Twitter Usage in the Developing World

Muhammad Raza Khan
University of California, Berkeley
Berkeley, CA, USA
mraza@berkeley.edu

Mahrukh Mehmood
University of Washington
Seattle, WA USA
mahrukh@uw.edu

Abstract
The penetration of mobile phones and social networks have made the forums like Twitter and Facebook an ideal way of collecting data about the populations at an unprecedented scale. The availability of this data provides an opportunity to take a deeper look into the patterns of interactions of people with technology. Furthermore, it gives the researcher an ability to see social phenomena at an unprecedented scale as well. In this work, we describe an exploratory study about the analysis of tweets from a developing country (Pakistan). Our study shows that contrary to expectations, not all the districts with higher human development index and literacy have the similar penetration of Twitter which highlights the presence of subtle patterns of technological preferences of the people in the developing countries. The data that we have collected gives us an ability to ask some interesting research questions about the preferences of the people of the developing countries when it comes to interacting with technology for exchange of information.

Author Keywords
Twitter, ICTD, HCI4D, Social Computing

Introduction
The popularity of social networks like Facebook and Twitter provide an unprecedented opportunity to analyze human activities at scale. The value of these mediums as a way

of data collection is much more advantageous in the developing world as extensive data collection surveys cannot be conducted in the developing countries. As a result, research using social networks data (Facebook data, Twitter data and Mobile Call Detail Records (CDR)) to analyze human behaviors is increasingly getting popular. Some of the interesting applications of using social networks data to analyze human behavior and population include Prediction of Poverty using CDR Data [1]; Improving Population mapping using CDR Data [2] & Twitter data [3]; and the use of CDR data to analyze gender disparities in communities [4]. In this work, we use Twitter data from 22 districts of Punjab province of Pakistan to analyze the patterns of microblogging in these districts.

Research Questions

The aim of our current research is to show that how the Twitter data can be used to analyze different communication preferences of the population of the developing countries. The research questions that we are interested in answering include

- Whether the people in the more developed districts of the developing countries have different tweeting patterns as compared to the people of lesser developed districts?
- How do the tweeting patterns across various districts vary when differentiated by language, gender, etc.
- What are the common topics of tweets across different districts?

The answer to these questions can help in a) better understanding of the social behavior in these districts; b) deeper insights into the interaction patterns of the users and Twitter and c) Improved user interfaces for the Twitter client through a better understanding of the user requirements and preferences.

Current Work

The techniques and methodologies that we are using in this project can be applied to the Twitter data from any country. However, we were interested in analyzing the technological preferences of people of the developing countries. Having first-hand experience and knowledge of Pakistan made it our obvious choice. Furthermore, our ultimate aim is to analyzing the significant latent topics and patterns of tweeting of developing countries, and many of the content on Twitter from the developing countries may be having non-English content. As such, understanding of the local content was one of the pre-requisites of our future planned work.
One of the big challenges in this project was to get a significant number of tweets from different districts of Pakistan. For the purpose of this research, we have used the Twitter Search API to get the tweets of all the user with a particular district name in the location field of their profile. Though we wanted to get the tweets of the user belonging to all the 36 districts of Punjab, only the data from 22 districts was available during the collection period (first six months of 2013). Figure 1 shows the population of each district as the percentage of the total population of all the districts being analyzed in the study. Figure 2 shows the distribution of the Human Development Index(HDI) for the year 2013 [5] for the 22 districts of the Punjab province of Pakistan. Figure 3 shows the distribution of the total no of tweets in our analysis while the figure 4 shows the distribution of the number of unique Twitter users across different districts.

Table 1 shows some other results of our analysis. As expected, the districts with the higher education score and higher HDI have a higher percentage of tweets and twitter users. However, looking deeply there are some other

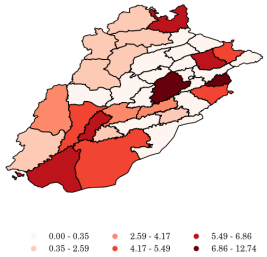

Figure 1: Distribution of population

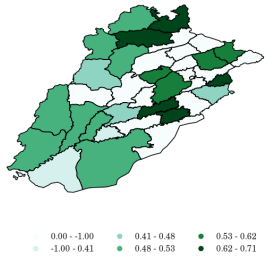

Figure 2: Distribution of human development index

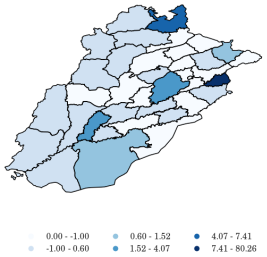

Figure 3: Distribution of total number of tweets

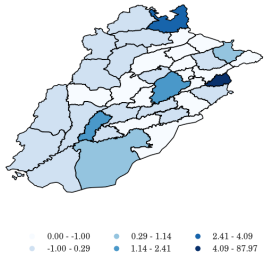

Figure 4: Distribution of nummber of twitter users

District	HDI	Population	Total Twitter Users	Average Tweets Per User	Non English Tweets (%)
Lahore	0.71	6,318,745	74,371	48.01	65.13
Chakwal	0.67	1,083,725	38	15.21	53.63
Rawalpindi	0.66	3,363,911	3,460	95.26	54.58
Sahiwal	0.66	1,843,194	35	4.8	66.67
Pakpattan	0.65	1,286,680	11	1.36	46.67
Sialkot	0.62	2,723,481	960	70.34	63.42
Faisalabad	0.61	5,429,547	2,037	88.92	56.12
Gujranwala	0.57	3,400,940	127	6.46	68.09
Chiniot	0.56	965,124	81	91.28	76.81
Multan	0.53	3,116,851	1,984	80.25	70.21
Muzaffargarh	0.53	2,635,903	51	99.53	64.42
Bahawalpur	0.53	2,433,091	654	77.85	68.66
Dera Ghazi Khan	0.52	1,643,118	1	1	100
Khushab	0.52	9,057,11	62	114.69	79.89
Attock	0.51	1,274,935	204	128.5	73.14
Rajanpur	0.51	1,103,618	27	89.41	85.09
Mianwali	0.51	1,056,620	247	109.03	72.63
Bhakkar	0.48	1,051,456	167	75.47	87.3
Khanewal	0.47	2,376,000	14	4.5	74.6
Lodhran	0.41	1,171,800	7	4.29	63.33
Rahim Yar Khan	0.37	3,141,053	3	3	88.89

Table 1: Twitter usage across different districts.

interesting trends in the data unravel. For example, all the districts with high HDI and population do not have a high number of Twitter users and tweets. For example, Chakwal district has got the second highest HDI, but it has got a quite low number of Twitter users. Similarly, the population of Lahore (HDI: 0.71) is almost twice as much as that of Rawalpindi (HDI: 0.67), but the number of Twitter users from Lahore is 20 times more as compared to that of Rawalpindi. Some of the districts with low HDI (For example, Mianwali) have got a unique Twitter users as compared to the districts with higher HDI.

We have also analyzed the language of the tweets and the Table 1, clearly shows that the districts with lower HDI have got more non-English tweets (Urdu tweets written in Roman script).

Conclusion & Future Work

In this exploratory work, we have tried to analyze the usage of Twitter in different districts of Punjab province of Pakistan. Though this study is exploratory in nature, still quite a few interesting patterns can be seen in our analysis. Twitter does not seem to have the same level of penetration in the districts with similar socioeconomic conditions. This trend indicates the gaps in marketing schemes of the company. However, more interestingly from a social perspective, we see that the people in the districts with lower HDI are interested in tweeting in the national language of Pakistan. Most of the people in these districts either tweet in Urdu us-

ing the Nastaliq script or the Roman script [1]. This fact also highlights the gaps and requirement of improvement when it comes to the interaction of the people of the developing world with technology. For instance, most of the word auto-completion techniques on the smartphones are not able to understand romanized Urdu words.

In the future, we want to analyze the communication patterns of males and females across different districts. We also want to analyze the content of the tweets to find out dominant topics and themes of tweets and whether there is a significant variation in these topics across the districts with different HDI. We also want to see the patterns of tweeting times across different districts and regions.

We are hopeful that our study once completed will not only help the social media companies like Twitter, but it will also help the designers and computational social scientists in general. We are hopeful that our study will result into better design guidelines for the ICTD and HCI researchers working in the developing countries. Furthermore, our results can also act as a proxy measure of educational and technological literacy in the developing world.

References

[1] Joshua Blumenstock, Gabriel Cadamuro, and Robert On. 2015. Predicting poverty and wealth from mobile phone metadata. *Science* 350, 6264 (2015), 1073–1076.

[2] Pierre Deville, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences* 111, 45 (2014), 15888–15893.

[3] Nirav N Patel, Forrest R Stevens, Zhuojie Huang, Andrea E Gaughan, Iqbal Elyazar, and Andrew J Tatem. 2016. Improving large area population mapping using geotweet densities. *Transactions in GIS* (2016).

[4] Philip J. Reed, Muhammad Raza Khan, and Joshua Blumenstock. 2016. Observing Gender Dynamics and Disparities with Mobile Phone Metadata. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development (ICTD '16)*. ACM, New York, NY, USA, Article 48, 4 pages. DOI: http://dx.doi.org/10.1145/2909609.2909632

[5] Social Policy and Development Center. 2016. *Social Development in Pakistan: Annual Review 2014-15*. Technical Report. http://www.spdc.org.pk/Data/Publication/PDF/AR2014-15.pdf

[1] https://en.wikipedia.org/wiki/Roman_Urdu