computing@computingonline.net
www.computingonline.net

Print ISSN 1727-6209
On-line ISSN 2312-5381
International Journal of Computing

TWITTER LOCATION-BASED DATA: EVALUATING THE METHODS OF DATA COLLECTION PROVIDED BY TWITTER API

Noor Ahmed Qarabash [1], Haneen Ahmed Qarabash [2]

[1] University of Information Technology and Communications, Baghdad, Iraq, noor.ahmed@uoitc.edu.iq
[2] University of Baghdad, Baghdad, Iraq haneenqarabash@gmail.com

Paper history:
Received 12 May 2020
Received in revised form 16 June 2020
Accepted 17 October 2020
Available online 30 December 2020

Keywords:
Social media;
Twitter;
location data;
data analysis.

Abstract: Twitter data analysis is an emerging field of research that utilizes data collected from Twitter to address many issues such as disaster response, sentiment analysis, and demographic studies. The success of data analysis relies on collecting accurate and representative data of the studied group or phenomena to get the best results. Various twitter analysis applications rely on collecting the locations of the users sending the tweets, but this information is not always available. There are several attempts at estimating location based aspects of a tweet. However, there is a lack of attempts on investigating the data collection methods that are focused on location. In this paper, we investigate the two methods for obtaining location-based data provided by Twitter API, Twitter places and Geocode parameters. We studied these methods to determine their accuracy and their suitability for research. The study concludes that the places method is the more accurate, but it excludes a lot of the data, while the geocode method provides us with more data, but special attention needs to be paid to outliers.

1. INTRODUCTION

Analysis of social media activity has become a popular tool for demographic studies, market research and analytics of social dynamics. Social media allow a diverse set of people to communicate on shared interests easily and create online communities. These platforms and the communities that are formed on them can offer a lot of insight into shared concerns, interests, activities, events, etc. Twitter in particular has emerged as a valuable tool for research due to its popularity, having around 342 million active users posting up to 400 million tweets a day [1]. Additionally, Twitter provides an application program interface (Twitter API) that supports collecting comprehensive data about tweets and users [2, 3].

Studies on Twitter data can vary in their applications and range, one important application is disaster response[4, 5]. There have been a number of studies in that field. For example, [6] used a mix of crowdsourcing manual labeling and machine learning to identify eye witness accounts from tweets regarding an incident or a disaster, using simple text and domain features of the tweet to classify witnesses as direct, indirect and vulnerable direct. Similarly, [7] some of the reactions were mapped across the world to the 2017 Iran-Iraq earthquake to determine what devices (such as android devices or IPhones) were typically used to respond to the event, what countries interacted the most and at what time as well as to investigate the significance of these parameters.

Another common application of Twitter data mining is sentiment analysis[8, 9], a field of research which attempts to understand the way a group of people feel regarding an issue to inform decision making [10]. For example [11] identified shared issues and concerns globally regarding climate change and sustainability by analyzing tweets using the hashtag #worldEnviromentDay. Another example is the study conducted by [12] that collected English and Turkish tweets to analyze the public sentiment toward the Syrian refugee crisis. The study also contrasted the differences between sentiment in Turkish tweets in comparison to tweets in English to highlight differences in the handling of

the issue between Turkey and other countries.

It can be noted that previous research was focused on the subject when collecting data, using the names of events or certain hashtags as query. However, in some cases it would be more logical to focus on the location the tweets originated from. The location of the original poster of the tweet can help in concentrating disaster relief efforts for crisis events and collecting more accurate demographic data for research. If we consider the example of a disaster response, being able to ensure tweets originated from the location of the disaster would help eliminate other chatter such as people from other areas showing concern or support.

The reason why most research often focuses on topic and not location is likely due to the fact that location (geo) tagging is optional for twitter users and approximately only 1% of users enable the option [13, 14] or state a specific location in their profiles. This motivated researchers to attempt a number of location prediction algorithms that involve a combination of word searches, inference, language, tweeting behavior, context and time zones [15-18]. However, these attempts involve analyzing data that has been previously collected to connect them into a location, rather than focus on the location while collecting the data. Twitter API does provide two methods to specify location during the data collection stage. These methods are described by the twitter API documentation [19]:

1. Twitter places: this method involves using a specialized place ID defined by Twitter to return tweets by users in that location. This method only returns tweets by users who opted to enable geo location services, and as stated before they present a very small percentage of users. This method has different granularity, or scale, to set it up for a country, city or neighborhood.
2. Geocode: this method requires latitude, longitude and a radius to return tweets that are either geo tagged as that location or belong to users who specified that location in their profiles.

From their description we can infer that Twitter places are more precise but exclude a lot of the data we may need, while the geocode offers a less reliable location but can offer us a more extensive range of research data, however this disparity has not been well explored in research. This paper aims to evaluate the accuracy of both methods by analyzing location data retrieved and mapping out coordinates to establish a guideline for location based Twitter analysis.

2. RELATED WORK

There have been several studies attempting to collect location data by predicting or inferring the origin location of tweets based on the text of the Tweet or on other features associated with it such as time or language. In [17] a survey of research was conducted that explored the idea of inferring the locations of users using the Tweets themselves. They found that most research used the message (Tweet) text to infer location. However, this approach is quite difficult due to the nature of Twitter messages. Those messages are short and usually contain many abbreviations, errors, hash tags, and other language uses and lack or rules on how to write on social media. Other methods were the users stated locations in their profiles, their geolocations provided by the devices they are using if they permitted it's sharing, and the user's social network and the locations of the other users they frequently interact with. Each one of these methods offers its advantages and disadvantages.

One research by [20] studied the possibility of predicting the location of the user from the content of the Tweet by looking for location indicative words. The study assumed that people are more likely to talk about their locations or places frequent in their tweets. For example, a person from London is more likely to mention London than they are to talk about Tokyo. They are also more likely to use words like British or name places in or people from the UK. Of course, it's possible for people from other places to use such words when talking about the UK but the frequency of use of these words and their appearing together more often can be an indicator that the person is from the UK.

Similarly, [14] proposed a new approach to infer the user's country by using the text of the message (Tweet) and Google trends. The method they suggested assumes that a person from a specific country is more likely to tweet about topics that are important and concerns that country. So, by looking at users who frequently tweet about trending topics we can infer the user's country. This approach can be useful in research that conducts sentiment analysis or explores the public perception of political, environmental, economical, etc. issues. However, this method does not take into account users who regularly tweet about several countries or have accounts dedicated to issues that relate to other countries. So, any research that requires knowing that information cannot use this method.

For this paper, we will investigate Twitter's own methods of determining the location offered for developers in their API. We will explore the suitability of these methods for research utilizing users' location.

3. METHODOLOGY

For this study, we used the Twitter search API to

collect six sets of tweets from five cities: Vancouver (Canada), Amsterdam (the Netherlands), Kuala Lumpur (Malaysia), Doha (Qatar) and Providence (USA), these cities were chosen for having a medium population density and active twitter user base.

For every city two datasets were created. The first dataset titled "Places" was collected using the first method provided by Twitter, the places search query is based on the city name. The granularity (scale) was set as "City". This search query returned tweets that have geo-location attached due to the users enabling the feature on their devices. The second dataset, titled "Geocode" was gathered using the second method known as "geocode parameter search", which involves querying the API with the specific coordinates of the city as well as a radius (for the circle the search query will cover). The radius was selected for each city based on its area. The two sets were collected over the same time of the day and all other search parameters, including date and language fixed to ensure that there is no outside variables influence on the analysis. The Twitter search API was not designed to be exact; it does not return all published tweets but a representative sample of them, which is not an issue since we are comparing both methods using with that same limitation. The search API offers the choice of focusing on recent, popular or a mix of recent and popular, the later was used in this study to provide variety to the dataset. Table (1) shows a breakdown of the cities and the collected datasets.

Table 1. Breakdown of the five cities captured datasets

City	Coordinates and Raduis	Dataset	Number of tweets
Vancouver, Canada	49.2827, -123.1207, 7km	Places	783
		Geocode	46,530
Amsterdam, Netherlands	52.3667, 4.8945, 9km	Places	974
		Geocode	56,008
Kuala Lumpur, Malaysia	3.1390, 101.6869, 9 Km	Places	701
		Geocode	48,961
Doha, Qatar	25.2854, 51.5310, 9Km	Places	337
		Geocode	14,898
Providence, US	41.8240, -71.4128, 4.5Km	Places	205
		Geocode	35,447

Each dataset included three main fields:

Username: user handle chosen by the owner of the account. Usernames are unique and can be used as an ID for the user.

Location: The address set by the user for his profile. This can be a city name, district, state, country or a combination. Twitter allows the location to be entered as text by the users in their profile page.

Coordinates: latitude and longitude, which are only recorded if a user has enabled exact geolocation on his device. While most modern devices have this feature, many users choose to disable it for privacy reasons, or to conserve battery [21]. The analysis consisted of two main tracks:

1. Comparing what percentage of the tweets were captured by both methods for the five cities. To compare the two methods, we found duplicate tweets in both datasets for each city and calculated the percentage of those duplicates in relation to the total number of tweets in both the "Places" dataset and the "Gecode" dataset.

2. Evaluating the location data collected for one city by comparing the user location names and mapping the available coordinates.

For the location analysis, the city of Vancouver was selected because it is one of the largest datasets of the five and it is in an English speaking country, which would make location names easier to compare. The location data collected for the city of Vancouver was compared using the user location names and coordinates in both datasets.

However, before we can analyze location data, it is critical to identify unique users, since some users made multiple tweets in the same location, which can bias the location frequency in later analysis. Unique users were identified by searching for matching usernames and locations.

It's worth noting that the locations and addresses could sometimes contain joke answers like "my home" or 'nowhere' because the location in Twitter is free to input text box. It is estimated that 34% of locations users use in Twitter are not real[22]. Dropping any location that is not repeated more than two times for the "Places" dataset and five times for the "Geocode" dataset filtered some of these locations.

A very small percentage of users had coordinates attached to their tweets, however, even mapping that limited sample can provide insight to the way the two methods work. All maps and coordinates where drawn using Folium python library.

4. RESULTS

4.1 CAPTURE SIMILARITY

For each city duplicates between the datasets were identified and the percentages were calculated. Table (2) provides a breakdown of the similarity observed for each city.

Table 2 Dublicates captured by the two methods

City	places Tweets	Gecode Tweets	Duplicates	Duplicates percentage in Places Dataset
Vancouver	783	46,530	755	96,55%
Amsterdam	974	56,008	962	98.77%
Kuala Lumpur	701	48,961	453	64.62%
Doha	337	14,898	327	97.03%
Providence	205	35,447	200	97.56%

4.2 FREQUENCY OF LOCATION NAMES

For the "Places" dataset of Vancouver City, there were 352 unique user profiles. To eliminate locations that were written as a joke or may be very personalized, any location that was not used by at least three users was eliminated. The result shows 245 entries for location data. 20 users (8%) left the location data empty and were discarded. Fig.1 demonstrates the breakdown of the remaining 225 locations used by users to refer to their whereabouts.

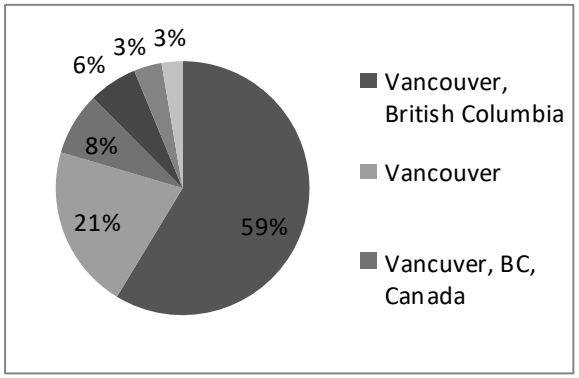

Figure 1 – frequency of location names occurring in the dataset for Vancouver

The most common location name was "Vancouver, British Columbia", though written in few different ways such as "Vancouver, BC", or "Vancouver, B.C.". The second most common location was Simply "Vancouver" Followed by "Vancouver, BC, Canada". All 225 were a variation of Vancouver, British Columbia, and Canada, in different combinations.

Of the locations that were dropped for occurring only once or twice there is a broad range of locations. Some are also variations of Vancouver, sometimes misspelled or with describers like "North Vancouver" or "The Drive, Vancouver", though others are from nearby cities or countries.

For the "Geocode" dataset for Vancouver City, there were 19,922 unique users. Because of the considerable number, the locations were filtered to locations that appeared five or more times, resulting in 12,863 locations. 4342 of users did not provide a location, which constitute around 33.8% of users, a considerably more significant percentage than in the "Places" dataset. The remaining 8,521 location names included many variations of words and cities, so they were grouped into five categories:

– Vancouver: for any location containing the word Vancouver such as "Vancouver, BC", "Vancouver, Canada" and "Greater Vancouver".
– Other Canadian cities and territories: this refers to any cities or territories that are not Vancouver, such as "Toronto" and "Nova Scotia". It additionally includes general location names like "British Columbia" or "Canada".
– The United States: The dataset contained a considerable number of locations that referred to the United States and some of its cities, which warranted the creation of its own category.
– Other countries: This refers to any city or country that is outside Canada or the USA, such as "Chile", "London, UK", and "Singapore".
– Other: this includes location names that do not refer to a specific place but are common enough not be filtered, for example "World", "Earth" and "Hell".

Fig. 2 illustrates the breakdown of these categories.

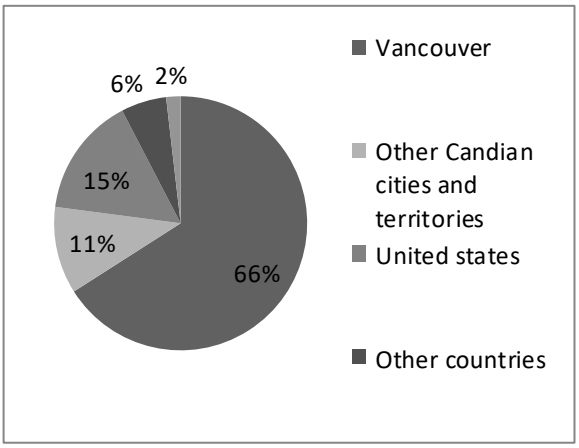

Figure 2 – Frequency of location name groups in geocode dataset for Vancouver

4.3 COORDINATES

To further explore the location data captured we can use the coordinates that were made available by a small percentage of users who enabled applications to post their exact coordinates presented by latitude and longitude.

From the "Places" dataset for Vancouver City, only 52 unique coordinates were provided by users pointing to the origin of the tweet. These coordinates were mapped to see how much they match the area we want to capture data from. Fig.3.A shows the coordinates from the "Places" dataset over a map of Vancouver, represented with dots.

The same process was repeated for the "Geocode" dataset, 50 unique coordinates were identified, of which 45 are in common with the "Places" dataset. Fig.3.B illustrates the mapping of coordinates from the "Geocode" dataset. The blue circle represents the radius of the search query used in collecting the data. As Fig.3.B demonstrates, the radius of the search can influence the results. The search radius in contrast to the cities' borders for the other 4 cities can be seen in Fig. 4.

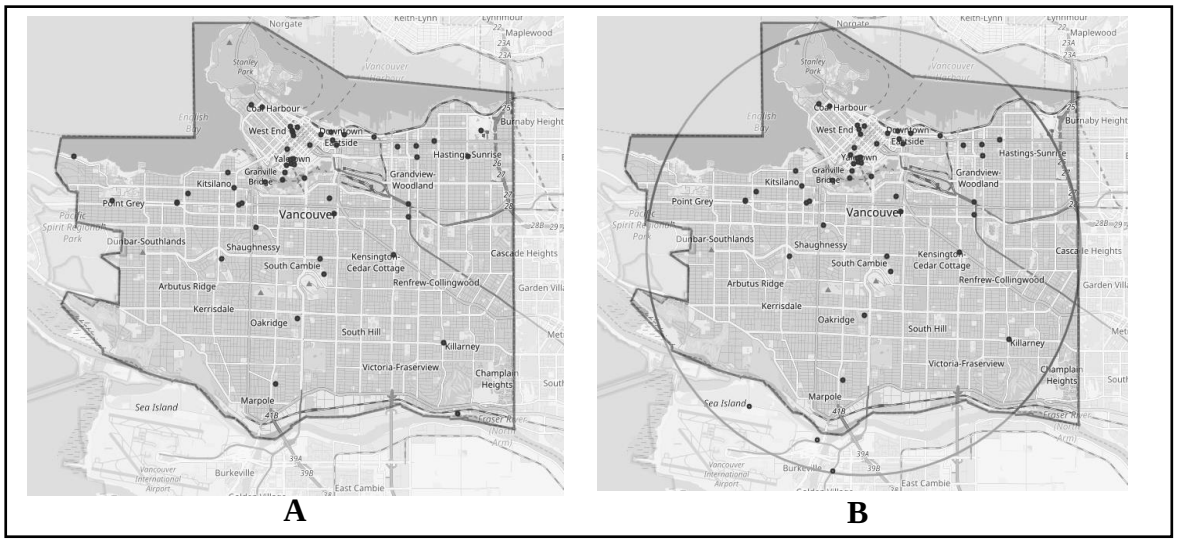

Figure 3 – A. mapped coordinates found in the places dataset
B. mapped coordinates found in the geocode dataset

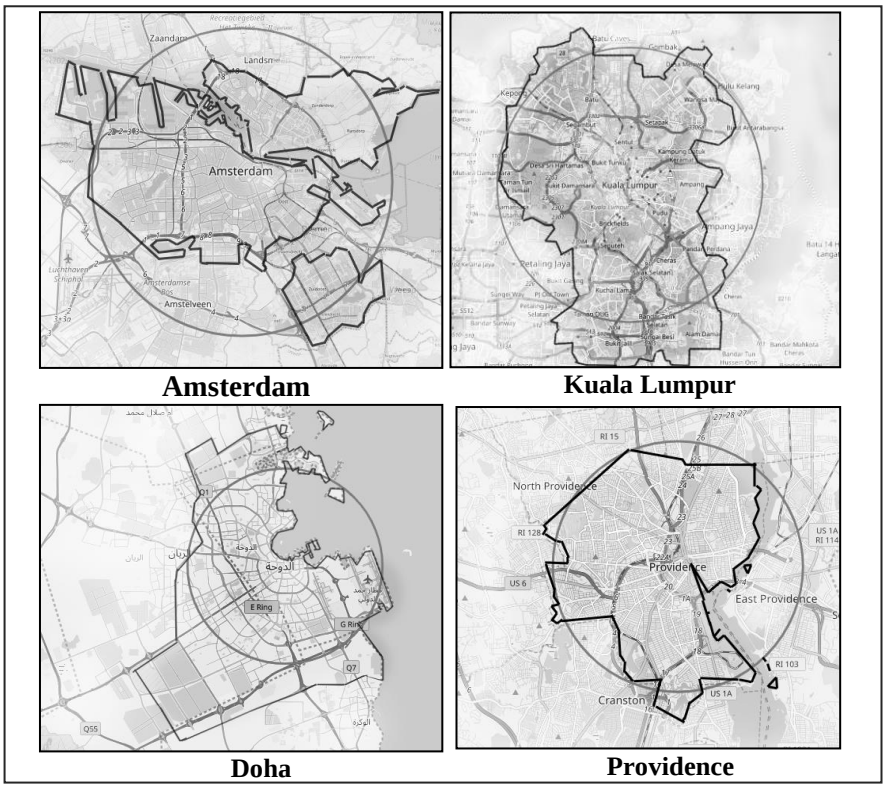

Figure 4 – Maps of 4 cities illustrating the search radius

5. DISCUSSION

The location data analysis of the Vancouver city datasets illustrates that the places method is more accurate in collecting tweets only from the designated area. This can be seen in both Fig. 1 and Fig. 3.A, where the location names and coordinates all fit generally within the area of Vancouver. The geocode method on the other hand contains a lot of uncertainty. For one thing, Fig. 3.B clearly shows some of the coordinates mapped were outside the city limit and a good percentage of the location names referred to other cities or countries. It is important to note that the later issue could be due to tourists and visitors who tweet from the designated area but have locations on their profiles that refer to their home countries.

One of the issues that became noticeable when applying the geocode method was the difficulty of selecting the correct radius. The five cities had unique shapes and surrounding areas. The incorrect radius could exclude some important areas outside the radius or include other surrounding areas inside the radius that make our results less exact.

As we can see in Fig.4, the radius doesn't always include the entire city or contains areas outside of it. Since the places method captures all tweets with known location within the city, the excluded data can be estimated by the percentage of tweets captured by both methods in relation to the places dataset. For four out of the three cities, the average of the mutual tweets captured is 97.48%. Leading to a 2.62% estimated percentage of tweets excluded when using the geocode method. However, for one city, Kuala Lumpur, the mutual tweets percentage rose up to 64.62%, making the excluded percentage of tweet 35.48%, one possible explanation for this is that the excluded areas in the other cities were not densely populated while in Kuala Lumpur it had more of an active twitter user base that became noticeable.

The areas outside the city that are included by a large radius are more challenging to quantify, but by controlling the radius we can make certain to only include locations that are relevant for our research. Future research could investigate collecting the data with multiple small radiuses to cover the exact area of a city.

6. CONCLUSION

Collecting Twitter data remains a valuable tool for researchers. However, a major concern of research is ensuring the tweets originated from a desired location. This paper examined the two methods provided by Twitter API for collecting data based on location by comparing the collected tweets and analyzing location names and coordinates returned by both. We found out that both methods offer a good approximation of data from the required location. However, the places method has a limited scope because it excludes any tweets that are not geo enabled. Meaning it can be used only in applications where the target demographic is users who enable geolocation. This makes this method more suitable for research that requires exact location information such as studies concerning a specific city's population. The geocode method offers a more extensive range of tweets that may contain a percentage of tweets from outside the desired area or exclude tweets if they fall outside its range. One of the limitations of that method is selecting the correct radius of the search to obtain the best results. Even though, the geocode method does provide a more expansive dataset that might be more suitable for general research applications where the exact location is not the focus, such as disaster effects and response or disease outbreaks.

7. REFERENCES

[1] M. J. Cumbraos-Sánchez, R. Hermoso, D. Iñiguez, J. R. Paño-Pardo, M. Á. A. Bandres, and M. P. L. Martinez, "Qualitative and quantitative evaluation of the use of Twitter as a tool of antimicrobial stewardship," *International Journal of Medical Informatics*, vol. 131, p. 103955, 2019.

[2] P. Rafail, "Nonprobability sampling and Twitter: Strategies for semibounded and bounded populations," *Social Science Computer Review*, vol. 36, pp. 195-211, 2018.

[3] Z. Saaya and T. W. Hong, "The development of trust matrix for recognizing reliable content in social media," *International Journal of Computing*, vol. 18, pp. 60-66, 2019.

[4] T. H. Nazer, G. Xue, Y. Ji, and H. Liu, "Intelligent disaster response via social media analysis a survey," *ACM SIGKDD Explorations Newsletter*, vol. 19, pp. 46-59, 2017.

[5] J. R. Ragini, P. R. Anand, and V. Bhaskar, "Big data analytics for disaster response and recovery through sentiment analysis," *International Journal of Information Management*, vol. 42, pp. 13-24, 2018.

[6] K. Zahra, M. Imran, and F. O. Ostermann, "Automatic identification of eyewitness messages on twitter during disasters," *Information Processing & Management*, vol. 57, p. 102107, 2020.

[7] I. T. Hamdan and A. Malik, "Demographic analysis of Twitter users during the 2017 Iran-Iraq earthquake," *Proceedings of the 2018 Fifth*

HCT Information Technology Trends (ITT), 2018, pp. 149-153.

[8] V. Kharde and P. Sonawane, "Sentiment analysis of twitter data: a survey of techniques," *arXiv preprint arXiv:1601.06971*, 2016.

[9] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," *Proceedings of the 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 452-455.

[10] C. Puschmann and A. Powell, "Turning words into consumer preferences: How sentiment analysis is framed in research and the news media," *Social Media+ Society*, vol. 4, p. 2056305118797724, 2018.

[11] A. Reyes-Menendez, J. Saura, and C. Alvarez-Alonso, "Understanding# WorldEnvironmentDay user opinions in Twitter: A topic-based sentiment analysis approach," *International Journal of Environmental Research and Public Health*, vol. 15, p. 2537, 2018.

[12] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics*, vol. 35, pp. 136-147, 2018.

[13] P. A. Longley, M. Adnan, and G. Lansley, "The geotemporal demographics of Twitter usage," *Environment and Planning A*, vol. 47, pp. 465-484, 2015.

[14] P. Zola, P. Cortez, and M. Carpita, "Twitter user geolocation using web country noun searches," *Decision Support Systems*, vol. 120, pp. 50-59, 2019.

[15] J. Mahmud, J. Nichols, and C. Drews, "Home location identification of twitter users," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, p. 47, 2014.

[16] A. Rahimi, T. Cohn, and T. Baldwin, "Semi-supervised user geolocation via graph convolutional networks," *arXiv preprint arXiv:1804.08049*, 2018.

[17] X. Zheng, J. Han, and A. Sun, "A survey of location prediction on Twitter," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, pp. 1652-1671, 2018.

[18] T. H. Do, D. M. Nguyen, E. Tsiligianni, B. Cornelis, and N. Deligiannis, "Multiview deep learning for predicting twitter users' location," *arXiv preprint arXiv:1712.08091*, 2017.

[19] Twitter. (2019, 12/12/2019). *Twitter developer documentation*. [Online]. Available at: https://developer.twitter.com/

[20] M. N. Y. Utomo, T. B. Adji, and I. Ardiyanto, "Geolocation prediction in social media data using text analysis: A review," *Proceedings of the 2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 84-89.

[21] K. Lin, A. Kansal, D. Lymberopoulos, and F. Zhao, "Energy-accuracy trade-off for continuous mobile device location," *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, 2010, pp. 285-298.

[22] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 237-246.

Noor Ahmed Qarabash *assistant lecturer at the University of Information Technology and Communi-cations in Baghdad, Iraq. She graduated with a Bachelor's degree in Information Technology Engineering in 2011 and obtained Master's degree in Cybersecurity and Management in 2014 from Warwick University. Her research interests include Cybersecurity, Data analysis, Internet of Things and Electronic governance.*

Haneen Ahmed Qarabash *lecturer at the Computing Science Department at the College of Science in the University of Baghdad. She graduated with a bachelor's in Computing Science in 2007 and obtained a master's degree in Web Development in 2009 from Al-Nahrain University. She was awarded with a Doctor of Philosophy (PhD) from Newcastle University in Human-Computer Interaction in 2018. Her research interests are in educational technology, students experience and engagement, alternative assessment, web development, human-computer interaction and user experience.*