Inferring Nationalities of Twitter Users and Studying Inter-National Linking

Wenyi Huang* 
Information Sciences and Technology 
Pennsylvania State University 
University Park, PA 16802 
harrywy@gmail.com

Ingmar Weber 
Qatar Computing Research Institute 
Doha, Qatar 
iweber@qf.org.qa

Sarah Vieweg 
Qatar Computing Research Institute 
Doha, Qatar 
svieweg@qf.org.qa

ABSTRACT
Twitter user profiles contain rich information that allows researchers to infer particular attributes of users’ identities. Knowing identity attributes such as gender, age, and/or nationality are a first step in many studies which seek to describe various phenomena related to computational social science. Often, it is through such attributes that studies of social media that focus on, for example, the isolation of foreigners, become possible. However, such characteristics are not often clearly stated by Twitter users, so researchers must turn to other means to ascertain various categories of identity. In this paper, we discuss the challenge of detecting the nationality of Twitter users using rich features from their profiles. In addition, we look at the effectiveness of different features as we go about this task. For the case of a highly diverse country—Qatar—we provide a detailed network analysis with insights into user behaviors and linking preference (or the lack thereof) to other nationalities.

Categories and Subject Descriptors
Applied Computing [Law, social and behavioral sciences]: Sociology; Computing methodologies [Machine Learning]: Supervised learning by classification

Keywords
Twitter; Qatar; nationality inference; user classification

1. INTRODUCTION
The availability of large amounts of social media data has created new possibilities to study social phenomena at large scale through the lens of online behavior. To obtain insightful results and to “link” these online data to offline, “real-world” variables, it is often useful to have detailed social media user attributes such as gender, age or nationality. Though inferring gender and age from a user’s social media presence has been studied before, the latter has, to the best of our knowledge, not been explored. A likely reason for this gap is that in most countries the population is dominated by the “native” nationality. Even in the US, which is often perceived as a country of immigrants, only about 13% of the population are foreign-born1 and of those, about 45% are American citizens,2 leaving just over 7% of “foreigners.” Due to this nationality skew, even a trivial American-or-Not classifier would have an accuracy of 93% by always reporting “American.” However, in Qatar, the majority of the population is foreigners, exceeding 85%. This creates a range of challenges related to the identification of national identity. We are interested in potential correlations between national identity and social capital [2], and are curious to know if asking questions about nationality as it manifests on Twitter can lead to different or better understandings of how the two are linked (or not).

To explore this question, we created a classifier that detects the likely nationality of Twitter users based on a range of features, including language, the hashtags they use, and the geographical location of their social ties. We then perform a feature analysis, which leads to an interesting discovery regarding patterns related to the use of particular hashtags, such as #disappointed or #takeemback. These hashtags have a negative connotation, and are linked to people of particular nationalities.

Regarding a definition of a “nationality,” we take a simplified approach given our goal: you (a Twitter user) have the nationality that others (CrowdFlower workers) believe you have. We argue that this approach is acceptable for several reasons. First, challenges regarding perception of national origin and impartial treatment are known [22]. Regardless of how individuals self-identify as one nationality or another, the impression others have of them factors into conduct and actions (ibid.) Second, we also evaluated the quality of crowdsourced data on a subset of users who explicitly state their nationality in their Twitter profile as in “An American living in Doha” (details in Section 4.3). The agreement between self-stated nationality and crowdsourced labels was 91.86%, which validates the reliability of our crowdsourced data.

2. RELATED WORK
Related work involves user classification on Twitter for attributes ranging from political orientation to gender. In addition, researchers have explored the use of Twitter messages (and similar data) to study a wide range of cultural phenomena.

2.1 Twitter User Classification
Rao et al. [16] introduced the work of classifying latent user attributes including gender, age, regional origin, and political orientation using simple features such as n-gram models, presence of emoticons, number of followers/following and retweet frequency.

---

1http://www.census.gov/hch/infographics/foreign\born.html
2http://www.migrationinformation.org/datahub/\state.cfm?ID=US\%3
In [9], the authors extended this work by introducing a two phase architecture for classifying Twitter users. The first part is using basic features for classification, and the second part is to use social graph-information to update classification result. The interesting result is that the first phase alone achieves good performance which is hard to improve by the social graph information. The idea of using information from neighborhood context for classification was first introduced by Pennacchiotti and Popescu [12], Zamal, Liu, and Ruths extended the previous work by augmenting the user features and neighboring features and boosted the performance of gender, age, political orientation classifications [23]. Burger, Henderson, and Zarrella [3] conducted a study that used only text content for determining the gender of Twitter users. They also included a human performance study using Amazon Mechanical Turk and claimed that the trained model performed better than human assessment. In addition, Cheng et al. [4], Hecht et al. [7] and Mahmoud et al. [8] have looked at inferring Twitter user location based on tweet text. Unlike previous research, our study begins to explore the question of nationality detection—a new classification task.

2.2 Social Media and Social Studies

Poblete et al. [13] analyzed and compared Twitter user language, sentiment, content, and network properties in the ten most active countries. In [6], the authors study behavioral patterns on Twitter and associated them with three different cultural dimensions: pace of life, individualism, and power distance. They found that country-level behavior derived from Twitter strongly correlates with said dimensions. Santani and Gatica-Perez [17] presented an analysis of languages used in Switzerland to examine multiculturalism. They used Foursquare data, and presented a descriptive analysis of linguistic differences and similarities in multiple cities. In our analysis we also came across negative sentiments and worries. This is related to work that looks at the geographical differences in well-being and happiness [14, 15, 18, 10].

3. DEMOGRAPHICS OF QATAR

Geographically, Qatar is located in the center of the Middle Eastern Countries. Although it is a small country, Qatar is one of the wealthiest countries in terms of per-capita income in the world due to oil and gas exploitation. Qatar is currently in a job and construction boom, partly related to the 2022 FIFA World Cup, which is attracting a lot of foreign workers. As a result, the country has experienced a significant shift in its population composition, doubling in size to about 2 million in 7 years. Figure 1 shows the population as of 2012, according to the Qatar Statistics Authority and Qatar’s Permanent Population Committee [11]. We can see that the demographic composition is very diverse with only 15% of the total population being Qataris, while the rest is mixed, and includes large fractions of Indians, Nepalese, and Filipinos. These statistics show the “offline” demographics of Qatar. Conversely, in our study, we analyze “online” nationalities as found on Twitter.

4. DATASET CONSTRUCTION

To study the online nationality distribution of Qatar, we chose Twitter as a platform due to its wide popularity and relative ease of data access through public APIs. To find a significant number of Twitter users based in Qatar, we made two constraints when querying the APIs: Users should either (i) explicitly state in their profile that they are located in Qatar (in the free text “location” field) or (ii) have at least one geo-tagged tweet originating from Qatar.

A total number of 51,449 candidate Twitter user profiles were collected between April 2013 and June 2013. For these users, we queried the API to collect all their publicly available tweets (up to 3,200). In total, 54,075, 860 tweets were collected. In addition to the tweets’ content, we obtained additional metadata information such as the country tags, location. We run the code on all collected tweets. We calculated tweet language distributions for each Twitter user. In the crowdsourcing jobs, we show the top 3 languages used in tweets for every user.

- We use language detection tools by Shuyo, Nakatani [19] and run the code on all collected tweets. We calculated tweet language distributions for each Twitter user. In the crowdsourcing jobs, we show the top 3 languages used in tweets for every user.

- We use an R library to convert latitude/longitude pairs from geo-tagged tweets to countries. In the crowdsourcing jobs, we show the top 3 locations along with a small sample of geo-tagged tweets from each location for every user.

Since most people state their location in natural language, e.g. “NYC,” “New York City,” and so on, we employed a state-of-the-art geo-coder to map these freetexts to explicit countries. We used the Yahoo! Placemaker API for this extraction. We show the top 3 self-stated countries of a user’s followers and friends in the crowdsourcing jobs.

3We are using Twitter’s terminology where a “friend” denotes a “followee,” the compliment of a follower.

4http://developer.yahoo.com/yql/console/

Figure 1: A Glimpse into Qatar’s “offline” Demographic.
We also show other information about the Twitter users available from the Twitter API, such as their name, screen name, profile picture, biography, a link to a homepage, location, time zone, and interface language.

4.2 Data Labeling

We use the Crowdflower platform\(^5\) for crowdsourcing. To create as-easy-as-possible micro-tasks for the contributors, we recreate Twitter profile pages and provide the preprocessed information about each user in the labeling interface. Figure 2 shows an example of our crowdsourcing job.

![Example of a crowdsourcing task](http://crowdflower.com/)

To ensure that the contributors correctly understood instructions, and to remove bots and spammers, we created 100 “gold” samples for the so-called “quiz mode.” Potential contributors will first only see “gold” samples and must correctly tag at least 8 out of 10 gold samples before they can access non-gold units. For each contributor, we also require at least 3 truthful labels with higher than 66.6% agreement, which means at least two people agreed with each other. If such an agreement is not attained on the first 3 tags, we will require more people to tag the data until the agreement reaches 66.6%.

With respect to nationality group, we divided the Twitter users into 6 groups: Qatari (QA), non-Qatari Arab (ARA), Westerner (WES), Indian Subcontinent (IN), Southeast Asia (SA) and others (OTH). These simplified groups are based on the Social and Economic Survey Research Institute (SESRI) of Qatar University and (OTH). These simplified groups are based on the Social and Economic Survey Research Institute (SESRI) of Qatar University and in other statistics concerning the country’s population. We assigned a group Unclear (UN) for Twitter profiles where Crowdflower contributors think there is not enough information to classify them into any of the 6 defined groups.

The tagging process lasted about a week. Figure 3 provides the results of the crowdsourcing tagging: The online demographics of Twitter users in Qatar is very different from the statistics in Figure 1. We attribute this difference to the fact that many expatriates from the Indian Subcontinent or from Southeast Asia come to Qatar to perform manual labor. Their salaries are often insufficient to permit purchases like smartphones or computers, and access to such devices is difficult. In addition, there is a possibility that unawareness and/or disinterest in Twitter leads to lack of tweet activity, and to remove bots and spammers, we created “hidden gold” data randomly.

4.3 Data Validation

To validate the correctness of the crowdsourced data, we introduced “hidden gold” data to our crowdsourcing job. Among 35,780 users, 1,210 stated their own nationalities in the profiles. We randomly selected and hand-checked a subset of these users and got 467 profiles which we were 100% sure about their nationalities. For these users, we replaced the nationality words that appeared in their profiles with “XXX” and shuffled them into our crowdsourcing job. Contributors on Crowdflower.com will be assigned to tag these “hidden gold” data randomly.

After the tagging process was complete, we evaluated the correctness of these “hidden gold” data, which shows that 91.86% of the validation set were correctly tagged.

5. CLASSIFICATION MODEL

In this section, we describe how to build classifiers using the ground truth labels from the crowdsourced data. First, we introduce the features we use for training the classifiers.

5.1 Features

Twitter offers a range of user information with different characteristics, so we chose to first include as much information as possible when generating the features.

Location-related features. There are 4 types of location-related features: 1) Followers’ Locations, 2) Friends’ Locations, 3) Self-stated Location and 4) geo-tagged tweets’ locations. Each feature vector is a 196 dimensional vector with each dimension representing a country;

Time zone. A vector of 24 dimensions with each dimension representing a possible time zone;

Language-related features. Due to the limitation of language detection tools, there are 20 possible dimensions for the language-related vector covering the most prominently spoken languages, but it is missing many less-common languages;

HashTags. We collected 7,057 different HashTags which appear more than 5 times in our dataset;

Profile picture features. We use Faceplusplus\(^6\)\(^,\)\(^6\) facial recognition API, to estimate basic information from profile pictures. The feature is a three dimensional vector with one dimension representing gender, one representing race, and one for age;

Name ethnicity. We use a name ethnicity detection toolkit by [20] and get a 10 dimensional vector with each dimension representing an ethnicity;

\(^5\) http://crowdflower.com/

\(^6\) http://www.faceplusplus.com/
UTF-8 charset type. We check the charset used by each user in their profile and their tweets. A vector with 209 dimension was formed, with each dimension representing the percentage of a type of charset (in UTF-8).

Tweet source. We collected 571 different utilities that Twitterers used to post tweets.

Mentioned users. Twitter users mentioned in Tweets (people you actually interact with)—this feature includes 3 sub features: 1) self-located mention of mentioned user, 2) interface language of mentioned user, 3) time zone of mentioned user.

Table 1 shows a complete list of features.

Table 1: Feature descriptions and examples.

5.2 Gradient Boosted Tree

Gradient Boosted Tree [5] is an effective off-the-shelf procedure for classification. The main advantages of Gradient Boosted Tree are that 1) it can handle data of mixed-type features, and 2) it is very robust regarding outliers in input space. In our case, we may have a lot of noisy data in our constructed feature space, and Gradient Boosted Tree can perform robustly in both training and prediction. Because the dataset is highly unbalanced, we performed our experiments and evaluations using stratified 5-fold cross-validation.

6. RESULT

The best performance of Gradient Boosted Tree was achieved with a number of trees= 300 and the best overall accuracy is 83.8%. In Table 2, we show the overall classification performance (Precision, Recall, F1 score) for each nationality group. Due to the unbalanced data distribution, we can see that the performance for less populated groups is not very high.

Table 2: The average Precision, Recall and F1 scores for each nationality group.

For a detailed analysis of the trained model, we show the confusion matrix of the classification results in Table 3. Combining the results from Table 2 we can see that the low performance of classifying non-Qatari Arabs is due to the confusion with the group of Qatari citizens. In the following section, we explain why such classification is so challenging.

Table 3: The confusion matrix of the trained classifier.

6.1 Feature Analysis

All features do not contribute equally to the classification model. In many cases, the majority of the features contribute little to the classifier and only a small set of discriminative features end up being used. Here, we discuss the importance of different features.

The relative depth of a feature used as a decision node in a tree can be used to assess the importance of the feature. Here, we use the expected fraction of samples each feature contributes to as an estimate of the importance of the feature. By averaging all expected fraction rates over all trees in our trained model, we could estimate the importance for each feature. It is important to note that feature spaces among our selected features are very diverse. The impact of the individual features from a small feature space might not beat the impact of all the aggregate features from a large feature space. So apart from simply summing up all feature spaces within a feature space (i.e. sum of all 7,057 importance scores in hashtag feature), which is referred to as un-normalized in Figure 4, we also plot the normalized relative importance of each features, where each feature’s importance score is normalized by the size of the feature space.
It is not surprising that the hashtag feature—with the largest ranked feature space—ranked first in the un-normalized plot. However, in the normalized plot, this feature ranked very low. One explanation is that some hashtags will be quite useful in classification, but the majority of hashtags are often not useful. To justify the explanation, we study the most popular hashtags when training the classifiers in Section 6.1.1. Location-related features like followers/following location and mentioned users’ locations are often dominant features. These features, however, are only ranked second (for follower locations) and fifth (for following locations) in un-normalized ranking, being ranked fifth/sixth in normalized ranking. The relative ranking of these can potentially be explained by the fact that many people follow pop stars or international news agencies, but non-famous users are unlikely to be followed by many international users. Tweet language is the third most important feature in both evaluation metrics. We presume that most expatriates from non-Arab countries are unlikely to understand or tweet in Arabic. In the following sections, we take a look at the details of some of the most influential features.

By comparing these two figures, we provide a guideline of feature selection: if you want one feature (or small feature group) to work (to some degree) all the time, then name ethnicity and race are the best choice. But if you want more accurate result, features with a large feature space size (i.e. hashtags) will help improve classification models.

6.1.1 Hashtags

In Figure 5, we plot the most influential hashtags for training the classifier using the Gephi toolkit. Red dots represent Twitter users labeled as Qatari, green dots represent non-Qatari Arabs, blue dots represent Westerners, cyan dots represent Southeast Asians, and yellow dots represent people from the Indian subcontinent. The large gray nodes (mostly overlapped by the textual hashtags) represent each hashtag. This color scheme will be used throughout the remainder of the paper.

From this figure, we observe that certain hashtags are only used by certain groups. For example #IPL (Indian Premier League) only appears among people from the Indian subcontinent, #ihatequotes is used among people from Southeast Asia, and #No thirsty in Qatar are most used among Qataris. Such hashtags serve as “sufficient” features, which means that if such hashtags appear in one’s tweets, the classifier will have sufficient information to infer your nationality group. However such “sufficient” features...
8. REFERENCES