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

MOTIVATION

- It is often useful to have detailed social media user attributes such as gender, age or nationality.
- In most countries, the population is dominated by the “native” nationality.
- However, in Qatar, the *majority* of the population is foreigners, exceeding 85%. (other Gulf counties ,Singapore, Switzerland as well)
- We are interested in potential correlations between national identity and social capital.

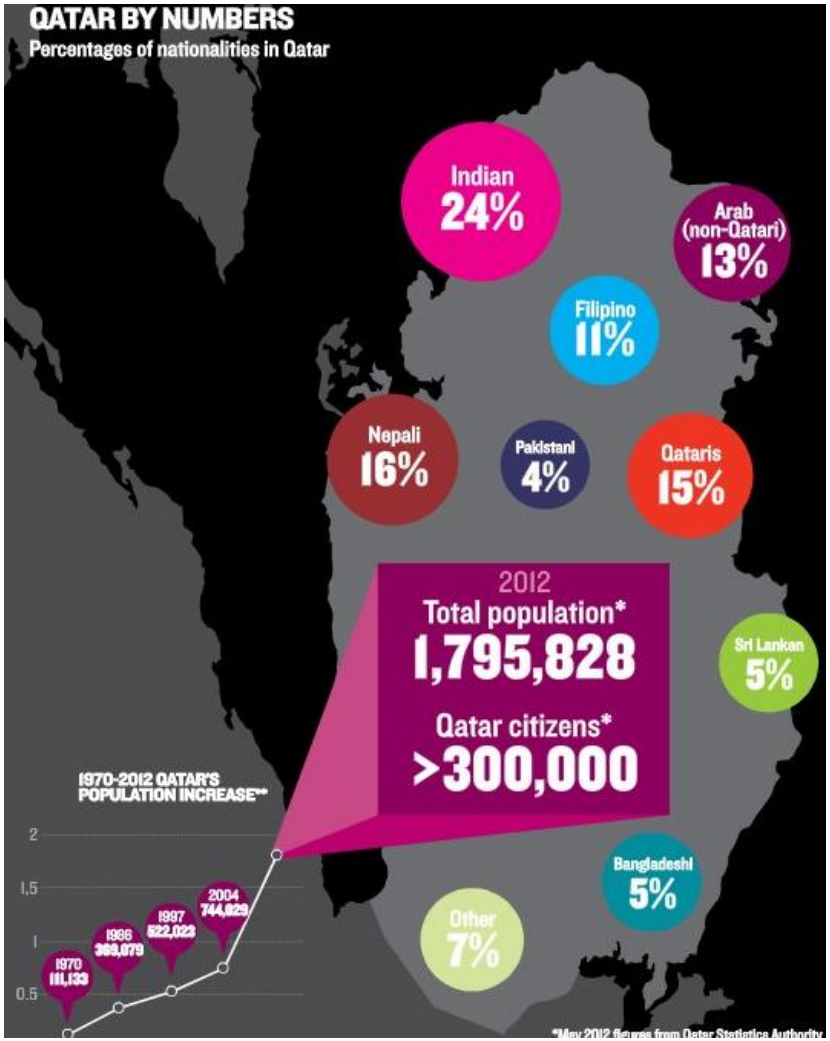


Figure 1: A Glimpse into Qatar’s “offline” Demographic.

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.

- We made two constraints: users should
  - explicitly state in their profile that they are located in Qatar (in the free text “location” field);

  OR
  - have at least one geo-tagged tweet originating from Qatar.

- 51,449 candidate Twitter user profiles were collected between April 2013 and June 2013.

DATASET CONSTRUCTION

- For each Twitter users, we collected:
  - their profiles information;
  - their publicly available tweets (up to 3200 each user);
    - In total, 54,075,860 tweets were collected.
  - latitude and longitude for geo-tagged tweets;
  - Device used to post tweets;
  - Followers' and friends' Twitter profiles;
    - 5,572,765 profiles including their self-declared location.
  - Profile pictures.

- We restricted our study to 35,780 users who had at least 10 tweets and at least 5 followers and 5 friends

Preprocessing

- We use language detection tools by Shuyo, Nakatani. We calculated tweet language distributions for each Twitter user.
  - SEVİYORUM BU ŞEHRİ -> Turkish
  - دوحة -> Arabic

- We use an R library to convert latitude/longitude pairs from geotagged tweets to countries.
  - (25.2867, 51.5333) -> Qatar

- Since most people state their location in natural language, e.g. “NYC,” “New York City,” and so on. We used the Yahoo! Placemaker API to extract their locations.
  - UAE, Emirates, دولة الإمارات العربية المتحدة -> United Arab Emirates

- 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.

Data Labeling

- We use the Crowdflower platform for crowdsourcing.
- We divided the Twitter users into 6 groups: Qatari (QA), non-Qatari Arab (ARA), Westerner (WES), Indian Subcontinent (IN), Southeast Asia (SA), others (OTH) and unclear (UN).
- We created 100 “gold” samples for the so-called “quiz mode.“
- For each job, we also require at least 3 trustful labels with higher than 66.6% agreement

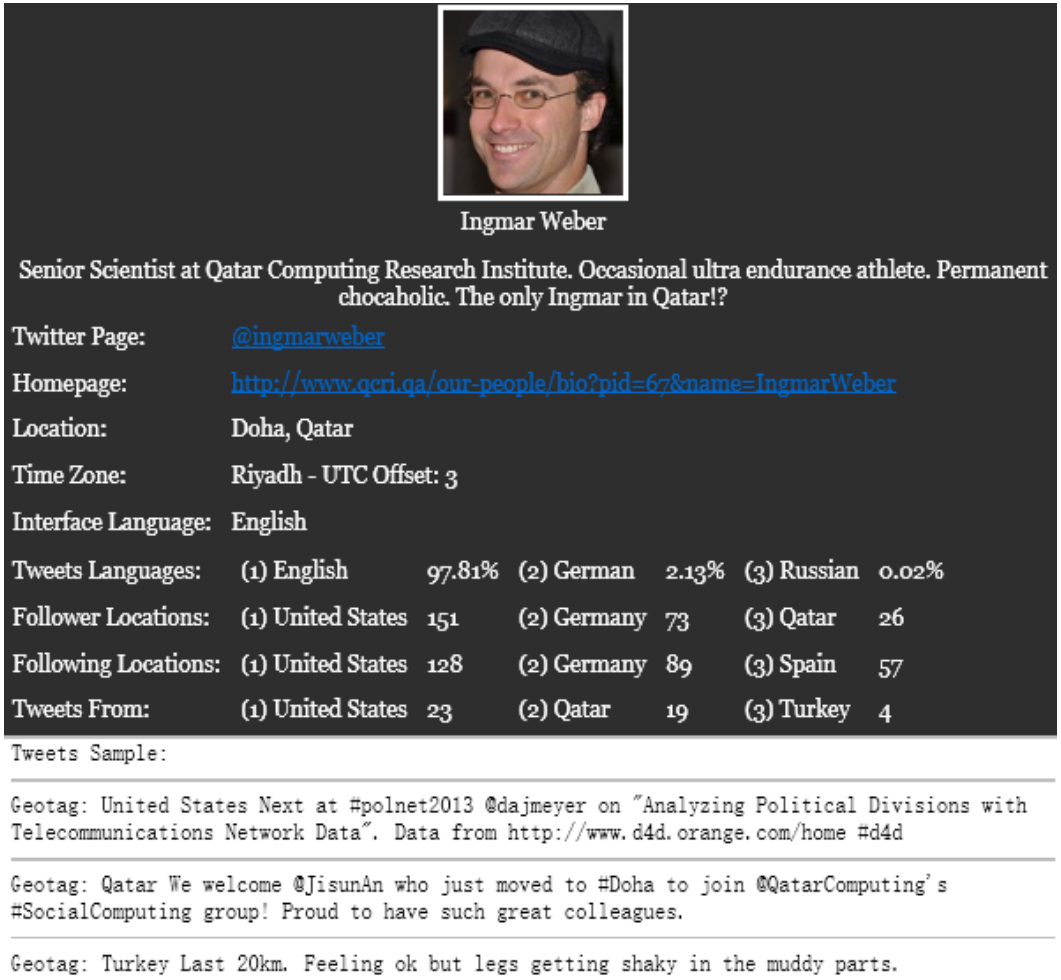


Figure 2: An example of a crowdsourcing task.

Data Validation

- We introduced “hidden gold” data to our crowdsource job.
- 1,210 users (among 35,780) stated their nationalities in the profiles.

- We randomly select 467 profiles.
- Nationality words that appeared in the profiles are replaced with “XXX”.
- 92% of these “hidden gold” data were correctly labeled.

- The online demographics of Twitter users in Qatar is very different from the statistics in Figure 1

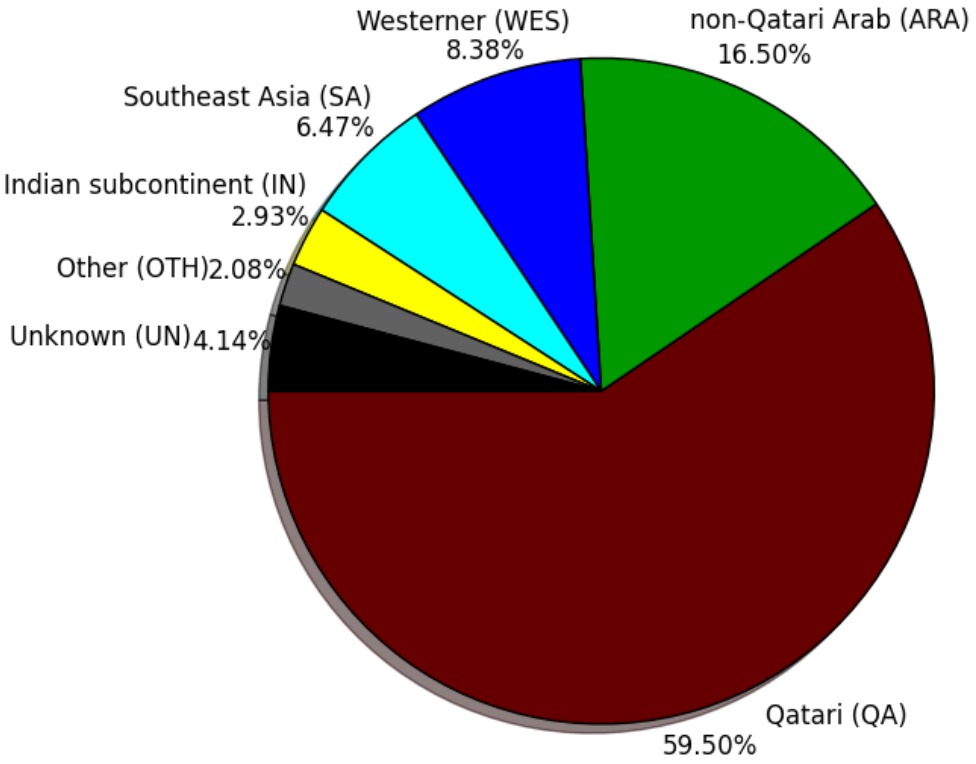


Figure 3: “Online” demographics of Twitter users in Qatar.

CLASSIFICATION MODEL

- Features:
  - Location-related features.
    - 196 dimensional vector with each dimension representing a country.
  - Time zone.
    - 40 UTC offset and time zone names.
  - Language related features.
    - 20 possible dimensions for most prominently spoken languages.
  - HashTags.
    - 7, 057 different HashTages which appear more than 5 times;
  - Profile picture features.
    - We use Faceplusplus to get gender, race and age information from the profile picture.
- Name ethnicity.
  - We use a name ethnicity detection toolkit and get a 10 dimensional vector with each dimension representing an ethnicity;
- UTF-8 charset type.
  - A vector with 209 dimension 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.
  - users mentioned in Tweets (people you actually interact with)
  - Three sub features: 1) self-stated location of mentioned user, 2) interface language of mentioned user, 3) time zone of mentioned user.

CLASSIFICATION MODEL

- Feature examples:

- Gradient Boosted Tree:
  - it can handle data of mixed-type features;
  - it is very robust regarding outliers in input space.

Feature	Description & Example
follower loc	[QA: 20, US: 1 , ...]
following loc	similar to follower loc
self loc	[UAE, UK, ...]
geo tag loc	similar to follower loc
time zone	[Abu Dhabi]
tweets lang	[EN: 70.7%, ES: 20.5%, UN: 8.8%]
interface lang	[EN: 1]
hashtag	[#love: 1 , : 1, #Mubarak: 1, ...]
race	[White: 91%, Yellow: 7%, Black: 2%]
age	[Age: 31, age_confidence: 83%]
gender	[Male: 1, gender_confidence : 98%]
name eth	[English: 89%, German: 9%, French: 2%]
charset	[Arabic: 25.1%, Basic Latin: 45.8% , ...]
source	[Twitter for iPhone: 312, Mobile Web: 3 ...]
mention loc	similar to follower loc
mention time zone	similar to time zone
mention lang	similar to interface lang

Table 1: Feature descriptions and examples.

RESULT

- Our experiments and evaluations are performed using stratified 5-fold cross-validation.
  - The overall accuracy is 83.8%.

	Precision	Recall	F-measure
QA	86.67%	95.37%	90.81%
ARA	82.96%	71.16%	76.56%
WES	70.86%	70.62%	70.64%
SA	93.35%	90.48%	91.89%
IN	82.19%	71.13%	76.00%
OTH	78.67%	40.72%	53.54%
UN	30.78%	15.13%	20.16%

Table 2: The average Precision, Recall and F-measure scores for each nationality group.

Confusion Matrix

Table 3: The confusion matrix of the trained classifier.

		Predicted Label						
		QA	ARA	WES	SA	IN	OTH	UN
True Label	QA	**5439**	143	63	10	9	2	37
	ARA	404	**1125**	25	1	5	4	17
	WES	125	28	**567**	12	12	14	45
	SA	24	5	16	**561**	2	0	12
	IN	41	2	16	2	**200**	1	19
	OTH	27	10	66	4	0	**81**	11
	UN	216	45	48	11	16	1	**60**

- the low performance of classifying non-Qatari Arabs is due to the confusion with the group of Qatari citizens.

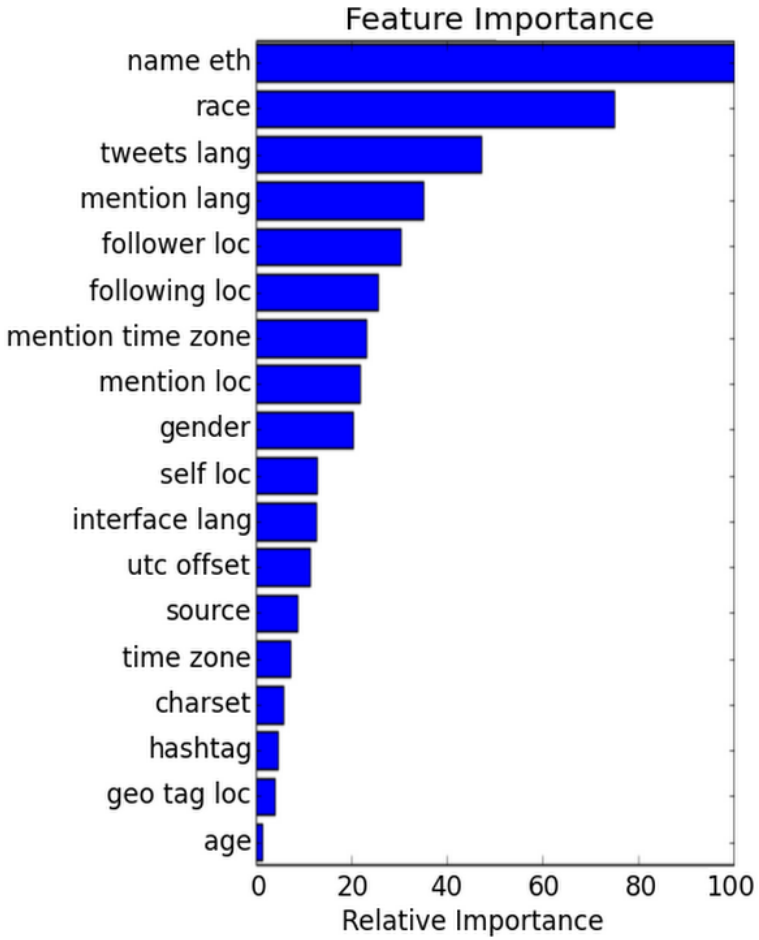
Table 4: The normalized confusion matrix of the human labelling.

		Predicted Label						
		QA	ARA	WES	SA	IN	OTH	UN
True Label	QA	**5158**	259	92	8	7	7	169
	ARA	86	**1418**	16	1	1	2	53
	WES	27	8	**721**	3	1	7	32
	SA	13	1	10	**578**	0	2	12
	IN	8	2	11	2	**239**	2	12
	OTH	5	3	8	0	0	**172**	8
	UN	76	30	40	11	7	5	**224**

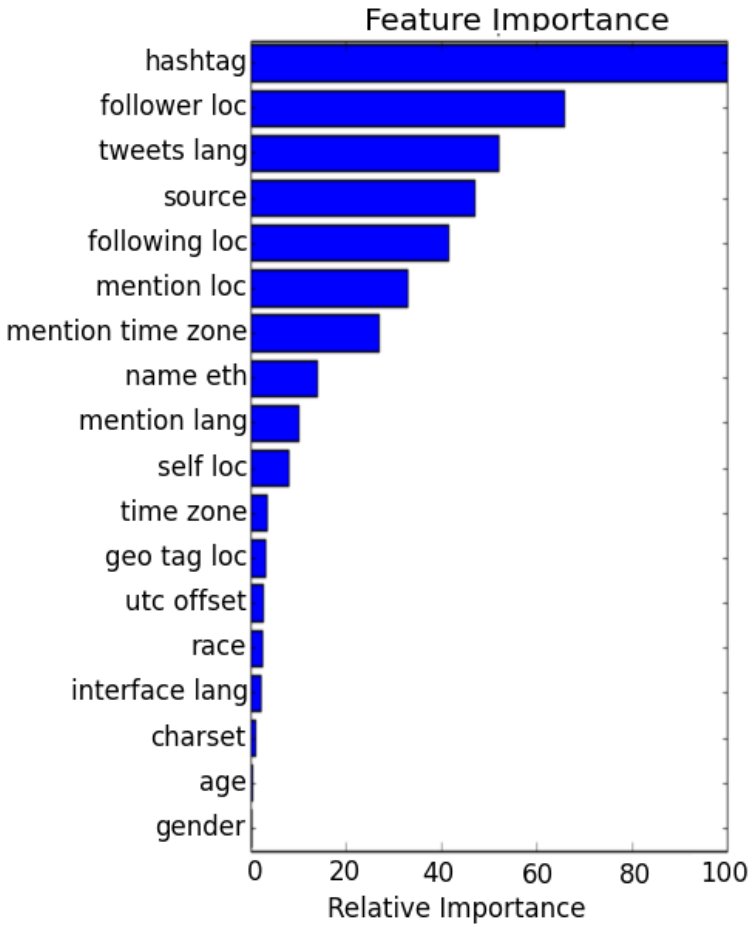
- human labeling is better than our classification model.
- it is difficult for humans to distinguish between the Qatari group (QA) and non-Qatari Arabs (ARA).

Feature Analysis

- The relative importance of different features:



(a) Normalized

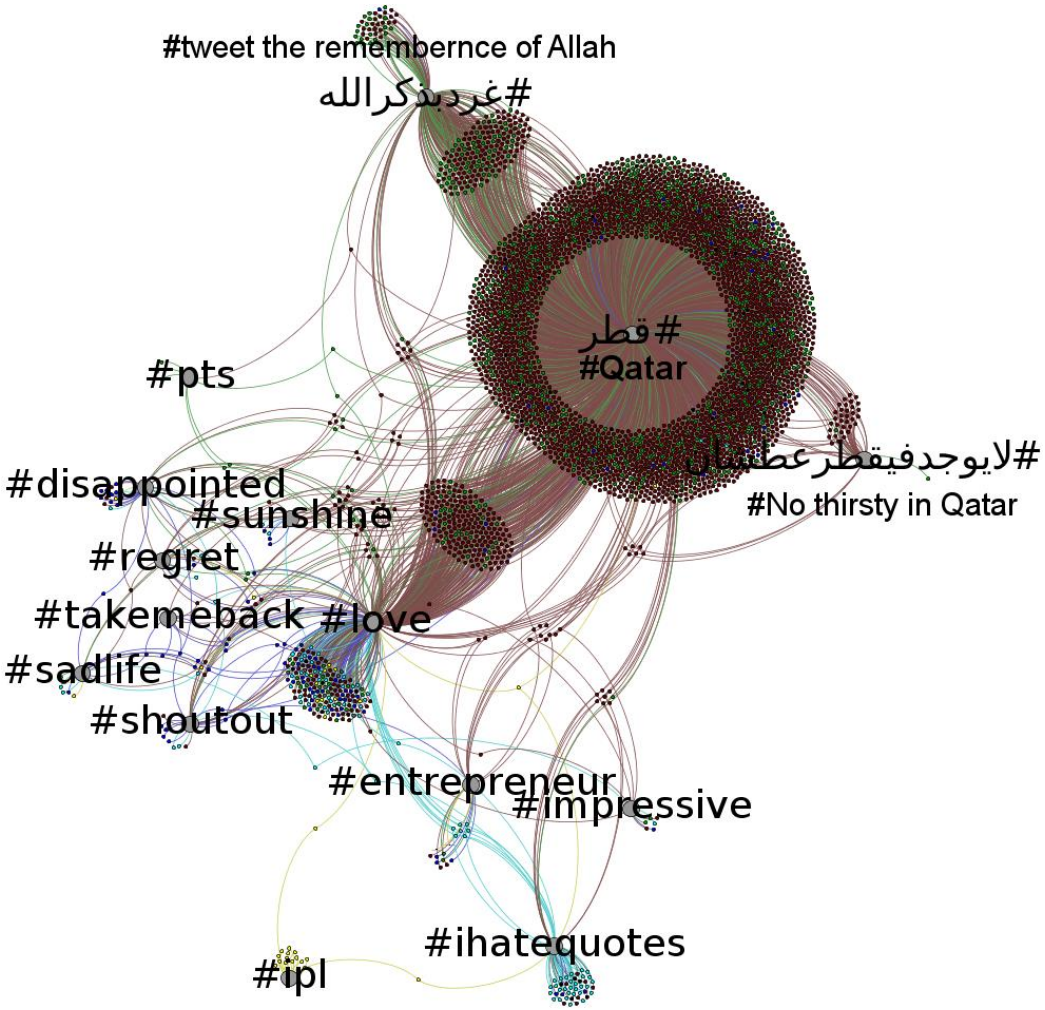


(b) Un-normalized

Figure 4: Relative Feature Importance when training the models.

Hashtags

- Certain hashtags are only used by certain groups.
  - #IPL (Indian Premier League)
  - #Ihatequotes
  - #No_thirsty_in_Qatar
- Non-Arab people rarely use Arabic hashtags, and they rarely retweet Arabic hashtags.
- Compared to Arabs, others are more willing to express personal feelings - at least in English
  - #love
- Foreign expatriates have a much higher probability of expressing negative emotions in tweets.
  - #sadlife
  - #disappointed
  - #takemeback



CONCLUSIONS

- We built a classification model to address the question of how to identify nationalities of Twitter users.
- A feature analysis study was performed, and we discovered some interesting patterns of user features.
- Our methodology serves as a foundation for future work.
- Exploring the link between social capital, cultural capital, and Twitter use and relationships in Qatar are all rich areas of study.