exeley INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Article | DOI: 10.21307/ijssis-2020-012 Issue 1 | Vol. 13 (2020)

A novel algorithm for estimation of Twitter users location using public available information

Yasser Almadany[1], Khalid Mohammed Saffer[2], Ahmed K. Jameil[1] and Saad Albawi[1,*]

[1]College of Engineering, University of Diyala, Baquba, Iraq.

[2]College of Science, University of Diyala, Baquba, Iraq.

*E-mail: saad.albawi@engineering.uodiyala.edu.iq

The paper was edited by Joyanta Kumar Roy.

Received for publication December 10, 2019.

Abstract

Social media networks are an attractive and hot research area in the big data community because of their numerous active users. One of the most widely studied topics in social networks is the prediction from the public available data. Recently, researchers have successfully predicted many statistical and human properties from social media networks using different machine learning algorithms. In this paper, a new efficient and accurate algorithm is proposed to predict the country location of a Twitter user using his or her public information only. The proposed algorithm employs the public information of the Twitter user and that of his or her followers and friends to predict his or her location without using GPS data. A convenient data set of Twitter users is gathered and used to test our proposed algorithm using KNIME software. The proposed algorithm is compared with other state-of-the-art algorithms, and results showed that our proposed algorithm significantly outperforms other location detection algorithms by using Twitter users from different countries.

Keywords

Twitter, Social media, KNIME, Followers, Time zone, Location, Friends.

Big data analysis is significant and widely used in many real-world applications in different fields. With the increasing volumes of unstructured data and the multiplicity of their sources, such as the internet, social media, and other large data sources, using a single computer processor may not be able to deal with such a large amount of data. By contrast, the need for new technologies that can deal with large and unstructured data is in high demand, such as non-traditional databases NoSQL, given that traditional databases cannot efficiently deal with such data (Anber et al., 2016).

Recently, predictive analyses of big data gained increasing attention in the computer science research society. The analysis of big data sets employs statistical methods or machine learning models to predict future results or unknown outcomes (Brown et al., 2015). Moreover, data mining techniques are used to search unstructured data based on current and historical data to predict future or unknown information on behavior, direction, and activity of these data. Probabilistic analyses require many experts to construct predictive models for the prediction process (Jayanthi et al., 2017). After the emergence of social media such as Facebook and Twitter, big data analysis and prediction techniques became one of the most important elements for identifying human behavior and predicting unknown aspects such as location, gender, and nationality. Social media data analysis is important for scientists, researchers, and governments (Tufekci, 2014) to understand and analyze many daily issues.

Social networking, characterized by its great potential and openness to data sharing, is a main platform for exchanging information and ideas. The availability of large data through social communication

greatly impacted the occurrence of qualitative breakthrough in the study and analysis of the methodology of human behavior, which requires careful studies on strengths and weaknesses (Tufekci, 2014). Many cases of big data analyses require real-time data analyses with comprehensive understanding of the impress factors for some cases, such as influence, reach, and relationship information. To perform correct and accurate analysis of social media, the data require a clear and accurate understanding of the facts they contain (Gentry, 2015). In most cases, we want to know the connections between social data and another event or obtain motivation results from social data analytics to predict some events. A number of outstanding articles in this area include Twitter Mood Predicts, The Personality Predicting, and The Present with Google Trends.

Twitter is one of the largest social media that enables researchers to verify their hypotheses, provide them access to enormous data, which they could use to build realistic applications (Liu, 2013). Twitter allows subscribers to socialize via text messages with a length of 140 characters. The Twitter application contains various types of data, namely, static and dynamic data, such as profile data and Twitter messages, respectively. Tweets can be images, videos, text messages, and more. Twitter has recently been used as a source of information for predicting and/or monitoring real-world results (Atefeh and Khreich, 2015). The large information provided by Twitter, such as Twitter messages, user profile information, and number of followers in the network, plays an important role in data analysis.

In this study, a new efficient and robust algorithm is proposed to predict the country location of Twitter users. The new proposed algorithm uses the available public information of Twitter users and that of their friends and followers. No GPS information is used in this study given that many Twitter users disable this feature due to privacy concerns. To test our proposed algorithm, a big data set is downloaded and gathered from the Twitter site using its API and KNIME software. Results show that our algorithm is efficient and robust against wrongly written and meaningless information by many Twitter users.

The rest of this paper is organized as follows. The second section presents a brief literature review on the prediction techniques from social media with a focus on Twitter. The third section explains the data gathering methods and data statistics. Our proposed algorithm for predicting the home location of Twitter users is explained in the fourth section. Results and discussions are provided in the fifth section. The sixth section concludes the paper and discusses future work.

Related works

In literature, there are many algorithms and methods proposed to predict several human and social patterns and information from social media sites. In this section, we will briefly review some of these algorithms.

Huang et al. (2014) discussed the challenge of the using rich personality traits to identify the nationality of the Twitter users from their profiles. Authors made two constraints when querying the application programming interface (APIs), that they either explicitly stated in their profile that they are located in Qatar or have at least one geo-tagged tweet originating from Qatar. The best performance of Gradient Boosted Tree was achieved with a number of trees equal to 300 and the best overall accuracy is 83%. This algorithm gives good results but the authors applied it only on one country and needs to be further examined in other country locations.

Another new algorithm for predicting Twitter user location is proposed in the study of Chang et al. (2012). In the present research, the authors aimed to improve the quality of predicting the home location of a Twitter user using probability frameworks. The authors proposed a new approach to estimate the spatial word usage probability with Gaussian mixture models. Furthermore, unsupervised measurements are used to rank local words which effectively remove noises that harm prediction. The authors showed that their approach can achieve a comparable and enhanced performance to the state-of-the-art in some cases, using less than 250 local words selected by the proposed method. In this paper, the authors utilized 3183 local words selected through supervised classification based on 11,004 hand-labeled ground truths. Using only 250 selected local words, the proposed approach can predict the home locations of selected Twitter users within an area of 100 miles with an accuracy of 0.499 or 509.3 miles of average error distance at best.

Mcgee et al. (2013) proposed a novel network-based approach (Friendly Location System) for location estimation in social media that integrates evidence of socialite strength between users to improve location estimation based on an examination of over 100 million geo encoded Tweets and 73 million Twitter user profiles. Several identified factors include the number of followers and how users interact which could reveal the distance between a pair of users. The system reduces the average error distance for 80% of Twitter users from 40 miles to 21 miles using only the information from the user's friends and their friends. In this study, fivefold cross validation is used on the target users to evaluate the system and the Friendly

Location System against 249,584 target users. The basic Friendly Location System predicts location within 25 miles or 63.9% of the time, whereas the basic system has an average error distance of 21.4.

In the study of Tsujioka et al. (2016), a study estimated the location of a person according to the Tweets of a Twitter user, and this approach is used to develop tourism. The estimation method for generating decision trees is based on machine learning, and it aims to classify Tokushima and non-Tokushima Twitter users. The accuracy of the estimation method is at 60%, and the method is based on analyzing Twitter messages with the distinction between local citizens and tourists. The proposed method generates a decision tree using machine learning, and four algorithms are used for the decision tree production method, namely, Forest algorithm, C4.5, NBTree, and REPTree, and the correct classification of Tweets for these algorithms are 82.58, 80.95, 78.89, and 66.35, respectively.

Culotta et al. (2015) showed that Twitter follower information is a strong data source for performing demographic inference. From six days' worth of data (December 6–12, 2013), authors randomly sampled 1000 profiles, categorized them by analyzing the profile, Tweets and profile image for each user and categorized 770 Twitter profiles into one of four ethnicities (Asian, African–American, Hispanic, and Caucasian). The authors fit a regression model to predict the demographics using information about the followers of each website in Twitter. The approach outperforms a fully supervised method for gender classification, and it is competitive for ethnicity classification. The authors found that the identities of only 10 randomly chosen followed accounts per user are sufficient to achieve 90% of the accuracy obtained using 200 followed accounts.

In the study of Hecht et al. (2011), Twitter user profiles are reviewed for user behavior relative to the location field. A total of 34% of users do not provide their true location information, in which the user should write the place where he or she is organically located. The manner in which data from Twitter is collected, as well as the characterization work and effects and machine learning work are explained in the present study. This work is descriptive and does not address causal explanations for the natural behavior of users. A total of 62 million Tweets are collected and set in English using Ling Pipe in text Classification 3 and two-stage combination of Google Language. Explicit user behaviors are shown by implicit position sharing behavior. User and country statuses are estimated with a classifier of user's Tweets. The methodology behind the creation of training sets for the classifier is highlighted, and how we divided a subset of these data for validation purposes is explained.

A web-based demo in the study of Kong et al. (2014) presents the large-scale user location estimation system (SPOT), which showcases different location estimating models on real-world data sets. The demo shows three different location estimation algorithms: friend-based, social closeness-based, and energy and local social coefficient-based. The first algorithm is a baseline, and the other two new algorithms utilize social closeness information traditionally treated as binary friendship. The two algorithms are based on the premise that friends are different and that close friends can help estimate location better. The demo shows that all three algorithms benefit from confidence-based iteration method, and provides two data sets: Twitter (148,860 located users) and Gowalla (99,563 located users).

The SPOT system uses several location estimating models on different data sets, and demonstrates both estimation accuracy and selected samples to help users compare these models under various settings and observe the estimation process clearly. We present statistics such as the average error distance and average number of friends.

In the studies of Zhang et al. (2011), Bothos et al. (2010), Tumasjan et al. (2010), Golbeck et al. (2011), De Choudhury et al. (2013), Lazarus et al. (2011), Volkova and Yarowsky (2014), many other methods and studies have predicted or determined other features such as predicting the stock market and future events, estimate election results, user information and personalities from their public data like personality traits and privacy concerns, determine diseases including depression and flu trends and identify gender.

Data gathering

The data gathering procedures and steps are described in this section. Moreover, a set of figures and tables are included to present our collected data and their properties in detail.

Kenime

Once the public data sets of Twitter and other social media websites were restricted, collecting the suitable and necessary data set from the internet became one of the most important issues in the big data research community. The collection of a big data set with convenient properties is a significant step to correctly predict and extract unknown information in any prediction system. The use of free and available

public information grabbers takes a long time and waste great efforts in cleaning and managing the collected data. Therefore, the utilization of Twitter application programming interface (API) is preferred (■). Twitter API is a package (■) developed by Twitter to allow developers and companies to interact with data including Tweets, user information, and several other attributes. To use Twitter API easily, a scripting language such as Perl or Python must be used to send data request to Twitter servers and then receive and gather the returned data in the required shape. Different public packages or tools are used to simplify the data gathering process. In this research, a quick and easy to use software is employed to connect to the Twitter API request the necessary information and gather them through KNIME software (■). This software is built for rapid and intuitive access to advanced data science, which helps organizations and companies make effective and fast data collection and analysis. The KNIME Analytics Platform, which is designed for discovering potential hidden data through data mining or predicting new futures, is one of the leading open solutions for data-driven problems. The KNIME software includes many important and widely used features such as data gathering and analysis, feature extraction, pattern classification, and clustering techniques. The software contains several nodes that can be connected to one another to perform sequential tasks on a data set, in which each task starts only after finishing the previous one. Figure 1 shows a part of the KNIME project from our implemented workspace that uses numerous nodes to search in Twitter public data, collect the results and then handle it using different 'Row Filter' nodes.

Data set collection

The Twitter data set is collected in this paper, which includes the Twitter information of 129,234 users using the tools described in the previous subsection. The collected information for each Twitter user includes id, user name, location, numbers of friends and followers, time zone, description, and languages. These data are gathered with different convenient keywords for each target country. We targeted five countries to ensure a diverse data set. Table 1 shows the countries and words used in each to gather random users.

Users are collected from five countries using different keywords through the KNIME search node, as shown in Table 1. It is preferable to use the main language in the targeted country to obtain the best results. For example, we used Arabic for Saudi Arabia and Turkish for Turkey. Figure 2 shows the number of collected Twitter users from each country.

The R node of the KNIME software is used to collect information from friends and followers of each Twitter user because our algorithm methodology depends on the information of the user and that of his or her friends and followers. We can execute R scripts within the KNIME environment. Therefore, a small R code is written to retrieve the information of the friends and followers of each Twitter user.

Proposed algorithm for location detection

The current paper proposes an efficient algorithm to predict the location country of Twitter users using their public available information (friends and followers' data). The proposed algorithm does not use GPS information which can accurately determine the location of Twitter users, but unfortunately, many people deactivate the GPS feature from their Twitter user setting. Therefore, the proposed algorithm employs all available public data to enhance the performance of location detection. The features

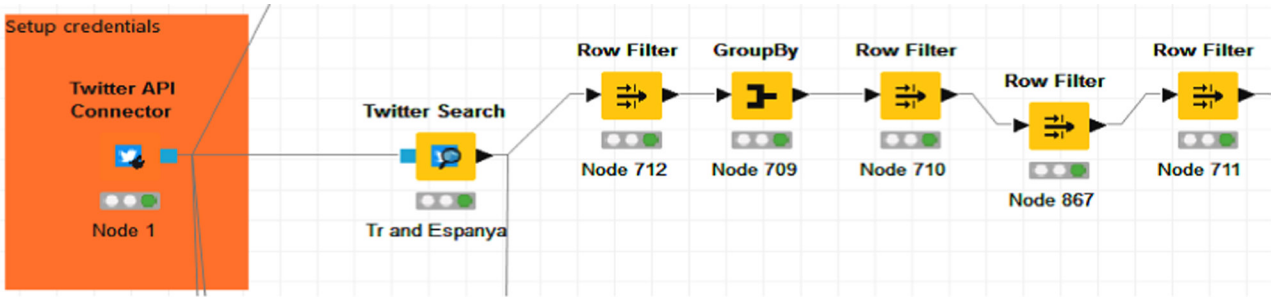

Figure 1: A series of KNIME nodes that used in our data gathering and analyzing processes.

Table 1. The used words in each country.

Country	Sample keyword for search
USA	'USA', 'health'
Spain	'Spain', 'moda'
Turkey	'Turkey', 'moda'
France	'France', 'paris'
Saudi Arabia	,السعودية, ,عربي,

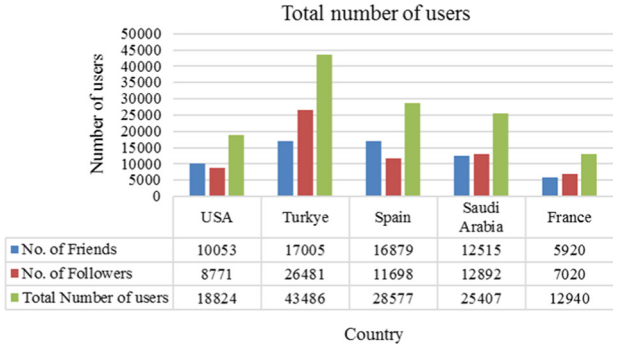

Figure 2: The total number of users.

incorporated in our detection algorithm are summarized in the following points:

1. Twitter user location;
2. Twitter user time zone;
3. Twitter user language;
4. Language of all friends of the considered Twitter user;
5. Language of all followers of the considered Twitter user;
6. Location of all friends of the considered Twitter user; and
7. Location of all followers of the considered Twitter user.

According to previous research (Abbas et al., 2017), most Twitter users do not write meaningful information on their social media accounts, making it difficult to depend on one field of information such as user location. Therefore, all previously listed features are integrated to predict the location of any Twitter user. We will not consider the information regarding the location of the considered Twitter user because we assume that this location is either unknown or wrongly written. It is important to mention that the proposed algorithm can find the home country that the considered user belongs to and spends most of his life in. Therefore, we do not consider the short period locations where the user spent some days in it and leave them quickly. The following steps summarize the procedures of our proposed location prediction algorithm.

Assuming that our proposed algorithm aims to predict the location of the considered Twitter user (denoted by TU); friends of TU are denoted by Fri, where *i* is between 1 and *N* (number of TU friends); followers of TU are denoted by FLi, where, *i* is between 1 and *M* (number of TU followers):

- Step 1: Collect public information of the considered Twitter user TU which includes the text written in his Location (TU_{loc}), Time zone (TU_{tz}), and Language (TU_{lang}) fields.
- Step 2: Gather the language fields of the friends of TU, denoted by LG_{FR1}, LG_{FR2},…, LG_{FRN}.
- Step 3: Collect the language fields of the followers of TU, denoted by LG_{FL1}, LG_{FL2}, …, LG_{FLM}.
- Step 4: Gather the location fields of the friends of TU, denoted by L_{FR1}, L_{FR2}, …, L_{FRN}.
- Step 5: Collect the location fields of the followers of TU, denoted by L_{FL1}, L_{FL2}, …, L_{FLM}.
- Step 6: Compute the most used languages for the friends of TU using the following equation:

$$\text{Friends Language Factor (FrLF)} = \frac{1}{N}\sum_{K=0}^{N} LG_k,$$
$$\text{where } LG_k = 0 \text{ or } 1,$$

LG_k indicates if the considered language is used for this friend or not. We compute this value for each language used in the TU friends in which the highest value is selected.

- Step 7: Compute the most used languages for the followers of TU using the following equation:

$$\text{Followers Language Factor (FLLF)} = \frac{1}{M}\sum_{K=0}^{M} LG_k,$$
$$\text{where } LG_k = 0 \text{ or } 1,$$

LG_k indicates if the considered language is used for these followers or not. We compute this value for each language used in the TU followers in which the highest value is selected.

- Step 8: Determine the most used location for each friend of TU utilizing the same method

Table 2. Samples of location keywords that used to classify the countries of Twitter users.

Country	Sample keywords
USA	USA – Miami – Los Angeles – California – Chicago – Houston
France	France – Landau –Melnibone – Bordeaux – Tours – Lyon – Paris – Nice
Saudi Arabia	Saudi Arabia – Dammam – جدة – مكة – القصيم ـ السعودية ـ الرياض
Turkey	Turkey – Istanbul – Izmir – Samsun – Adana – Antalya – Ankara
Spain	Spain – Barcelona – Madrid – Agitando – Granada – Barna

described in Abbas et al. (2017). To use this method, we should employ some of the keywords shown in Table 3. The value FrLoc, which indicates how many friends seem to write the same location as the user TU, is then computed, as described in the next equation (Table 2):

$$\text{Frloc} = \frac{1}{N}\sum_{K=0}^{N} L_k \cdots (1) \text{ where } L_k = 0 \text{ or } 1.$$

- Step 9: Repeat Step 7 for followers of TU to compute the most used location for each follower and to denote the resulted value as FLLoc.
- Step 10: Find the location of the Twitter user TU with the following procedures:
  - According to the values of FrLoc and FLLoc for the friends and followers of user TU, select the first three countries with the highest values.
  - Select the highest three most used languages for user TU utilizing the computed values of FrLG and FLLG.
  - Choose the country w10ith the highest sum of factors as the home location of the user TU.
  - If two countries have close values, then use the Location (TU_{loc}), Time zone (TU_{tz}), and Language (TU_{lang}) fields of TU user to select the closest location to the Twitter user.

An example is provided to further explain the proposed algorithm. If the computed values of the user are as shown in Table 3, then the user location will be considered as Turkey because it obtained the highest sum. The strong point in our proposed algorithm is it uses more than one field of information, thereby ensuring the correctness of the predicted country and decreasing the error rate.

Experiments and discussion

A data set of Twitter users was downloaded to investigate the performance of our proposed location detection algorithm, as shown in the third section. The collected data set includes all required information such as the public information of the Twitter user and that of his or her followers and friends. In the first experiment, we examine the impact between using the information of friends and followers by exploiting the location and language fields for each Twitter user. Figure 3 shows the values of FrLoc and FLLoc for the friends and followers of 20 randomly selected Twitter users from five different selected countries. The figure shows some interesting notes regarding each country. First, the information of friends is more valuable than that of the followers because friends are selected by the user and can directly indicate his or her knowledge and directions, which provides us

Table 3. Example for determining the best predicted country using proposed algorithm.

Country	FrLoc	FLLoc	FrLG	FLLG	Sum
Turkey	0.4	0.3	0.5	0.4	1.6
USA	0.2	0.3	0.3	0.4	1.2
Spain	0.1	0.2	0.2	0.1	0.6

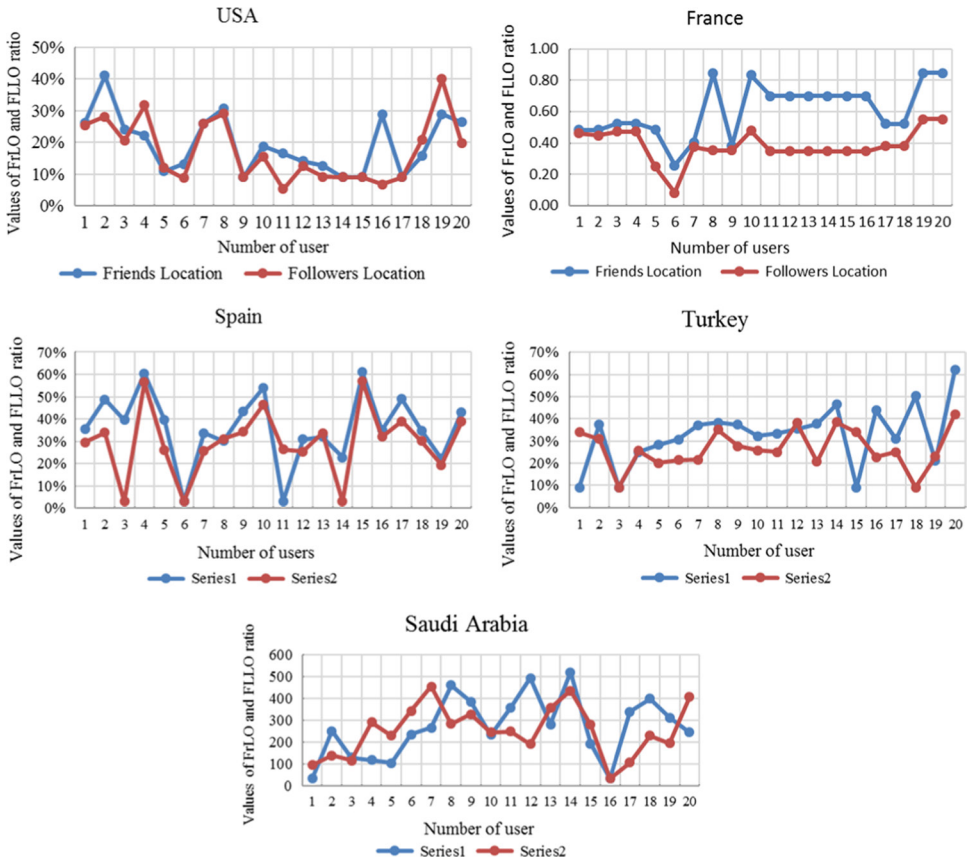

Figure 3: Samples of values for FrLoc (friends location) and FLLoc (followers location) of the friends and followers of 20 randomly Twitter users from five different countries.

with accurate information regarding the location of the Twitter user. Second, the figure of each country is different than the others given that the users of each country have their own habits and traditions strongly raised in the figures. For example, users of France, Turkey, and Spain usually write correct information in the location field, making the value of FrLoc between 0.4 and 0.5 better compared with the FrLoc of USA and Saudi Arabia between 0.2 and 0.3 in average.

Figure 4 indicates the values of FrLG and FLLG for the friends and followers of another 20 randomly selected Twitter users from five different selected countries. The FrLG and FLLG values differed from that of FrLoc and FLLoc considering that they are high because writing the language is easier than writing the countries, given the limited writing language options. In Figure 4, most of the FrLG and FLLG are between 0.7 and 1 which are good results, but it means that we can only use the languages of the friends and followers to predict the location of the considered Twitter user. Unfortunately, even if the values of FrLG and FLLG are high, then we cannot depend on them because most users in some countries prefer to write another language other than their own on the Twitter language field.

For example, we can see from the results that people from Saudi Arabia prefer to write English ‘en’ than their main language Arabic ‘ar.’ Furthermore, results in Table 5 show that the languages of followers and friends have approximately the same impact on the location of the considered Twitter user. Dissimilarity between the values of FrLG and FLLG over different countries is presented in Figure 3.

In the second experiment, we applied the proposed location detection algorithm on the collected data set presented in the third section. Users of each country were collected in separate groups and studied and analyzed individually. We eliminated the accounts of companies and institutions from our data set by excluding users with numerous followers or friends. Moreover, Twitter users with small number of friends were also excluded because they may cause false accept errors and are unsuitable for our algorithm as described in other research (Abbas

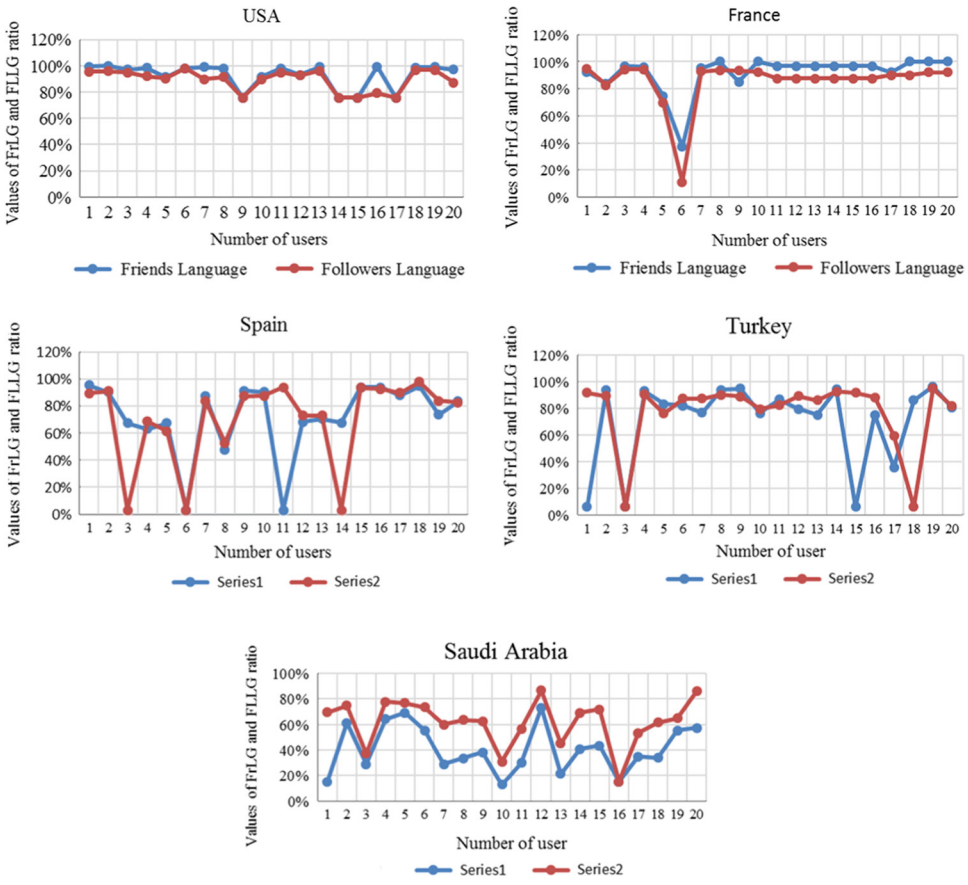

Figure 4: Samples of values for FrLG (friends language) and FLLG (followers language) of the friends and followers of 20 randomly Twitter users from five different countries.

et al., 2017). Table 4 shows the results of applying our location detection algorithm on Twitter users from five countries, which are USA, Turkey, Spain, Saudi Arabia, and France. Findings show that our algorithm can accurately detect the true location of Twitter users from different countries with the highest detection value from Turkey users because most of them write their country name and language correctly. The second best result is from France followed by Spain and USA. In addition, results indicate that users from Saudi Arabia have relatively low detection rates because most of them write the wrong location and main language in the Twitter public field.

Table 5 compares our algorithm and previously proposed algorithms for detecting the location of Twitter users. The proposed algorithm clearly benefits from using as much public information as possible (self-related information and that of friends and followers) and achieves the highest performance value over all other algorithms. Furthermore, our proposed algorithm compared the highest number of countries during testing and training processes. Note that the performance of our algorithm at 92.8 is the average value that can reach a success rate of more than 98% in some countries, as shown in Table 4.

Table 4. Comparison between the accuracy of the proposed algorithm in different countries.

Country	Accuracy
USA	90%
Turkey	98%
Spain	94%
Saudi Arabia	86%
France	96%

Table 5. Comparison between the accuracy of the proposed algorithm and previous algorithms.

Algorithm	No. of country	Accuracy
Huang et al. (2014)	1	83.8%
Culotta et al. (2015)	1	90%
Abbas et al. (2017)	4	90%
Proposed	5	92.8%

Conclusion

In this study, a new efficient and robust algorithm was proposed to predict the location of the Twitter user from the available public data on his or her account. The new proposed algorithm uses the public information of the Twitter user, such as time zone and language, as well as the information of his or her friends and followers. No GPS information was used in this paper as many Twitter users disable this feature due to privacy concerns, thereby making it difficult to depend on it. To test our proposed algorithm, a large data set was downloaded and gathered from the Twitter site from five different countries using Twitter API and KNIME software. Results show that our algorithm is efficient and robust against wrongly written and meaningless information written by many Twitter users. In addition, our proposed algorithm shows highly competitive performance compared with the location detection algorithms in the literature.

Literature Cited

Abbas, A. K., Bayat, O. and Nuri Ucan, O. 2017. Estimation of Twitter user's nationality based on friends and follower's information. *Computers & Electrical Engineering*, 66: 517–530, doi: 10.1016/j.compeleceng.2017.06.033.

Anber, H., Salah, A. and El-aziz, A. A. 2016. A literature review on Twitter data analysis. *International Journal of Computer and Electrical Engineering* 8(3): 241–249, doi: 10.17706/ijcee.2016.8.3.241-249.

Atefeh, F. and Khreich, W. 2015. A survey of techniques for event detection in Twitter. *Computational Intelligence* 31: 132–164, doi: 10.1111/coin.12017.

Bothos, E., Apostolou, D. and Mentzas, G. 2010. Using social media to predict future events with agent-based markets. *IEEE Intelligent Systems* 25(6): 50–58.

Brown, D. E., Abbasi, A. and Lau, R. Y. K. 2015. Predictive analytics. *IEEE Intelligent Systems* 30(2): 6–8.

Chang, H. -W., Lee, D., Eltaher, M. and Lee, J. 2012. @ Phillies Tweeting from Philly? Predicting Twitter user locations with spatial word usage. pp. 111-118, doi: 10.1109/ASONAM.2012.29.

Culotta, A., Ravi, N. K. and Cutler, J. 2015. Predicting the demographics of Twitter users from website traffic data. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 72–78.

De Choudhury, M., Gamon, M., Counts, S. and Horvitz, E. 2013. Predicting depression via social media. ICWSM, 13: 1–10.

Gentry, J. 2015. R Based Twitter Client Description Provides an interface to the Twitter web API, available at: http://lists.hexdump.org/listinfo.cgi/twitter-users-hexdump.org.

Golbeck, J., Robles, C. and Turner, K. 2011. Predicting personality with social media. Conference on Human Factors in Computing Systems – Proceedings, 253–262, doi: 10.1145/1979742.1979614.

Hecht, B., Hong, L., Suh, B. and Chi, E. H. 2011. Tweets from Justin Bieber's heart: the dynamics of the "Location" field in user profiles. CHI '11 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, May 7–12, 237–246.

Huang, W., Weber, I. and Vieweg, S. 2014. Inferring nationalities of Twitter users and studying inter-national linking. Proceedings of the 25th ACM Conference on Hypertext and Social Media, 237–242.

Jayanthi, N., Babu, B. V. and Rao, N. S. 2017. Survey on clinical prediction models for diabetes prediction. *Journal of Big Data*, 4(26), doi: 10.1186/s40537-017-0082-7.

Kong, L., Liu, Z. and Huang, Y. 2014. SPOT: locating social media users based on social network context. *Proceedings of the VLDB Endowment* 7: 1681–1684.

Lazarus, R., Achrekar, H., Lazarus, R. and Park, W. C. 2011. Predicting flu trends using Twitter data. IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS, 702–707, doi: 10.1109/INFCOMW.2011.5928903.

Liu, H. 2013. Some computational challenges in mining social media. 2013 IEEE/ACM International Conference in Advances in Social Networks Analysis and Mining (ASONAM), Niagara Falls, August 25-28, doi: 10.1109/ASONAM.2013.6785672.

Mcgee, J., Caverlee, J. and Cheng, Z. 2013. Location prediction in social media based on tie strength. *CIKM '13 Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 459–468.

Package based Twitter client description provides an interface to the Twitter web API. Version 1.1.8, February 20, 2015.

Tsujioka, S., Kondo, A. and Watanabe, K. 2016. Estimation of residence information of Twitter users based on their posted messages: data for tourism development. *International Journal of Research in Chemical, Metallurgical and Civil Engineering* 3(1): 180–183.

Tufekci, Z. 2014. Big questions for social media big data: representativeness, validity and other methodological pitfalls. ICWSM 14, 505–514.

Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welpe, I. M. 2010. Predicting elections with Twitter: what 140 characters reveal about political sentiment. Conference: Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM, Washington, DC, May 23-26, 178–185.

Volkova, S. and Yarowsky, D. (2014), "Improving gender prediction of social media users via weighted annotator rationales", In NIPS 2014 Workshop on Personalization, San Diego, CA, December 8–13.

Zhang, X., Fuehres, H. and Gloor, P. A. 2011. "Predicting stock market indicators through Twitter "I hope it is not as bad as I fear", *Procedia – Social and Behavioral Sciences* 26: 55–62.