Dietary Habits of an Expat Nation: Case of Qatar

Yelena Mejova,[1] Hamed Haddadi,[1] Sofiane Abbar,[1] Azadeh Ghahghaei,[2] Ingmar Weber[1]

[1]Qatar Computing Research Institute, [2]Freie Universität Berlin

***Abstract*—**

We introduce an exhaustive collection of Instagram posts tied to locations in the Gulf nation of Qatar, and explore its potential as a source for the study of the native and expat populations, in particular their dietary and health activity. Obesity has taken an epidemic proportion in such fast-developing countries, and we show that, for example, for Arabic-speakers posting from restaurants strongly correlates with posting from sweets shops. Furthermore, posts in Arabic attract almost three times as much liking and commenting as posts in English. However, we also show that social media has substantial limitations in accurately reflecting the expat population, with languages of India and Nepal – whose expat populations outnumber the locals – being drastically under-represented.

I. Introduction

Rapidly developing countries which have high technological penetration rates present a unique opportunity to study changing cultural norms, and integration of new expatriates. In the past 30 years the population of Qatar grew 13-fold [1], and today it is a prominent player both in political and cultural international spheres. The combination of the highest internet penetration rate in the region at 86% in 2011 [2] and a largely expatriate population create a unique opportunity to study a diverse society using online behavioral data. Unfortunately, rapid development comes at a price, with a recent survey putting Qatar as fifth in adult obesity worldwide [3]. Culture, as it develops in this dynamic environment, is a central guide to creating effective local intervention campaigns.

In this paper we explore the culture of food and activity of the multinational population of Qatar, focusing on Arabic versus English speaking users of Instagram. Exhaustively covering Instagram locations in Qatar, we collect all checkins made in Qatar since 2010. Each checkin consists of an image, description text, hashtags, as well as other users' comments associated with a visit to a particular location. We then manually classify locations into different activity categories such as Food, Outdoor & Recreation, and Nightlife spots, and look for correlations between different types of activities. Compared to English-speaking population (though some of which may be Qatari), we find differences in the behavior of these users, with fewer posts from locations associated with high physical activity, as well as a significant difference in the other users' interaction with the content (with posts in Arabic receiving three times the number of likes of English posts). We hope that this preliminary study encourages the use of social media for culturally-relevant health research, especially in rapidly developing nations.

II. Related Works

Physical activity and dietary intake have a proven relationship with the alarming global rise of diabetes, obesity, and many cardiovascular diseases [4], [5], [6]. Social networks have been a useful testbed for understanding health issues such as obesity [7]. Recently, there has been a growing interest in the use of online social media for health studies. Particularly, analysis of Twitter and Instagram has proven insightful for understanding health characteristics of the United States [8], [9], [10]. In the context of obesity and physical activity, lack of motivation for healthy eating and physical exercise have been highlighted as being the biggest barriers to exercise in Arab countries [11]. While traditionally weight-gain has not been a major concern for local Qataris [12], this has changed due to sudden growth in obesity rates with Qatar and other Arab countries taking top positions in the world obesity rankings [13]. Correspondingly, this has become an alarming concern for the governments and health professionals in the region. The increase in obesity has in turn lead to high rates of diabetes in Qataris, now over 20%.[1]

Globalization of food supply and changes in nutrition patterns have been cited to be the highest contributor to obesity in the Middle East [14]. Although there have been survey-based studies on obesity problems in Qatar [15], use of social media gives us a larger sample size in addition to enabling us to understand the social approvals of different networks. Hence in this paper we pay specific attention to the use of Instagram as one of the most popular social network in this region to understand these trends from a social perspective. We pay particular attention to the popularity of different types of food and activities amongst Arab and English speakers in Qatar.

III. Data

Here, we describe the collection of our dataset, which combines all Foursquare locations in Qatar with Instagram posts from these locations.

We begin by querying Foursquare `venues/search` API[2] with a bounding box around the whole of Qatar. We then check whether the number of responses hits the limit of 50, and if so, subdivide the area into four boxes and recursively proceed to collect their locations. This approach produced $21,376$ locations, each with the corresponding name, coordinates, country, and category. The categories, which we use extensively in our analysis, are uniquely identified in a hierarchy.[3] We also used

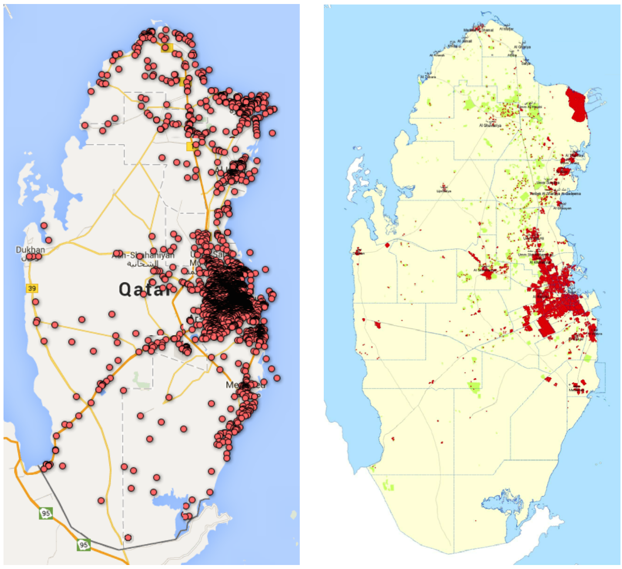
[1] http://blogs.nature.com/houseofwisdom/2013/03/qatar-surpasses-us-in-obesity.html

[2] https://developer.foursquare.com/docs/venues/search

[3] https://developer.foursquare.com/categorytree

the country field to make sure the data are from Qatar, and not the neighboring Bahrain or Saudi Arabia, which overlap with the initial box.

We then mapped these Foursquare location IDs to Instagram ones using the Instagram `location/endpoints` API[4], resulting in $16,280$ locations, with 76.2% finding a match. At the time of the collection, Foursquare was the location provider for Instagram queries, hence this conversion is highly accurate. Figure 1 shows the locations we have gathered, and a population census from 2010 as a comparison. The most populated area is around the capital city of Doha, with beaches and main roads being other popular posting sites.

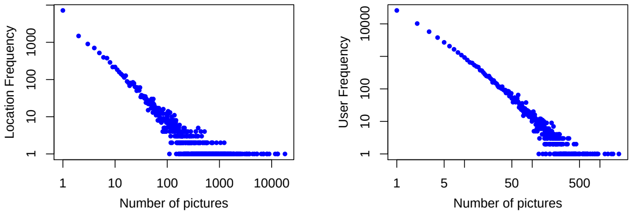


(a) Instagram (b) Census 2010

Fig. 1: Foursquare locations compared to population.

On March 4, 2015 we collected all of the Instagram posts for each location, exhaustively collecting the historical posts, with some dating back to October 10, 2010. The distribution of this data is shown in Figure 2, for locations (a) and users (b). The most popular location is *Souq Waqif* – a popular market with restaurants and shops – and in that it can be considered an aggregation of a whole area of potential locations. As far as it is evident to the authors, complete histories were retrieved for all locations, except *Souq Waqif*, which maxed out at 18,000, although still going back as far as 12/26/2010.

The collection encompasses 65,673 unique users, with 40,060 users posting more than once, and 11,684 posting 10 times or more. The posting behavior shows a familiar power law distribution, as seen in Figure 2b. Upon manual checks, only few instances of automated posts from bots were found, which leads us to believe the data to be of high quality in this respect. However note that since association with known location was used for querying this data, those checkins which may have GPS coordinates in Qatar but which are *not* associated with a location would not be present in this collection, and we leave the collection of such data for future research.



(a) Locations (b) Users

Fig. 2: Frequencies of locations and users having some number of Instagram posts.

IV. Results

A. *Location categories*

What kinds of places do Instagram users in Qatar visit? To address this question, we take advantage of the Foursquare location categorization hierarchy. The latest version of this hierarchy, consisting of 713 distinct categories and 3 levels, was manually converted to a computer-readable form, and is available online[5]. Table I shows the top-level categories with the accompanying number of Instagram posts in our dataset. The most popular *Food* category encompasses 228 categories ranging from coffee shops to 12 kinds of Turkish restaurants, illustrating the richness of Instagram as a source for diet-related research.

TABLE I: Number of Instagram pictures per top-level location category.

Category	N_{pic}
Food	$124,256$
Outdoors & Recreation	$111,035$
Travel & Transport	$100,441$
Shop & Service	$70,590$
Professional & Other Places	$69,069$
Arts & Entertainment	$53,824$
College & University	$26,768$
Residence	$26,764$
Nightlife Spot	$8,391$
Event	$1,248$

B. *Language detection*

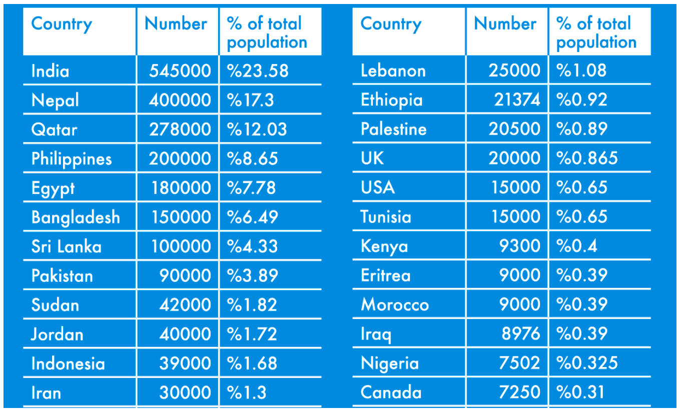
Another dimension we explore in this study is the multinational composition of Qatar, with native Qataris constituting a mere 12% of the total population (as of late 2013)[6] (see Figure 3). Expats from the nations of India, Nepal, and Philippines are the most common, as well as those from neighboring countries, and yet others from America and Europe. Since Instagram users do not provide their nationality, or often even withhold their real names, we use language as a proxy of national or regional belonging.

Three sources of text accompany Instagram posts: tags and comments provided by the posting user, and the comments

[4] https://instagram.com/developer/endpoints/

[5] https://sites.google.com/site/yelenamejova/resources

[6] http://www.bqdoha.com/2013/12/population-qatar



Country	Number	% of total population	Country	Number	% of total population
India	545000	%23.58	Lebanon	25000	%1.08
Nepal	400000	%17.3	Ethiopia	21374	%0.92
Qatar	278000	%12.03	Palestine	20500	%0.89
Philippines	200000	%8.65	UK	20000	%0.865
Egypt	180000	%7.78	USA	15000	%0.65
Bangladesh	150000	%6.49	Tunisia	15000	%0.65
Sri Lanka	100000	%4.33	Kenya	9300	%0.4
Pakistan	90000	%3.89	Eritrea	9000	%0.39
Sudan	42000	%1.82	Morocco	9000	%0.39
Jordan	40000	%1.72	Iraq	8976	%0.39
Indonesia	39000	%1.68	Nigeria	7502	%0.325
Iran	30000	%1.3	Canada	7250	%0.31

Fig. 3: Qatar population statistics as of October 2013.

posted by other users on the picture. To assist automated classification, we clean the text by removing punctuation, user names (preceded by an @), and hash symbols (#), but not hashtags themselves. Note that we were careful not to exclude too many special characters, as this would often remove non-Latin alphabets. We then applied a language identification tool for Python LangID[7], which identifies 97 languages, trained on data including Wikipedia, ClueWeb09, and Reuters RCV2 datasets. Language detection is known to be a challenging problem on microblogs, mainly due to the limited text length. However, Lui and Baldwin have shown that *langid* achieves an accuracy of 0.94, making it a sufficiently precise tool for the purpose of this paper [16]. In Table II we show the top languages identified for the tags and comments provided by the posting user, since these more likely represent the user's language than the comments posted by others. Column N_{pic} lists counts of individual posts, and N_{usr} counts the user's main language. English dominates the tags, followed by Arabic, and None, indicating no text was present. The comments were blank nearly 42% of the time, with Arabic being the most common language, followed by English.

TABLE II: Top 10 languages as identified for the posting user's tags or comments, including no text present (None), and average language used per individual users.

Tags			Comments		
Language	N_{pic}	N_{usr}	**Language**	N_{pic}	N_{usr}
English	169,992	21,558	None	249,024	30,550
Arabic	147,815	16,962	Arabic	185,412	21,401
None	128,193	16,634	English	56,525	60,02
Chinese	15,616	1,104	Swahili	13,602	789
Japanese	11,392	870	Latin	10,493	834
Urdu	9,043	544	Urdu	9,739	584
Farsi	8,641	523	Pashto	8,174	321
Maltese	6,635	380	Quechua	6,331	355
Quechua	6,190	182	Malay	5,820	541
Spanish	5,879	557	Farsi	4,672	276

Note that identifying the language of short text, especially which is not a sentence and lacks the stop words which identify a language, is difficult. Thus, we are more suspicious of the Chinese and Japanese labels for the tag classification, which may be provoked by a great variety of unicode characters in the data. However, the distribution of the languages for the comments is more like that in the latest population figures of Qatar. Still, we see a much greater penetration of the app in Arabic and English-speaking populations, indicating that, unsurprisingly, Indian and Nepalese migrants are not well represented. Thus, we restrict ourselves to populations of Arabic and English speakers. To be most selective, we examine the posts which contain both tags and comments in Arabic (87,849 posts, 16,618 unique users), and both in English (32,789 posts, 9,645 unique users).

C. Posting Behavior

We further delve into the data by extending the location category assignment to non-overlapping "super" categories, as listed below:

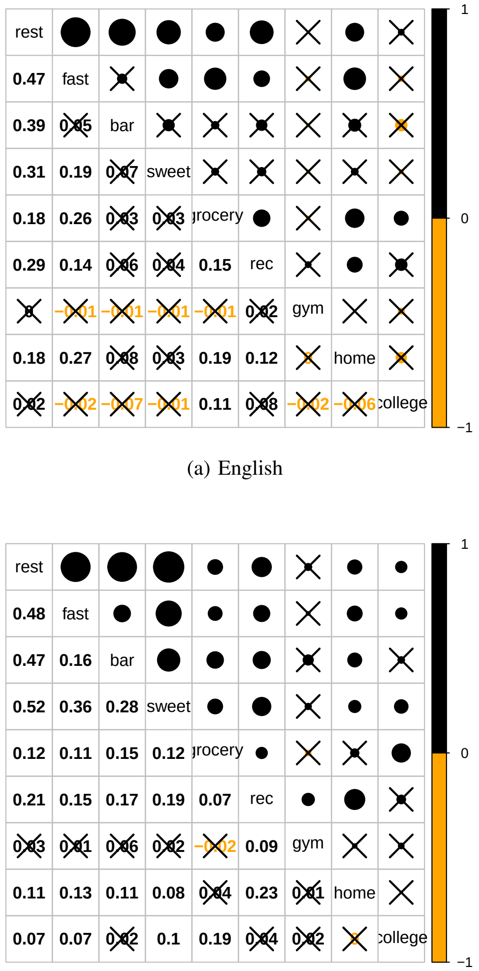
- *Restaurant* - most restaurants (under Food category), excluding those below
- *Fast food* - fast food places, burger and pizza places
- *Bar* - bars and nightlife spots
- *Sweet* - sweet shops, ice-cream parlors, donut shops
- *Grocery* - food grocery stores
- *Recreation* - parks, beaches, public recreational spaces
- *Gym* - fitness clubs, gyms, sports courts
- *Home* - residences, houses, apartment buildings
- *College* - university and college campuses

Note that these categories do not include all of the collected locations, for instance businesses not dealing with food, entertainment venues or transportation.

Now, we are interested in the interaction between these categories. For example, do users who post at restaurants also visit grocery stores (and cook at home)? Does the likelihood of one going to the gym decrease one's visitations to restaurants? Do college students cook more at home? To answer these questions, we compute correlations between each of these. To avoid biases and noise, we select users with $n \geq 10$ posts within a language (English or Arabic), then first aggregate the posts by user, and normalize by the total posts (all within a language). The correlation matrices for English and Arabic posts are shown in Figures 4(a,b), with correlations not meeting significance level of $p = 0.01$ crossed out. Note that the correlation value can be found in the lower triangle and visual representation in the upper triangle of the matrix.

The largest correlations the two groups have in common are the connection between Fast Food and other restaurants (0.473 for English and 0.478 for Arabic), indicating that for both cohorts fast food chains are still popular among users who also visit other restaurant-types. Among other establishments, Arabic speakers visiting restaurants tend to also visit sweet shops more at 0.522 (compared to 0.310 for English). Also Arabic speakers posting from college/university locations (potentially students or staff) are significantly likely to visit restaurants and fast food places, whereas English speakers do not show this relationship, although both tend to also post from grocery stores. There are few significant correlations with gym posts for either group – this is due to the low number of such posts, indicating uneven coverage of this data.

[7] https://github.com/saffsd/langid.py



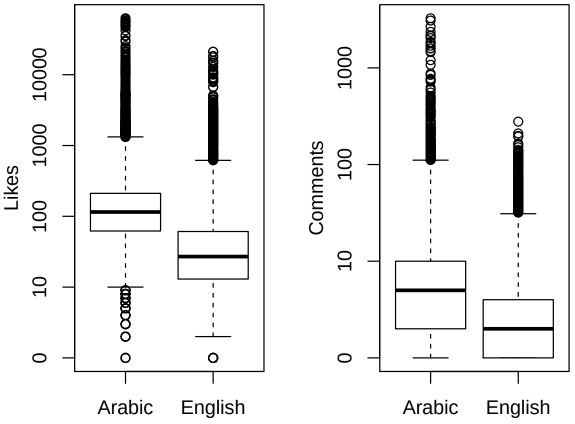
(a) English

(b) Arabic

Fig. 4: Correlation between location super-categories for English and Arabic language posts, aggregated and normalized per user, with correlations not meeting significance level of $p = 0.01$ crossed out.

We also see a marked difference between social behavior, as operationalized by *likes* and comments, shown in Figure 5. The difference is highly statistically significant, with mean comments at 9 for Arabic and 3.5 for English, and likes at 219 for Arabic versus 78 for English.

Looking beyond restaurant visitations, we proceeded to examine posts at locations associated with physical activity. Using the number of calories burnt during a physical activity[8] estimated by Harvard Heart Letter[9], we classified each location into slightly active, moderately active, and active. During aggregation, we then normalized posts for each user (preventing users who post more to dominate the statistics), and by overall number of users (resulting in a proportion from 0 to 1). Figure 6 shows the proportion of posts at the three classes of locations. We find that both English and Arabic speakers mostly favor locations associated with light activity (top locations dominated by the many beaches the country has to offer), but that English-speakers are more likely to also post from very active locations like sports clubs.



(a) Likes

(b) Comments

Fig. 5: Likes and comments of posts whose posters used Arabic vs. English

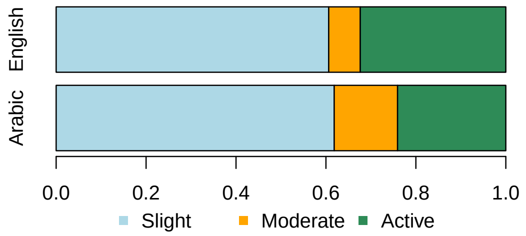
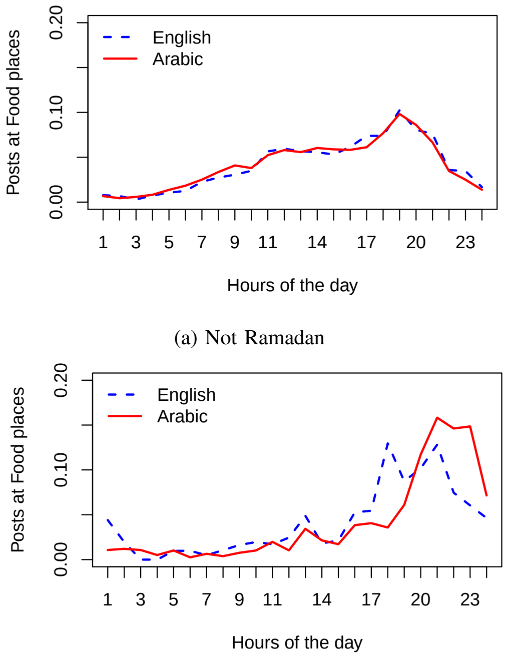


Fig. 6: Posts at locations with various activity levels

Finally, we can examine the behavior of the two cohorts during the month of Ramadan (in all 4 years our data spans), when fasting during the daylight hours is observed by practicing Muslims, but also at which days many restaurants are closed during the day. Figures 7(a,b) show the normalized hourly post volume throughout the day during the non-Ramadan days (a) and during Ramadan (b). Although the dining hours match very closely on most days, during Ramadan English-speakers begin posting earlier. However note that English-speakers also have very small number of posts during the day, suggesting that the holiday affects both cohorts. It is also possible that people avoid posting food-related content during the fasting hours out of respect for those in their networks who are fasting. As we see in Figures 8(a,b), English-speakers post in transport-related locations more during the lunch time in Ramadan, whereas Arabic-speakers more toward late evening (around 22), when evening meal *suhur* is shared with friends and family. Why users refrain from posting is an interesting challenge which will require inquiry outside Instagram to address.

[8] http://www.health.harvard.edu/diet-and-weight-loss/calories-burned-in-30-minutes-of-leisure-and-routine-activities

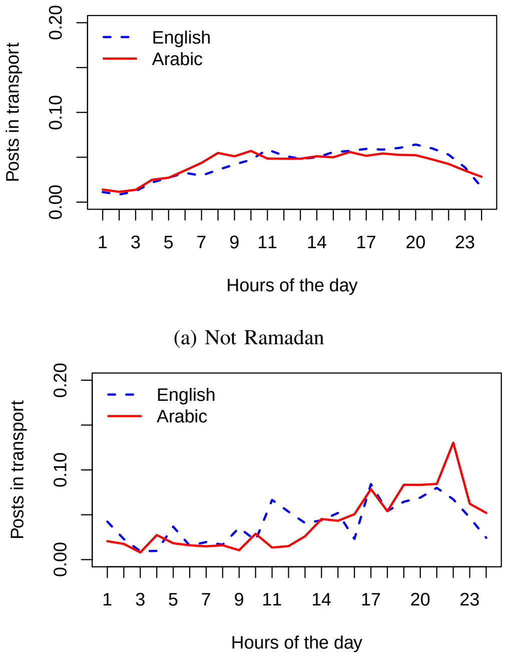
[9] http://www.health.harvard.edu/newsletters/Harvard_Heart_Letter



(a) Not Ramadan

(b) Ramadan

Fig. 7: Normalized distribution of posts at food locations, per hour of the day, during non-Ramadan days and during Ramadan.



(a) Not Ramadan

(b) Ramadan

Fig. 8: Normalized distribution of posts in transport, per hour of the day, during non-Ramadan days and during Ramadan.

V. Discussion

Much like previous studies on social media in Qatar, we show a vastly uneven penetration of the technology in the strata of society, with English and Arabic dominating the discourse, and Urdu, Pashto, and Malay at the tail of usage distribution. These results echo those on identifying nationality of Twitter users in Qatar by Huang *et al.* [17]. Because the use of English may be somewhat ubiquitous across many nationalities (a *lingua franca* of the internet), we will adopt more precise name and profile-based approaches to classifying Instagram users in this dataset.

In this preliminary data exploration, we performed only a rough analysis of the hashtags associated with posts. In particular, we found sweets to dominate the foods mentioned in both English and Arabic language posts. Looking at the top 200 most frequent tags, we labeled them for type of food, finding sweet-related tags (like "chocolate" or simply "sweet") to be quite popular, with 7 sweet out of 13 total food tags for Arabic and 7 out of 17 for English. However, only English-language posts had health-related tags (like "healthy" and "fresh"), with no such Arabic counterparts. Both mentioned social tags, including "friends" and "family" ("girl" appearing only in English ones), which could be used for discerning the nature and importance of social settings in food consumption or being physically active.

Next, our aim is to identify the actual foods which are mentioned in the tags and comments associated with this media, and produce a lexicon for automated dietary tracking annotated with additional information, including calorie content and other healthiness factors. In ongoing work, we are collecting and labeling food-related images, in order to enable the conversion of these images to their potential caloric value for automated intake analysis.

The reflection of a country's culture in social media presents a new alternative in the study of dynamic cultural changes which affect food consumption. As posited by Abdulrahman Musaiger some two decades ago [18], the rise of obesity in Arab countries can be attributed to many reasons: food preferences, religion, beliefs, education, gender roles, and change in women's employment. As we can see from the temporal analysis, religious practices alter behavior patterns, both having to do with food consumption and travel. Future study examining individual profiles, language use, and social media interaction may further show the development of the dietary behavior in demographic slices of population.

Ultimately, as the use of social media continues to proliferate in the Arab world, the temporal extension of the current dataset (which already spans four years) may capture the development of the country and track the change in dietary habits of its inhabitants. As Qatar grows and increasingly attracts expatriates from around the world, its dietary culture may drift away from its traditions. For example, how will the increase in the urbanized culture, public transport, and proliferation of health clubs affect the different local and expat communities? Social media, including Instagram, will be a valuable resource for answering these questions.

Finally, it is important to supplement social media datasets with other resources, including both linking to other social media and more standard approaches using surveys and questionnaires.

References

[1] The world bank. population data. [Online]. Available: http://data.worldbank.org/indicator/SP.POP.TOTL

[2] International telecommunication union. percentage of individuals using the internet 2000-2011. [Online]. Available: http://www.itu.int/ITU-D/ict/statistics/material/excel/2011/Internet_users_01-11.xls

[3] "Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the global burden of disease study 2013," *Institute for Health Metrics and Evaluation*, 2014.

[4] R. R. Pate, M. Pratt, S. N. Blair, W. L. Haskell, C. A. Macera, C. Bouchard, D. Buchner, W. Ettinger, G. W. Heath, A. C. King *et al.*, "Physical activity and public health: a recommendation from the centers for disease control and prevention and the american college of sports medicine," *Jama*, vol. 273, no. 5, pp. 402–407, 1995.

[5] L. Gillis, L. Kennedy, A. Gillis, and O. Bar-Or, "Relationship between juvenile obesity, dietary energy and fat intake and physical activity." *International journal of obesity and related metabolic disorders: journal of the International Association for the Study of Obesity*, vol. 26, no. 4, pp. 458–463, 2002.

[6] I. Janssen, P. T. Katzmarzyk, W. F. Boyce, C. Vereecken, C. Mulvihill, C. Roberts, C. Currie, and W. Pickett, "Comparison of overweight and obesity prevalence in school-aged youth from 34 countries and their relationships with physical activity and dietary patterns," *Obesity reviews*, vol. 6, no. 2, pp. 123–132, 2005.

[7] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England journal of medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[8] S. Abbar, Y. Mejova, and I. Weber, "You tweet what you eat: Studying food consumption through twitter," pp. 3197–3206, 2015. [Online]. Available: http://doi.acm.org/10.1145/2702123.2702153

[9] Y. Mejova, H. Haddadi, A. Noulas, and I. Weber, "# foodporn: Obesity patterns in culinary interactions," *ACM conference on Digital Health 2015*, 2015.

[10] M. J. Widener and W. Li, "Using geolocated twitter data to monitor the prevalence of healthy and unhealthy food references across the us," *Applied Geography*, vol. 54, pp. 189–197, 2014.

[11] A. O. Musaiger, M. Al-Mannai, R. Tayyem, O. Al-Lalla, E. Y. Ali, F. Kalam, M. M. Benhamed, S. Saghir, I. Halahleh, Z. Djoudi *et al.*, "Perceived barriers to healthy eating and physical activity among adolescents in seven arab countries: a cross-cultural study," *The Scientific World Journal*, vol. 2013, 2013.

[12] M. El-Islam, "Cultural aspects of morbid fears in qatari women," *Social Psychiatry and Psychiatric Epidemiology*, vol. 29, no. 3, pp. 137–140, 1994. [Online]. Available: http://dx.doi.org/10.1007/BF00796494

[13] M. Ng, T. Fleming, M. Robinson, B. Thomson, N. Graetz, C. Margono, E. C. Mullany, S. Biryukov, C. Abbafati, S. F. Abera *et al.*, "Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: a systematic analysis for the global burden of disease study 2013," *The Lancet*, vol. 384, no. 9945, pp. 766–781, 2014.

[14] O. Galal, "Nutrition-related health patterns in the middle east." *Asia Pacific journal of clinical nutrition*, vol. 12, no. 3, pp. 337–343, 2002.

[15] A. T. Soliman, "Obesity epidemic in qatar," *Hamad Medical Center, Qatar*, 2014. [Online]. Available: http://www.researchgate.net/profile/Ashraf_Soliman8/publication/261759173_Obesity_Epidemic_in_Qatar/links/00b7d5356b3df7a618000000.pdf

[16] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, 2012, pp. 25–30.

[17] W. Huang, I. Weber, and S. Vieweg, "Inferring nationalities of twitter users and studying inter-national linking," in *Proceedings of the 25th ACM conference on Hypertext and social media*. ACM, 2014, pp. 237–242.

[18] A. O. Musaiger, "Socio-cultural and economic factors affecting food consumption patterns in the arab countries," *The Journal of the Royal Society for the Promotion of Health*, vol. 113, no. 2, pp. 68–74, 1993.