$  and  $Point_4 = (\lambda_1, \varphi_2)$ , then Eq. 1 can be used to calculate the area of this bounding box.

$$S = R^2 \cdot |(\lambda_2 - \lambda_1) \cdot (\sin \varphi_2 - \sin \varphi_1)| \quad (1)$$

where  $R$  refers to the earth radius;  $\lambda_1$  and  $\lambda_2$  represent the longitudes of the bounding box, and  $\varphi_1$  and  $\varphi_2$  refer to the latitudes of the bounding box.

Equation 1 can be used to calculate the size of the bounding box. The bounding box’s centroid can be reckoned as the predicted location of a tweet, therefore, if the bounding box’s area is smaller, it can provide a relatively more accurate prediction. For city, admin and country, the bounding box is too large to be used to predict the geolocation.

#### 4.4 Modelling

The location-related information is obtained from the four sources: textual content, location of user profile, place labels, and bounding box. Three coordinate datasets of counties are constructed based on gazetteers and digital boundaries of the US.

**United States Gazetteers.** The national gazetteers of the US were used as the data source and called GA in this study. It is a dataset including county’s names and information related to geography in the US. This data is provided by the United States Census Bureau, and researchers can download it for free [42]. There are totally ten fields in the dataset, and some of them are displayed in Table 2. The field of “NAME” can provide duplicate names, but they locate in different states which means they have different values of “USPS”. Fields of “INTPTLAT” and “INTPTLON”, respectively, refer to latitude and longitude of the specific county.

**Table 2.** Data fields of US gazetteers.

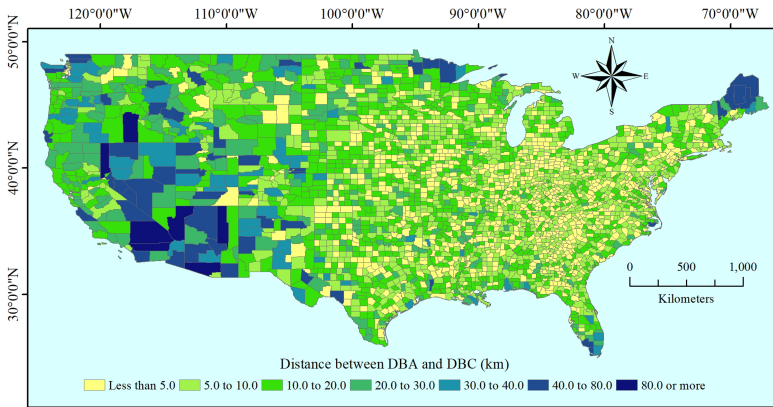
Field	Description
USPS	United States Postal Service state abbreviation
GEOID	Unique geographic identifier for each feature
NAME	Name of the feature
INTPTLAT	Latitude of the feature in decimal degrees
INTPTLON	Longitude of the feature in decimal degrees

**Digital Boundaries of the United States.** Digital boundaries of the US are in the format of Environmental Systems Research Institute (ESRI) *lpk*. This group layer can be freely downloaded from the website of ESRI and presents counties of the US in the 50 states, the District of Columbia, and Puerto Rico. The detailed datasets are represented as polygons with over 40 fields [43].

In this paper, we only used digital boundaries of US counties due to the coarse granularity of location inference based on the city and state level. In order to obtain geographic coordinates of each county, we developed two ways to compute them and named them Digital Boundary’s Centroid (DBC) and Digital Boundary’s Average (DBA). DBC is calculated based on geometric properties of every county’s polygon, and the value can be calculated by the centroid of the polygon. On the other hand, DBA is calculated by tweets falling into the county’s polygon and the value can be calculated by their average latitude and longitude. For instance, suppose there are  $m$  counties in the contiguous US which are  $County_1, \dots, County_j, \dots, County_m$  and  $P\_tweet_1 = (\lambda_1, \varphi_1), \dots, P\_tweet_i = (\lambda_i, \varphi_i), \dots, P\_tweet_n = (\lambda_n, \varphi_n)$  are geographic coordinates of  $n$  tweets located in  $County_j$ , then the predicted coordinates of  $County_j(P\_county_j)$  can be calculated by Eq. 2. This method can help compute the average longitude and latitude of geotagged tweets falling into the county’s polygon.

$$P\_county_j = (\bar{\lambda}, \bar{\varphi}) = \left( \frac{\sum_{i=1}^n \lambda_i}{n}, \frac{\sum_{i=1}^n \varphi_i}{n} \right) \tag{2}$$

After calculating all polygons’ coordinates based on DBA and DBC, Fig. 5 shows the distribution of distances between DBA and DBC of counties in the contiguous US. This figure illustrates that the distance difference is less than 20 km in most countries, especially for the smaller ones, while for some larger counties in the west and northeast corner, the difference is about 40 km or more. Smaller distance difference means two predicted methods are close to each other. When the distance difference is larger, the better method of coordinates prediction can achieve a better performance.



**Fig. 5.** Distribution of distance difference of counties in the contiguous U.S.

**Modelling.** As demonstrated in Fig. 2, the model is on the basis of four location-related attributes of the tweet’s metadata: textual content (T), user location (U), place label (P) and bounding box (B). Location entities are extracted from T, U, and P by NER techniques, and then query them through coordinate datasets of GA, DBC, and DBA.

Equation 3 illustrates how the three predicted matrices are computed. The value will be stored as “null” if there is no county found based on NER. When we use NER to query the specific county’s name, sometimes several results will be found since there are duplicate names of different counties. Therefore, the distance between the predicted point and centroid of the tweet’s bounding box should be computed first, if it is within the specific threshold range, the predicted point can be reckoned as a valid result, otherwise will be discarded.

$$\begin{matrix} \left[ \begin{array}{ccc} \text{Text}_1 & \text{UserLoc}_1 & \text{Place}_1 \\ \vdots & \vdots & \vdots \\ \text{Text}_i & \text{UserLoc}_i & \text{Place}_i \\ \vdots & \vdots & \vdots \\ \text{Text}_n & \text{UserLoc}_n & \text{Place}_n \end{array} \right] \end{matrix} \xrightarrow{\text{NER} \left\{ \begin{array}{l} \text{GA} \\ \text{DBC} \\ \text{DBA} \end{array} \right\}} \left\{ \begin{array}{l} \left[ \begin{array}{ccc} T_{GA_1} & U_{GA_1} & P_{GA_1} \\ \vdots & \vdots & \vdots \\ T_{GA_i} & U_{GA_i} & P_{GA_i} \\ \vdots & \vdots & \vdots \\ T_{GA_n} & U_{GA_n} & P_{GA_n} \end{array} \right] \\ \left[ \begin{array}{ccc} T_{DBC_1} & U_{DBC_1} & P_{DBC_1} \\ \vdots & \vdots & \vdots \\ T_{DBC_i} & U_{DBC_i} & P_{DBC_i} \\ \vdots & \vdots & \vdots \\ T_{DBC_n} & U_{DBC_n} & P_{DBC_n} \end{array} \right] \\ \left[ \begin{array}{ccc} T_{DBA_1} & U_{DBA_1} & P_{DBA_1} \\ \vdots & \vdots & \vdots \\ T_{DBA_i} & U_{DBA_i} & P_{DBA_i} \\ \vdots & \vdots & \vdots \\ T_{DBA_n} & U_{DBA_n} & P_{DBA_n} \end{array} \right] \end{array} \right. \quad (3)$$

where  $\text{Text}_i$ ,  $\text{UserLoc}_i$ , and  $\text{Place}_i$  are respectively textual content, user location, and place label of a tweet;  $T_{GA_i}$ ,  $U_{GA_i}$ , and  $P_{GA_i}$  are predicted coordinates corresponding to  $\text{Text}_i$ ,  $\text{UserLoc}_i$ , and  $\text{Place}_i$ , respectively, based on GA;  $T_{DBC_i}$ ,  $U_{DBC_i}$ , and  $P_{DBC_i}$  are predicted coordinates corresponding to  $\text{Text}_i$ ,  $\text{UserLoc}_i$ , and  $\text{Place}_i$ , respectively, based on DBC;  $T_{DBA_i}$ ,  $U_{DBA_i}$ , and  $P_{DBA_i}$  are predicted coordinates corresponding to  $\text{Text}_i$ ,  $\text{UserLoc}_i$ , and  $\text{Place}_i$ , respectively, based on DBA;

$T_{GA_i}$ ,  $U_{GA_i}$ , and  $P_{GA_i}$  can be “null” if corresponding counties are not found in GA;  $T_{DBC_i}$ ,  $U_{DBC_i}$ , and  $P_{DBC_i}$  can be “null” if corresponding counties are not found in DBC;  $T_{DBA_i}$ ,  $U_{DBA_i}$ , and  $P_{DBA_i}$  can be “null” if corresponding counties are not found in DBC.

Equation 4 shows how the area and centroid's coordinates are computed by the tweet's bounding box.

$$\begin{bmatrix} BBox_1 \\ \vdots \\ BBox_i \\ \vdots \\ BBox_n \end{bmatrix} \begin{cases} Area \\ Centroid \end{cases} \Rightarrow \begin{bmatrix} B_{AREA_1} & B_{CEN_1} \\ \vdots & \vdots \\ B_{AREA_i} & B_{CEN_i} \\ \vdots & \vdots \\ B_{AREA_n} & B_{CEN_n} \end{bmatrix} \quad (4)$$

where  $BBox_i$  is the tweet's bounding box;  $B_{AREA_i}$  and  $B_{CEN_i}$  are the area and centroid's lon-lat coordinates of  $BBox_i$ , respectively.

Because every tweet has the attribute of bounding box, every model in our study contains this attribute and is placed in the last position. UPTB is one model and designed according to the order of U, P, T, and B. Figure 6 illustrates a flow diagram of how UPTB works based on GA.

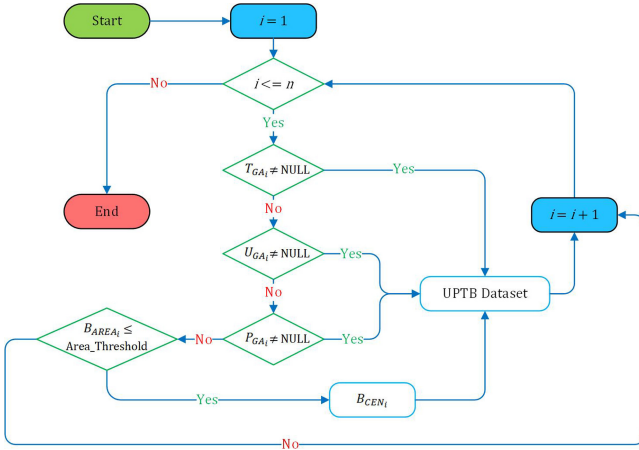


Fig. 6. Working principle of UPTB based on GA.

As shown in this flow chart,  $n$  elements are traversed in the outermost. Then, if  $T_{GA_i}$  is not “null”, it is passed directly to the UPTB dataset, otherwise indicated by  $U_{GA_i}$ . If  $U_{GA_i}$  is not “null”, it is passed directly to the UPTB dataset, otherwise indicated by  $P_{GA_i}$ . If  $P_{GA_i}$  is not “null”, it is passed directly to the UPTB dataset, otherwise indicated by  $B_{AREA_i}$ . If the value of  $B_{AREA_i}$  is not more than the  $Area\_Threshold$ ,  $B_{CEN_i}$  is passed to the UPTB dataset and then a new loop starts, otherwise a new loop starts directly and the final result will be set as “null”. When the predicted result is “null”, it means geo coordinates of this tweet cannot be predicted based on this model.

The other models are implemented with the same mechanism. That is, six models (i.e., TUPB, TPUB, UTPB, UPTB, PUTB, and PTUB) contain four parameters, six models (i.e., TUB, TPB, UTB, UPB, PTB, and PUB) contains three parameters, three

models (i.e., TB, UB, and PB) contain two parameters and one model (B) contains merely one parameter. A total of 16 models are implemented in this study.

## 5 Experimental Results

We applied models mentioned in Sect. 4 to the sample dataset and evaluated their performance based on different metrics.

### 5.1 Research Data

Table 3 shows the Twitter dataset that we used in this study. We collected these tweets from 10<sup>th</sup> to 30<sup>th</sup> of June 2020 in the contiguous US during the COVID-19 pandemic spreading around the world. The total number of collected tweets are 12.4 million and tweets with geo-tags account for 6%. Only geo-tagged tweets related to COVID-19 are applied to the models described in Sect. 4, and the number is 3,600.

**Table 3.** Statistical information about Twitter dataset.

Item	Content
Database size	61.0 GB
Date of data gathering	2020.06.10–2020.06.30
Total number of tweets	12,415,222 tweets
Total number of unique tweets	12,408,538 tweets
Total number of tweets from mobile devices	11,475,982 tweets
Total number of tweets from Instagram	401,610 (3.24%)
Total number of English tweets	10,056,767 tweets
Number of geo-tagged tweets	758,946 tweets (6.11%)
Number of geo-tagged tweets related to COVID-19	3,600 tweets (0.029%)

As shown in Table 3, geo-tagged tweets account for 6.11% of the total Twitter dataset. These tweets were extracted, then plotted with digital boundaries of the contiguous US. Figure 7(a) [42] shows the population distribution of the contiguous US counties (i.e., people per square kilometer of 2018), and Fig. 7(b) shows the geo-tagged tweets distribution based on the contiguous US counties (i.e., geotagged tweets per 1,000 square kilometers between June 10<sup>th</sup> and June 30<sup>th</sup>, 2020).

In statistics, the Pearson’s correlation coefficient (PCC) is a statistic that measures linear correlation between two variables. The value range of PCC is between  $-1$  and  $1$ , and the higher the value, the better the positive linear correction. Equation 5 shows how to calculate PCC based on two paired data  $\{(x_1, y_1), \dots (x_i, y_i) \dots (x_n, y_n)\}$  consisting of  $n$  pairs.

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (5)$$

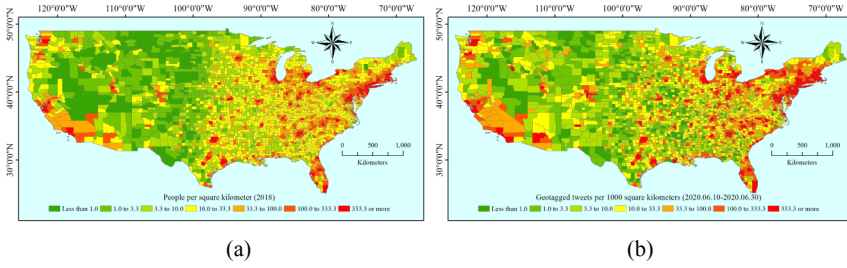


Fig. 7. Population and tweets distribution in the contiguous U.S.

In this study,  $x_i$  means people per square kilometer in every county, and  $y_i$  means tweets per 1,000 square kilometers in every county. PCC of the two variables in this study is 0.88, which indicates a strong positive correlation. Figure 7 also shows that population distribution and tweets distribution have a high correlation, hence we can detect real world events based on geo-tagged tweets or tweets with predicted geolocation.

## 5.2 Evaluation Metrics

Models' performance can be evaluated by the distance between the predicated geolocation and the real geolocation of a tweet. The actual distance between two points on the earth's surface can be calculated by the great circle distance. For instance, the great circle distance of two points,  $p_1 = (\lambda_1, \varphi_1)$  and  $p_2 = (\lambda_2, \varphi_2)$ , can be calculated by Eq. 6.

$$Dist(p_1, p_2) = 2R \arcsin \left( \sqrt{\sin^2 \left( \frac{\varphi_2 - \varphi_1}{2} \right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2 \left( \frac{\lambda_2 - \lambda_1}{2} \right)} \right) \quad (6)$$

where  $R$  is the earth radius;  $\lambda_1$  and  $\lambda_2$  refer to the longitudes of points, and  $\varphi_1$  and  $\varphi_2$  refer to the latitudes of points.

Mean error distance (MED) and median error distance (MDED) are two metrics to evaluate models in our research, and are implemented by Eq. 7 and Eq. 8, respectively.

$$MED = \frac{1}{n_{tweets}} \sum_{i=1}^{n_{tweets}} Dist(\hat{p}_i, p_i) \quad (7)$$

$$MDED = median_{i=1}^{n_{tweets}} Dist(\hat{p}_i, p_i) \quad (8)$$

where  $\hat{p}$  represents the predicted geolocation and  $p_i$  refers to the real geolocation of a tweet.

The tweet's metadata indicates that the value of bounding box is always not null, therefore, it can be used to predict the geo coordinates of the tweet. But its area varies a lot among different tweets and the error distance can be affected dramatically. Figure 8 shows the variation of MED and its percentage based on different area thresholds of the bounding box. For example, if the area threshold is set to 1,000,000 km<sup>2</sup>, almost 100% of tweets can predict the geo coordinates, but the MED is almost 25 km. When the area

threshold is set to 5,000 km<sup>2</sup>, almost 90% of tweets can be valid to predict, and the MED improves to 5 km. As shown in Fig. 8, when the area threshold is set to 5,000 km<sup>2</sup> and 10,000 km<sup>2</sup>, the MED and percentage can achieve a relatively better performance, thus the following experiments were conducted by these two values.

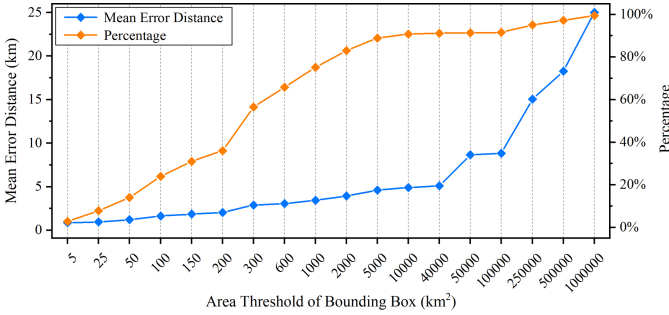


Fig. 8. MED and percentage of different area thresholds.

Sometimes users mention some other location names rather than the place where tweets are posted. But in most cases, users are more likely to be within or around the place. In addition to this, there often exist duplicate names of different counties in the datasets of GA, DBC, and DBA. Therefore, sometimes several counties were extracted by NER from a tweet. To resolve this issue, we only focus on the predicted location in the bounding box and the distance between it and the bounding box’s centroid is within the specific range. In this study, we chose the distance threshold from 1 km to 10 km. For example, when the distance threshold is set to 6 km, only the first result with distance of predicted point and bounding box’s centroid no more than 6 km has been kept. Figure 9 shows MED of TUPB in three datasets with different distance thresholds, when the area threshold is set to 5,000 km<sup>2</sup>. As illustrated in this figure, the distance threshold has no obvious effect on datasets of DBC and GA, but it has a significant impact on DBA. When distance is set to 6 km, the MED is lowest, hence we chose 6 km as the distance threshold in this study.

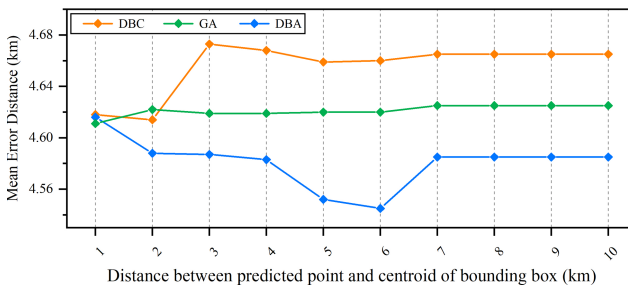


Fig. 9. MED based on different distance thresholds.

### 5.3 Results

Combining models mentioned in Sect. 4, three coordinate datasets of counties of the US and Eq. 6, MED ( $B_{AREA_i} \leq 5,000 \text{ km}^2$  and  $B_{AREA_i} \leq 10,000 \text{ km}^2$ ) can be computed and shown in Table 4 and Fig. 10. When the area threshold is set to  $5,000 \text{ km}^2$ , about 88.9% of sample tweets are successfully predicted, and the percentage has improved to 90.8% when the area threshold is set to  $10,000 \text{ km}^2$ .

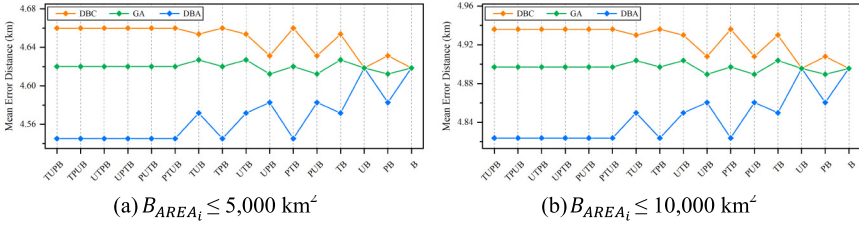
**Table 4.** MED of models based on two area thresholds.

Models	MED ( $B_{AREA_i} \leq 5,000 \text{ km}^2$ )			MED ( $B_{AREA_i} \leq 10,000 \text{ km}^2$ )		
	DBC	DBA	GA	DBC	DBA	GA
TUPB	4.660	4.545	4.620	4.936	4.824	4.897
TPUB	4.660	4.545	4.620	4.936	4.824	4.897
UTPB	4.660	4.545	4.620	4.936	4.824	4.897
UPTB	4.660	4.545	4.620	4.936	4.824	4.897
PUTB	4.660	4.545	4.620	4.936	4.824	4.897
PTUB	4.660	4.545	4.620	4.936	4.824	4.897
TUB	4.654	4.572	4.627	4.930	4.850	4.904
TPB	4.660	4.545	4.620	4.936	4.824	4.897
UTB	4.654	4.572	4.627	4.930	4.850	4.904
UPB	4.631	4.583	4.612	4.908	4.860	4.890
PTB	4.660	4.545	4.620	4.936	4.824	4.897
PUB	4.631	4.583	4.612	4.908	4.860	4.890
TB	4.654	4.572	4.627	4.930	4.850	4.904
UB	4.619	4.619	4.619	4.896	4.896	4.896
PB	4.631	4.583	4.612	4.908	4.860	4.890
B	4.619	4.619	4.619	4.896	4.896	4.896

From Fig. 10(a), one can see that GA has a relatively steady performance for all models, and all values of MED are around 4.62 km. DBC has a similar performance to GA, but the models with four sources have relatively worse performances compared to other models. While DBA has a clear trend of variation based on different models, the models with three or four sources have better performances than other models. Figure 10(b) shows MED's variation with respect to DBC, GA, and DBA based on 16 models when the area threshold of the bounding box is set to  $10,000 \text{ km}^2$ . One can see that three lines from Fig. 10(b) have similar trend patterns as those from Fig. 10(a).

There often exist some abnormal values in the dataset, and these values can pose a significant impact on the mean value, hence the median value can reduce the impact of abnormal values. Table 5 and Fig. 11 show the median error distance with the bounding box's area of  $5,000 \text{ km}^2$  and  $10,000 \text{ km}^2$ .





**Fig. 10.** MED of models based on two area thresholds.

**Table 5.** MDED of models based on two area thresholds.

Models	MDED ( $B_{AREA_i} \leq 5,000 \text{ km}^2$ )			MDED ( $B_{AREA_i} \leq 10,000 \text{ km}^2$ )		
	DBC	DBA	GA	DBC	DBA	GA
TUPB	3.239	3.095	3.245	3.327	3.233	3.373
TPUB	3.239	3.095	3.245	3.327	3.233	3.373
UTPB	3.239	3.095	3.245	3.327	3.233	3.373
UPTB	3.239	3.095	3.245	3.327	3.233	3.373
PUTB	3.239	3.095	3.245	3.327	3.233	3.373
PTUB	3.239	3.095	3.245	3.327	3.233	3.373
TUB	3.183	3.135	3.244	3.280	3.243	3.367
TPB	3.239	3.095	3.245	3.327	3.233	3.373
UTB	3.183	3.135	3.244	3.280	3.243	3.367
UPB	3.239	3.195	3.239	3.324	3.239	3.259
PTB	3.239	3.095	3.245	3.327	3.233	3.373
PUB	3.239	3.195	3.239	3.324	3.239	3.259
TB	3.183	3.135	3.244	3.280	3.243	3.367
UB	3.239	3.239	3.239	3.255	3.255	3.255
PB	3.239	3.195	3.239	3.324	3.239	3.259
B	3.239	3.239	3.239	3.255	3.255	3.255

From Fig. 11(a), one can see that the line of GA is almost straight, and all values are around 3.25 km. DBC shows a similar performance to GA, but three models of DBC performed relatively better. While DBA performs vary depending on different models, especially the models with four sources show better performances than other models. Figure 11(b) shows MDED’s trend of DBC, GA, and DBA based on 16 models when the area threshold of the bounding box is set to 10,000  $\text{km}^2$ . One can see that the models with four sources have the same performance regardless of DBC, GA, and DBA. But the values of MDED change a lot when less than four sources are used.

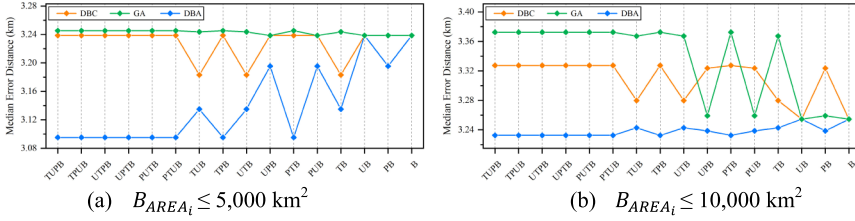


Fig. 11. MDED of models based on two area thresholds.

From Fig. 10 (MED of the models) and Fig. 11 (MDED of the models), it shows that DBA has the best performance in all cases, GA performs better in MED, and DBC performs better in MDED. Compared with MED, MDED have smaller error distances for all models.

## 6 Conclusion

Twitter has demonstrated its importance for gathering and publishing up-to-date information during a real-world event. Geographic information plays an important role in emergency response and event monitoring. However, only 2% of tweets are with geo-tags, hence geolocation inference of tweets is still a major challenge. In this study, we proposed various models to predict geolocation of tweets, as organized as follows: (1) Twitter data collection; (2) data cleaning and extract geo-tagged tweets related to COVID-19; (3) location entity extraction from location-related metadata of tweets based on NER; (4) construction of three coordinate datasets on the basis of gazetteers and digital boundaries of the US; (5) model implementation based on different area thresholds of bounding box; (6) model evaluation.

The proposed method has fully used all potential location-related attributes to predict tweets’ geolocation. When the area threshold of the bounding box is set to  $10,000 \text{ km}^2$ , the best model can successfully predict the geolocation of 90.8% of COVID-19 related tweets with the mean error distance of 4.824 km and the median error distance of 3.233 km. This method has achieved the best performance compared with previous methods.

There still exist some deficiencies in this study. Firstly, the library of NER is limited and does not contain every county’s name, which results in some useful information being filtered out. Secondly, even though the distance threshold is introduced to reduce the interference caused by duplicate county names, there still exist counties with the same name located in the same bounding box. Thirdly, in some cases, several location entities can be extracted based on NER, but in this study, only the first location entity that meets the criteria is chosen. However, the real location-related information does not always appear in the first position.

For further study, the proposed method can be applied to other emergency datasets (e.g., bushfires, typhoons, and earthquakes). When computing the average lon-lat coordinates of geo-tagged tweets located in a county, different weights can be added to each

tweet. In addition, techniques such as natural language processing and deep learning models can strengthen text analysis and promote the development of this research field.

**Acknowledgements.** This research is sponsored by China Scholarship Council (CSC).

## References

1. Singh, L., et al.: A first look at COVID-19 information and misinformation sharing on Twitter (2020)
2. Banda, J.M., et al.: A large-scale COVID-19 Twitter chatter dataset for open scientific research—an international collaboration (2020)
3. Worldometers. <https://www.worldometers.info/coronavirus/>
4. Prieto, V.M., Matos, S., Alvarez, M., Cacheda, F., Oliveira, J.L.: Twitter: a good place to detect health conditions. *PloS One* **9**, e86191 (2014)
5. Paul, M.J., Dredze, M.: You are what you tweet: analyzing Twitter for public health. In: Fifth International AAAI Conference on Weblogs and Social Media
6. Steiger, E., De Albuquerque, J.P., Zipf, A.: An advanced systematic literature review on spatiotemporal analyses of Twitter data. *Trans. GIS* **19**, 809–834 (2015)
7. Crooks, A., Croitoru, A., Stefanidis, A., Radzikowski, J.: #Earthquake: Twitter as a distributed sensor system. *Trans. GIS* **17**, 124–147 (2013)
8. Sinnenberg, L., Buttenheim, A.M., Padrez, K., Mancheno, C., Ungar, L., Merchant, R.M.: Twitter as a tool for health research: a systematic review. *Am. J. Public Health* **107**, e1–e8 (2017)
9. 50+ Twitter statistics & facts for 2020. <https://www.websitehostingrating.com/twitter-statistics/>
10. 10 Twitter Statistics Every Marketer Should Know in 2019. <https://au.oberlo.com/blog/twitter-statistics>
11. Ajao, O., Hong, J., Liu, W.: A survey of location inference techniques on Twitter. *J. Inf. Sci.* **41**, 855–864 (2015)
12. Huang, C., Tong, H., He, J., Maciejewski, R.: Location prediction for tweets. *Front. Big Data* **2**, 5 (2019). <https://doi.org/10.3389/fdata>
13. Laylavi, F., Rajabifard, A., Kalantari, M.: A multi-element approach to location inference of Twitter: a case for emergency response. *ISPRS Int. J. Geo-Inf.* **5**, 56 (2016)
14. Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., Gawron, J.-M.: Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. *PloS One* **11**, e0157734 (2016)
15. Gao, Y., Wang, S., Padmanabhan, A., Yin, J., Cao, G.: Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *Int. J. Geogr. Inf. Sci.* **32**, 425–449 (2018)
16. Li, W., Serdyukov, P., de Vries, A.P., Eickhoff, C., Larson, M.: The where in the tweet. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 2473–2476. ACM
17. Cheng, Z., Caverlee, J., Lee, K.: A content-driven framework for geolocating microblog users. *ACM Trans. Intell. Syst. Technol. (TIST)* **4**, 2 (2013)
18. Hecht, B., Hong, L., Suh, B., Chi, E.H.: 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, pp. 237–246. ACM

19. Ryoo, K., Moon, S.: Inferring Twitter user locations with 10 km accuracy. In: Proceedings of the 23rd International Conference on World Wide Web, pp. 643–648. ACM
20. Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., Ratti, C.: Geo-located Twitter as proxy for global mobility patterns. *Cartogr. Geogr. Inf. Sci.* **41**, 260–271 (2014)
21. Priedhorsky, R., Culotta, A., Del Valle, S.Y.: Inferring the origin locations of tweets with quantitative confidence. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing, pp. 1523–1536. ACM
22. Cheng, Z., Caverlee, J., Lee, K.: You are where you tweet: a content-based approach to geo-locating twitter users. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 759–768. ACM
23. Chandra, S., Khan, L., Muhaya, F.B.: Estimating Twitter user location using social interactions--a content based approach. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, pp. 838–843. IEEE (2011)
24. Chang, H.-W., Lee, D., Eltaher, M., Lee, J.: @Phillies tweeting from Philly? Predicting Twitter user locations with spatial word usage. In: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pp. 111–118. IEEE Computer Society (2012)
25. Ikawa, Y., Vukovic, M., Rogstadius, J., Murakami, A.: Location-based insights from the social web. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1013–1016. ACM
26. Abrol, S., Khan, L.: Tweethood: agglomerative clustering on fuzzy k-closest friends with variable depth for location mining. In: 2010 IEEE Second International Conference on Social Computing, pp. 153–160. IEEE (2010)
27. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th International Conference on World Wide Web, pp. 61–70. ACM
28. Bouillot, F., Poncelet, P., Roche, M.: How and why exploit tweet's location information? In: AGILE 2012: 15th International Conference on Geographic Information Science (2012)
29. Lingad, J., Karimi, S., Yin, J.: Location extraction from disaster-related microblogs. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 1017–1020. ACM
30. Li, R., Wang, S., Deng, H., Wang, R., Chang, K.C.-C.: Towards social user profiling: unified and discriminative influence model for inferring home locations. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1023–1031. ACM
31. Takhteyev, Y., Gruzd, A., Wellman, B.: Geography of Twitter networks. *Soc. Netw.* **34**, 73–81 (2012)
32. Li, C., Sun, A.: Fine-grained location extraction from tweets with temporal awareness. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 43–52. ACM
33. Tweeting Made Easier. [https://blog.twitter.com/official/en\\_us/topics/product/2017/tweetingmadeeasier.html](https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html)
34. An Introduction to JSON. <https://www.digitalocean.com/community/tutorials/an-introduction-to-json>
35. Tweet Location Metadata. <https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/overview/geo-objects>
36. Li, B., Chen, Z., Lim, S.: Geolocation prediction from tweets: a case study of influenza-like illness in Australia. In: GISTAM, pp. 160–167

37. Singh, J., Dwivedi, Y., Rana, N., Kumar, A., Kapoor, K.: Event classification and location prediction from tweets during disasters. *Ann. Oper. Res.* **283**(1–2), 737–757 (2017). <https://doi.org/10.1007/s10479-017-2522-3>
38. Regular Expression Language - Quick Reference. <https://docs.microsoft.com/en-us/dotnet/standard/base-types/regular-expression-language-quick-reference>
39. Marujo, L., et al.: Automatic keyword extraction on Twitter. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 637–643
40. Kouzy, R., et al.: Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus* **12**, e7255 (2020)
41. Chen, E., Lerman, K., Ferrara, E.: Tracking social media discourse about the COVID-19 pandemic: development of a public coronavirus Twitter data set. *JMIR Public Health Surveill.* **6**, e19273 (2020)
42. The National Counties Gazetteer File. <https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html>
43. USA Counties. <https://www.arcgis.com/home/item.html?id=a00d6b6149b34ed3b833e10fb72ef47b>