Identifying Influencers by Countries in Twittersphere*

Rong Pan[1], Guanqi Fang[1], and Aleksey Panasyuk[2,3]

[1] Arizona State University, Tempe AZ 85226, USA
{rong.pan, gfang5}@asu.edu
[2] Air Force Research Laboratory, Rome NY, USA
aleksey.panasyuk@us.af.mil
[3] University of Syracuse, Syracuse NY, USA
apanasyu@syr.edu

Abstract. Twitter user location detection is a research problem interested by many agencies for various purposes such as marketing, emergency management, counter terrorists, etc. In this study, we focus on identifying local influencers for each country and using machine-learning methods to classify countries for a potential influencer. We develop a novel approach that utilizes the publicly available account data of followers to infer the country being influenced by an influencer. On the datasets we collected, this approach has been shown to be very accurate. It is foreseen that the classifier models we developed can be applied on vast twitter users to enhance the geocoding of user location and the ranking of local influencers.

Keywords: Social media influencer · Twitter user network · Decision trees.

1 Introduction

Twitter is one of the most popular online social networking platforms worldwide. On Twitter, people communicate in short messages called tweets, and sending out tweets, or tweeting, is a mass broadcasting activity, because anyone who can access the twitter website or use a twitter app will be able to view any twitter user's tweets as long as he/she knows the handler or screenname of this user. A few countries block Twitter service (e.g., China, Iran, North Korea and Turkmenistan); however, people in these countries still can access Twitter by using VPNs or other backdoor internet connections. Therefore, many individuals, companies and organizations are using this platform to promote their messages and products, and some governments and nation-state players are using it to campaign their policies and political propaganda. Note that the a screenname on Twitter may refer to a real person or a real entity, but it could also be a fictional character, such as Homer J. Simpson, or even a computer algorithm or

* This research was supported by Air Force Research Laboratory at Rome, New York.

social bot, which is programmed to automatically broadcast status and converse with other twitter users. In fact, in a recent SEC filing Twitter admitted that up to approximately 8.5% of the accounts it considers active are automatically updated without any discernible additional user-initiated action. According to Pew Research Center, two-thirds of all tweeted links are shared by suspected bots and a small number of highly active bots are responsible for a large share of links to prominent news and media sites.

Twitter accounts are connected by a followers-friends mechanism. If account A befriends with account B, then the tweets posted by account B will be automatically pushed to account A, so A is one of B's followers. An influential account is an account that has a lot of followers. Identifying these influential accounts is important for various purposes. For example, a company wants to promote a product on social media, then it may try to recruit those influential accounts on specific market segments to broadcast its product commercial. Similarly, a government or nation-state player may want to monitor an influencer's online activities for any social or political reasons. In this research we are interested in knowing which accounts are influential and to which countries these accounts are influencing. Note that these two questions are intertwined, in the sense that an account that has a huge following globally could be just a celebrity account (e.g., a pop singer or sport star) and its followers do not concentrate in a particular country. Our research interest is on identifying (country-specific) local influencing accounts, instead of global celebrity accounts. It requires the information of which country most followers locate in and we will apply machine-learning methods on account profile data to acquire this knowledge.

The research presented in this paper extends previous studies on classifying twitter user home location (see, e.g.,[1], [2]) and it is also related to the research of location-aware influence maximization (LAIM) problem. To our knowledge, it is the first time that some unique and always available features of user account data, such as account creation time, are used for location inference. In addition, we validate our datasets by checking with a social media service provider, and explore the user-follower network, demonstrate the usefulness of network structure for location inference. The classifiers trained by our initial datasets can be applied on a much larger user-follower datasets that had been collected, thus an in-depth knowledge of country-wise influencer's characteristics and influencing segment distributions can be obtained, which are invaluable to many social, economical and behavioral applications.

The rest of the paper is organized as follows: The twitter user location problem is introduced in the next section and we review the existing literature that had discussed this problem. In Section 3 we describe how the data used in this research were collected, then an important feature, twitter account creation time, is investigated in Section 4. Next, we study several machine-learning algorithms and compare their performance on classifying countries based on account profile data. Finally, conclusions and future studies are drawn.

2 Literature Review

Location inference for tweets and twitter users has gathered a lot of attentions from academia and government in recent years for various reasons. In 2015, DARPA held a 4-week competition, in which multiple teams supported by the DARPA Social Media in Strategic Communications program competed to identify a set of influencing social bots in twitter. Geo-tag was one of the features in the dataset and tweet/user networks had been shown to be very important in detecting social bots [3]. In 2016, a shared task of the 2nd Workshop on Noisy User-generated Text called for geotagging over 1 million twitter users across 3362 metropolitan cities. The best performance was achieved by systematically using metadata embedded in tweets, which confirmed that metadata contains abundant high-quality location information and pure text-based methods are unable to provide a satisfactory result [4] [5] [6] [7].

There are several excellent survey papers of previous twitter geolocation inference studies in literature [1], [2], [8]. In particular, Zheng et al. [2] categorize the location prediction problem on Twitter into three categories – the user home location prediction problem, the tweet location prediction problem, and the mentioned location prediction problem. For each of these problems, the paper provides a comprehensive review of the inference methods, data sources, evaluation metrics and performance comparisons that have been discussed in previous studies. It compares the location inference based on tweet content, twitter network and tweet context. As in the influencer-follower database we created there are only the user profile and user network information, we may use them to infer a user's home location, but not a tweeting location.

Zheng et al. [2] shows that most earlier studies used only content, i.e., the appearance of country or city names in some fields in tweets or user profiles, for location inference; however, more recent studies began to pay attention to context and user network derived from twitter queries and combine them with content. After 2017, there have been more tools being proposed for pinpointing a user's exact location. For example, Hemamalini [9] proposed to take advantage of Google goecoding API to obtain a location's latitude and longitude. Similarly, Google geocoding API was used in Panasyuk et al. [10] for geocoding Twitter users' self-declared home locations too. The authors developed a heuristic to utilize the additional parameters returned from Google geocoder output to further process location texts and to determine city-level addresses. This is needed because the location texts posted by users are often incomplete, ambiguous and erroneous, sometimes even intentionally misleading.

Beside of additional external localization tools, more sophisticated data modeling and machine-learning techniques have been applied on these localization problems too. Ozdikis [11] presented a locality-adapted kernel density estimation method. This study utilized the textual features extracted from tweet contents only. Similarly, Xu [12] developed a deep neural network-based pipeline to recognize fine-grained location mentions in tweets and linked the recognized locations to location profiles.[13] proposed a semi-supervised factor graph model (SSFGM), in which a probabilistic framework that can combine various sources

of information (e.g., content and social network) together is established and an approximate learning algorithm, based on softmax regression, is developed. Ghaffari [14] also made the use of metadata of tweets to find home locations with high resolution for a subset of users, with high accuracy in the prediction. Luceri [15] employed a graph-based deep learning architecture to learn a model between the users' known and unknown geo-location during a considered period of time. They analyzed the dataset of twitter messages as a study case, because in tweets the user generated content (i.e., tweet) can embed user's current location. In [16] the authors introduced a community-based approach for inferring a user's geographic location. Here, the communities are detected by Infomap, then the geographical proximity and structural proximity metrics of these communities in the ego-net of a user are calculated and evaluated for their effectiveness on predicting the user location.

As to compare geolocation evaluation metrics, [17] presents a comparison study of several Twitter user geolocation models. These models are evaluated using ten metrics over four geographic granularities. Al Hasan Halder [18] provides an in-depth empirical comparison of eight representative prediction models using five metrics on four real-world large-scale datasets, including Twitter.

3 Data Collection

Through Twitter's free APIs, we are able to build a large database that consists of the user profiles of twitter influencers and their followers. Twitter maintains a special label that tracks verified users, @verified. As a starting point, we treat these @verified accounts as influencer accounts and, by querying the followers of @verified accounts, we obtain a collection of follower accounts. In addition, applying the local community detection method illustrated in [19], we collect a set of local influencers (note that these accounts may not be @verified) and their followers. Overall, this influencer-follower database consists of 322 thousand influencers and 377 million followers. Considering that Twitter has reported there were about 321 million active monthly users, this database covers almost entire users in twittersphere. However, due to the limitation of Twitter free API, only up to 5000 follower id's can be retrieved. (More followers can be retrieved, but it would be time costly. A sample of first 5000 followers for each influencer is large enough for our study.) All data are stored in a Mongo database.

3.1 Twitter Account Data

In our database, Twitter account data, no matter it is from an influencer or a follower, are stored. For example, one account is associated with a user named "Ellen Anderson" and this account is not verified. This user follows 2 twitter accounts and is followed by other 5 followers, the default language is English, the location field is not filled, and the creation time of this account is November 29, 2016 at 20:42:34. The time recorded is UTC (Universal Time Coordinated or Greenwich Mean Time, GMT) time. If we assume this user is living in New

York, US, then the local time is 16:42:34 as there is a -4 hours time offset. Understanding this time offset is important, because later we will utilize the account creation time to infer location. Also it is found that the time_zone field is empty. This is because, due to privacy concern, after 2017 Twitter had made this field private.

The home location information is not required for creating a twitter account, unless a user voluntarily fills in the information, which is available for less than 10% of all users. Sometimes, in the description field a user may write down some words related to his/her home location; however, this does not happen often. So, in this research, we try to utilize some auxiliary data that were automatically filled in by the system when a user signed up its twitter account to discern the location information. These auxiliary data include account creation time (converted to GMT in the 24-hour format) and account language. Machine learning techniques are applied on these data to classify the geographical location down to the country level.

4 Account Creation Time

We analyze the account creation time feature of the followers corresponding to an influencer. As some twitter users' home locations have been specified by the user themselves or revealed in the description field. We extract the information from the database of 377M followers and create a spreadsheet of country/city, the time zone of country/city, the number of records,and the number of accounts created during each one hour interval in the UTC 24-hour period.

4.1 Initial Data Analysis

When plotting the number of account creation times in the 24-hour cycle for each of these time zones, a clear sinusoidal pattern appears. This is because, for any user living in a certain time zone, it is unlikely that the user would create its twitter account in the hours of early morning when they are commonly the hours of sleep time. Therefore, features from these sinusoidal patterns can help us to identify the most likely time zone of followers.

We standardize the number of account creation for each time zone by the following equation:

$$Percentage\ at\ UTC\ Hour\ i = \frac{No.\ of\ Account\ Creation\ at\ UTC\ Hour\ i}{Total\ No.\ of\ Account\ Creation\ across\ all\ UTC\ Hours} \quad (1)$$

The average of percentages at UTC hour i in the same time zone is calculated and the curve of mean percentage of account creation time is shown in Fig. 4.1 for 12 time zones.

4.2 Local vs. Global Influencers

We keep the data of 247 time zones out of 339 time zones in total because in those time zones there are at least 240 creation time records. We treat these

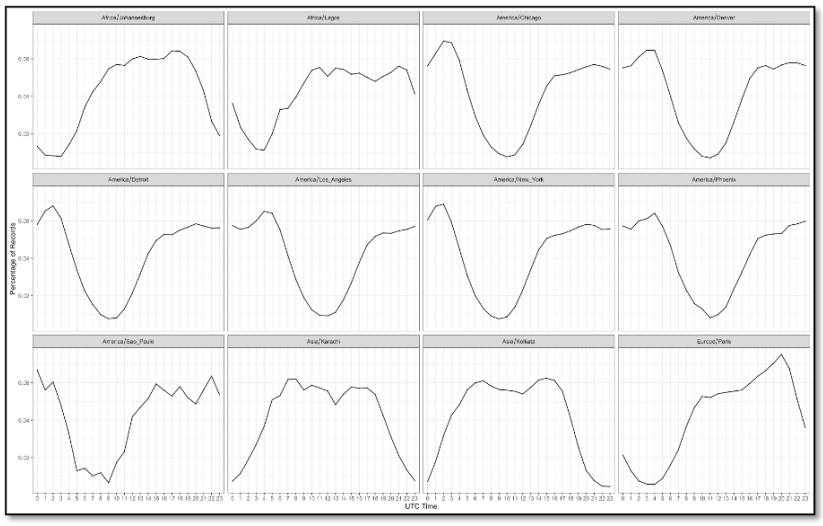

Fig. 1. Line plots of mean percentage of account creations over UTC time for 12 time zones.

account creation time curves of 247 time zones as the baseline data for locating a user given its followers' account creation times. Furthermore, we calculate the ranges and standard deviations of these 247 curves and plot the histogram of range/standard deviation. For an individual influencer, if its influencing region is globally spread, then its followers' account creation time curve should be flat and so the range/standard deviation would become smaller, falling into the lower side of the histogram. Therefore, we may use the 3-sigma lower limit on the histogram as a criterion to judge whether or not an influencer is a global or local influencer. Figure 4.2 provides these histograms and the 3-sigma lower limits. Figure 4.2 gives the followers' creation time curves of 12 influencer accounts with smallest values of ranges. By examining these accounts manually, we can deem these accounts to be global accounts. For example, the twitter account "goproANZ" is the official Australian and New Zealand GoPro Twitter. This account posts many GoPro extreme sport action photos, thus attracts GoPro enthusiasts around the globe.

From the above analysis one can see that the curves of followers' account creation times certainly contain the information of the time zone of an influenced region for a local influencer. Utilizing the nadir point of the curve and the time window of valley should help in identifying the time zone of the region, then narrowing down to possible countries under influence. In addition, by examining the flattness of this curve for a specific infuencer we may infer whether it is a local influencer or a global influencer.

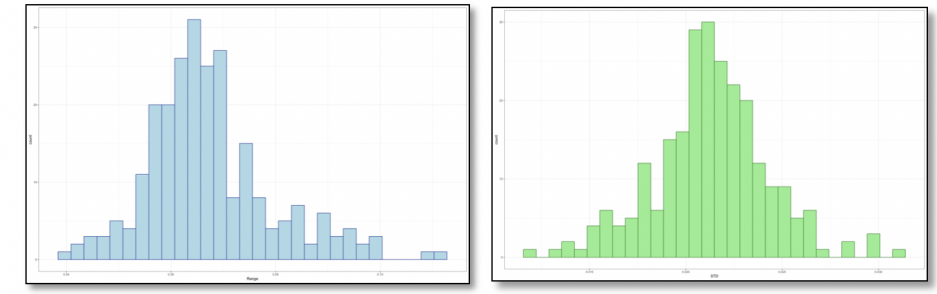

Fig. 2. Hitograms of range (left) and standard deviation (right) for 247 time zones. The range average is 0.06738 and the 3-sigma lower limit is 0.03126. The standard deviation average is 0.02130 and the 3-sigma lower limit is 0.01224.

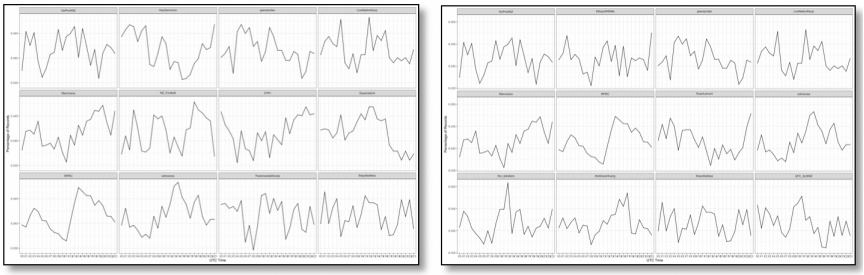

Fig. 3. Twelve influencer accounts with smaller range (left)/standard deviation (right) values. The account names in the left graph are $GoProANZ$, $HeyGeronimo$, $jasonjordan$, $LiveNationKpop$, $Marcmena$, $NZ_Football$, $pnau$, $Queensland$, $RPRC$, $sdmoores$, $TheAmandaWoods$, $thepotbelleez$. The account names in the right graph are $GoProANZ$, $ItStopsWithMe$, $jasonjordan$, $LiveNationKpop$, $Marcmena$, $RPRC$, $RyanLamont$, $sdmoores$, $thedubsters$, $theGraceHuang$, $thepotbelleez$, UFC_AUSNZ.

4.3 Verification with SocialBakers

To generate a list of local influencers for each country, we first screen all followers and determine their countries by examine their accounts' location field and description field (see ref. [10]). Then, for all influencers, the distribution of their followers' country names is computed and we choose only the influencers whose followers concentrate in one country ($> 50\%$). Note that as aforementioned, many twitter users do not provide location information and give no description to their twitter accounts. Thus, only a small subset of local influencers is obtained for each country. Nevertheless, this dataset can be used for building a successful country classifier, which can then be applied on any larger dataset later on.

SocialBakers is a social media marketing service provider. It uses a proprietary algorithm to provide the top influencer list on Twitter for each country. To validate the top influencer list we obtained, we compare our list with the top 100 influencers obtained from SocialBakers and identify the countries where our list has at least 50% overlap with the SocialBakers' list. Then, we will classify influencers for these countries.

5 Country Classification of Local Influencers

As previously described, the account creation time and language information are system generated, thus always available. For each country, we track down top influencers' followers and record the top three languages used by the followers and their distribution, as well as the followers' account creation times over a 24-hour period and their distribution. Multiple datasets are created and they are described in Table 1. These datasets are labeled training datasets and we will use them to train multiple machine learning (ML) algorithms.

Table 1. Datasets of local influencers.

Dataset	no. of countries	no. of influencers each country
Dataset 1	50	75
Dataset 2	42	100
Dataset 3	32	250
Dataset 4	17	1,000
Dataset 5	10	2,499
Dataset 6	2 (United States & Great Britain)	10,000

5.1 Machine Learning Algorithms

First of all, the account creation time data in the six training datasets are standardized per Eq (1). This results in 24 values (from Hour 0 to Hour 23 in UTC) for each influencer. From our initial analysis, it is understood that the time

of nadir point on the creation time curve is important for discriminating time zones, thus a feature, called 'MinPoint', is created to indicate the UTC time value of the nadir point. Similarly, 'MaxPoint' is created for the UTC time value of the peak point of creation time curve. In addition, we sum the four consecutive lowest values and the four consecutive highest values on the creation time curve and name them 'MinConSum' and 'MaxConSum', respectively. For the language feature, one-hot encoding is applied. In summary, there are 40 language features, 24 creation time features, and 4 augmented features that are derived from the creation time curve. Each dataset is randomly split 80-20 for training and testing purposes.

Naive Bayes Classifiers Naive Bayes (NB) classifier is a supervised classification algorithm in machine learning. This algorithm uses Bayes Theorem, which essentially concerns the conditional probability of an event given that some relevant events had already occurred. By using the basis of the Bayes theorem, the Naive Bayes Classifier formula can be written as $\pi(y|x_1,..,x_j) \propto l(x_1,\ldots,x_j|y)\pi(y)$, where $\pi(y|x_1,\ldots,x_j)$ is the posterior probability of label y given the values of features $x_1,\ldots,x_j$; $l(x_1,\ldots,x_j|y)$ is the likelihood function of these feature values if all labels are known; and $\pi(y)$ is the prior probability of labels. In the Naive Bayes classifier, we determine the classes (labels) of data points by maximizing the posterior probability. We assumed a multivariate normal distribution for the data from each class.

Tree-based Classifiers The second machine learning algorithm that we experimented with is decision tree (DT). Tree-based methods involve segmenting the prediction space into a number of simple regions. A classification tree conducts these splits to minimize the mis-classification error rate. In practice, the following two methods are used for selecting features for splitting:

Gini Index: $G = \sum_i \hat{p}_i(1-\hat{p}_i)$; Cross Entropy: $D = -\sum_i \hat{p}_i \log(\hat{p}_i)$
where $\hat{p}_i$ is the percentage of correct classification for class i. We implemented the entropy splitting criterion in our study and experimented the depth of decision tree from 1 to 10, and eventually decided to use the depth of 8. We implemented both single tree classifiers and random forest (RF) classifiers. Random forest is an ensemble method, in which the algorithm generates B different bootstrap training sets and all these training sets are used to build classification trees and their prediction results are summarized by majority votes. To decorrelate these trees. we used a random sample of m predictors from all p features at each splitting step. In this study we used the default m value, which is the square root of p.

Autoencoder and SVM Autoencoder (AE) is an unsupervised learning technique in which neural networks are leveraged for the task of representation learning. In previous tasks features, such as 'MinPoint' and 'MinConSum', were handcrafted based on our understanding of creation time curves, thus we wonder if an

autoencoder can help in automatically identifying useful features. In this task we used autoencoder to generate 11 auto-features, and then we applied a support vector machine (SVM) classifier on these features.

5.2 Cross-dataset Validation

The testing performance of the above-mentioned ML algorithms on the six datasets are given in Table 2. One can see that generally all of these algorithms have high accuracy, particularly when there are a small number of countries. This is because the size of training data for each class increases when the number of countries decreases. The testing results also indicate that the features extracted from creation time curves and languages are extremely useful for discriminating countries. Overall, RF provides the best performance.

Table 2. Comparison of classification performance of four machine learning algorithms.

Dataset	NB	DT	RF	AE-SVM
Dataset 1	**0.9227**	0.8533	0.9093	0.6107
Dataset 2	0.9167	0.8381	**0.9262**	0.6605
Dataset 3	0.9181	0.8713	**0.9663**	0.7258
Dataset 4	0.9824	0.9879	**0.9918**	0.7954
Dataset 5	0.9964	0.9992	**0.9994**	0.9289
Dataset 6	0.9995	**1.0000**	**1.0000**	0.9999

We are also interested in the performance of a classifier trained by a dataset with more country labels and tested by a different dataset with less country labels; that is, whether or not a trained classifier can be transferred to another scenario. We call it cross testing. Note that the dataset with fewer countries will have more labeled data per country; therefore, this cross testing can help us to determine if the model trained by the dataset with more countries but less labeled data per country is good enough for generalization. The rationale is that when we expand the dataset to all the twitter user profiles collected, we do not know the country labels for most users; however, we are interested in knowing whether or not these users should be labeled to a particular country or a small set of countries of interest. In order to do that, we test the classifier trained by the first dataset (with 50 country labels) on the remaining datasets (42, 32, 17, 10 and 2 country labels, respectively). See Table 3. The results turn out be much worse than the previous test. In particular, for Dataset 6, which has only two country labels, the classified tends to its entry to a different country that locates in the same time zone and/or speak the same language. The misclassification is so severe that we would like to investigate the importance of each feature to a classifier and how to improve the accuracy of cross testing.

The cross testing performance of RF is better than other classifiers. For the last dataset (Dataset 6 with two country labels), if we list the top 3 countries selected by the classifier and check whether or not the true country label is

Table 3. Comparison of cross testing performance of two machine learning algorithms.

Dataset	NB	RF
Dataset 2	0.9326	**0.9707**
Dataset 3	0.8935	**0.9044**
Dataset 4	0.8399	**0.8519**
Dataset 5	0.8745	**0.8948**
Dataset 6	0.5575	**0.7095**

included in the top-3 list, we find the accuracy is improved to 0.9386. In addition, using RF, we can provide the probability of a user being labeled to a particular country. Another benefit of RF is that it is able to show which feature is more important to the classification. Figure 5.2 depicts the importance of each feature on a bar plot. As one can see that the most important feature is the English language, then it is followed by 'T3', 'T6' and 'T4', which are creation time features. Many other creation time features, such as 'T10', 'T8', 'T2', 'T11', are relatively important too. Also, the features – 'MinPoint' and 'MinConSum' – are important, while 'MaxPoint' and 'MaxConSum' are relatively less important. Most language features, except English, Spanish and several others, are not important at all.

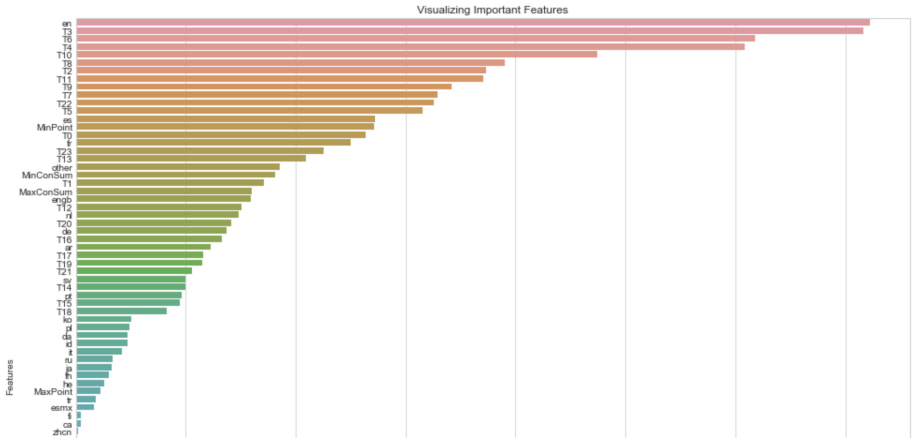

Fig. 4. Importance measures of features by RF.

To further understand why some features are selected by RF to be important, we look into three pairs of countries – USA vs. Canada, Great Britain vs. Australia, and Venezuela vs. Columbia. Table 4 summarizes the results of classification. The three most important features for the USA-Canada pair are all language features: 'en' (English), 'other' (other language) and 'fr' (France). This makes sense because both countries are in North America continent and they

Table 4. Classification performance of three pairs of countries.

Comparison	Accuracy	Top 3 Important Features
USA vs. Canada	1.0000	'en','other', 'fr'
Great Britain vs. Australia	1.0000	'MinPoint','en', 'engb'
Venezuela vs. Columbia	0.9000	'MinConSum','other', 'MaxConSum'

territories span over multiple time zones. Therefore, it is difficult to differentiate them by using account creation time features. However, many people in Canada are multilingual, speaking French beside of English, or preferring British English instead of American English. So, these language features naturally become the most important differentiators for separating these two countries. Similarly, the most important feature for the Great Britain-Australia pair is 'MinPoint', which is the nadir time point of account creation time curve, because these two countries locate on two opposite sides of the globe. It is a little bit more challenging to separate Venezuela and Columbia, as both countries speak Spanish and they are in the same continent, South America. It turns out that 'MinConSum', 'other' and 'MaxConSum' are the top-3 important features, so both creation time and language features are helpful in this case.

6 Future Work

Through this paper we have described an initial investigation of why and how twitter network and twitter follower's account profiles can be utilized for examining a social media influencer and locating the country that he/she exerts influence on. Note that the countries identified are the countries that most followers of a known influencer live in, but it is reasonable to use this information backward to infer the home location of influencer too, although we do see some examples of expats and foreigners that have a big follower crowd in a country he/she does not live in. The user home location identification problem is an outstanding problem in Twitter, particularly after Twitter made the home location field private. The traditional method uses text mining of description field to find country-city mentioned by the user, but this method is not reliable, as well as not widely applicable, because with privacy concerns many twitter users would not say their home location or may simply use some generic, or even teasing words, such as Earth, Internet, or Mars. Using the information obtained a user's followers' profiles is like generating the second-order knowledge in network analysis. This has not been explored for twitter user analysis so far. With the data we have, one challenge is that the available follower profiles are sampled from all followers up to 5000 unique accounts, thus the network structure obtained is not a complete structure, but rather a smaller, random sampled, structure. We plan to tackle this challenge in our future research.

We have explored only local influencers in this paper, but could focus on global influencers by looking at those with over 10 million followers or some similar threshold. An interesting topic is the differentiation of local and global

influencers. As mentioned in Section 4, a theoretical global influencer could have followers in any part of the world (or a larger region than a single country), so the account creation time curve of these followers should be flat. In addition, we plan to sample data from multiple countries and mix them to create artificial global influencers' creation time curves and language distributions, and then test different hypotheses regarding global influencers. This study will help us understand what features are more important for differentiating local and global influencers, and find some criteria that can be implemented online to quickly identify local influencers.

Lastly, machine learning techniques need to be studied more for this kind of applications. In this paper we have used autoencoders to automatically search features for classification; however, more efforts are needed for making an understanding of these auto-features and combining auto-features with original features for various classification tasks.

References

1. O. Ajao, J. Hong, and W. Liu. A survey of location inference techniques on twitter. *Journal of Information Science*, 1:1–10, 2015.
2. X. Zheng, J. Han, and A. Sun. A survey of location prediction on twitter. *IEEE Transactions on Knowledge and Data Engineering*, 30(9):1652–1671, 2018. https://doi.org/10.1109/TKDE.2018.2807840.
3. V. S. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer. The darpa twitter bot challenge. 2015.
4. B. Han, A. Rahimi, L. Derczynski, and T. Baldwin. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 213–217, 2016.
5. L. Chi, K. H. Lim, N. Alam, and C. J. Butler. Geolocation prediction in twitter using location indicative words and textual features. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 227–234, 2016.
6. G. Jayasinghe, B. Jin, J. McHugh, B. Robinson, and S. Wan. Csiro data61 and the wnut geo shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 218–226, 2016.
7. Y. Miura, M. Taniguchi, T. Taniguchi, and T. Ohkuma. A simple scalable neural networks based model for geolocation prediction in twitter. In *Proceedings of the 2nd Workshop on Noisy User-generated Text*, pages 235–239, 2016.
8. A. Zubiaga, A. Voss, R. Procter, M. Liakata, B. Wang, and A. Tsakalidis. Towards real-time, country-level location classification of worldwide tweets. *IEEE Transactions on Knowledge and Data Engineering*, 29(9):2053–2066, 2017. https://doi.org/10.1109/TKDE.2017.2698463.
9. S. Hemamalini, K. Kannan, and S. Pradeepa. Location prediction of twitter user based on friends and followers. *International Journal of Pure and Applied Mathematics*, 118(18):2817–2824, 2018.
10. A. Panasyuk, E. S.-L. Yu, and K. G. Mehrotra. Improving geocoding for city-level locations. In *2009 IEEE 13th International Conference on Semantic Computing (ICSC)*. IEEE, 2019. https://doi.org/10.1109/ICSC.2019.00081.

11. O. Ozdikis, H. Ramampiaro, and K. Nørvåg. Locality-adapted kernel densities of term co-occurrences for loation prediction of tweets. *preprint submitted to Journal of Information Processing and Management*, 2019.
12. C. Xu, J. Li, X. Luo, J. Pei, C. Li, and D. Ji. Dlocrl: a deep learning pipeline for fine-grained location recognition and linking in tweets. *arXiv:1901.07005v3*, 2019.
13. Y. Qian, J. Tang, Z. Yang, B. Huang, W. Wei, and K. M. Carley. A probabilistic framework for location inference from social media. *arXiv:1702.07281v3*, 2019.
14. M. Ghaffari, A. Srinivasan, and X. Liu. High-resolution home location prediction from tweets using deep learning with dynamic structure. *arXiv:1902.03111v2*, 2019.
15. L. Luceri, D. Andreoletti, and S. Giordano. Infringement of tweets geo-location privacy: an approach based on graph convolutional neural network. *arXiv:1903.11206v1*, 2019.
16. P. Wagenseller, A. Avram, E. Jiang, F. Wang, and Y. Zhao. Location prediction with communities in user ego-net in social media. In *ICC 2019 - 2019 IEEE International Conference on Communictations*. IEEE, 2019. https://doi.org/10.1109/ICC.2019.8761695.
17. A. Mourad, F. Scholer, W. Magdy, and M. Sanderson. A practical guide for the effective evaluation of twitter user geolocation. *arXiv:1907.12700v1*, 2019.
18. N. Al Hasan Haldar, J. Li, M. Reynolds, T. Sellis, and J. X. Yu. Location prediction in large-scale social networks: an in-depth benchmarking study. *The VLDB Journal*, July, 2019. https://doi.org/10.1007/s00778-019-00553-0.
19. A. Panasyuk, R. Pan, E. S.-L. Yu, and K. G. Mehrotra. Ranking influencers over a geographic area inspired by the election prediction problem. *manuscript*, 2019.