Vol. 3, No. 1 | January – June 2019

Current Status of Urdu on Twitter

Saqib Muhammad Ghulam[1] Tariq Rahim Soomro[2]

Abstract:
Language is the medium of communication and interaction in words or sign format. Easy access to mobile held devices and popularity of social media has revolutionized the way people used to communicate with each other in the past. The information generated in the past two years is many times greater than the information generated since data storage technology inceptions. The Urdu language has emerged since the 6th century from other languages like Arabic, Hindi, Persian, Prakrit and Sanskrit, and it is the national language of Pakistan. Even after being used widely in all public, official and media use of this language is limited to speaking and reading. The study to analyze the use of Urdu language on social media was conducted during the general elections of Pakistan, 2018. With the expansion of the internet, a new colony of the digital world has come into existence which communicates. This study utilized Twitter as a source of information to assess how Urdu as a language has flourished over this Social Media platform. The results showed the varying use of Urdu and Roman Urdu. The study was able to prove that the language had evolved and integrated even into Social Media and its usage is mostly during occasions and varying durations.

Keywords: *Twitter, Semantics, Sentimental Analysis, Urdu, Social media data*

1. Introduction

Like Facebook, Twitter [1] is one of the most popular social media networking services that began in 2006. Any registered Twitter user can use this platform to post 140-character information called tweet. This tweet has the potential to generate a large number of retweets that can help reach the target audience. Twitter helps registered user to post in any format like audio, video or text. Also, the practice of sharing information from different platforms has added a new format of posting tweets. According to the study [2] researchers analyze these tweets for sentiment analysis whereas another study [3] has tried to predict suggest an appropriate mechanism in predicting election results.

Every registered user communicates in a common understandable language to make his/her tweet reach masses. The medium of communication the registered user uses is known as the language. Every user on twitter if free to post information in any language. The Urdu language is not only the national language of Pakistan but also an official language of India. More than 60.6 million people speak the language, Twitter does not bind them to use their native language thus giving birth to social media language learning (SMLL) [4]. The Urdu language evolved from Arabic, Hindi, Persian, Prakrit and Sanskrit, during 6th to the 13th century. Urdu [5] is also the official language of Pakistan. Urdu uses Arabic writing script, known as 'Nastalique' and resembles Arabic, Persian, and Turkish [6]

As per official statistics of Pakistan, six percent (6%) of the population of Pakistan speaks Urdu language. Urdu started impacting subcontinent when the British Empire in early 1837 abolished the Persian language along with other northern emirates' languages and declared Urdu and English languages as official languages of India [7].

This study was conducted during the Pakistan general election 2018, which was the perfect time to evaluate how much this 18th century language has integrated into digital society and people use this language to express their sentiments as other languages like Arabic has shown during 2012 Egypt uprising

[1] *SZABIST Dubai Campus, United Arab Emirates*
[2] *CCSIS, Institute of Business Management (IoBM), Karachi, Pakistan*

[8] and Turkish during an attempted coup to oust the Turkish president.

Twitter tweets corpse were used as sample in this research to study and develop the required information. However, the major challenge during this study remained as most people prefer Roman Urdu over traditional Urdu in all digital mediums which will make identification of tweets corpse difficult. Selection of common words strategy helped us identify these words in the sample data.

2. Literature Review

During the 1980's social media known as Bulletin Board System (BBS) started its development. The first social media site was SixDegrees.com whose model was later adopted by sites such as Friendster, Myspace and Facebook. Social Media sites have transformed how people communicated over the internet [9]. The term user generated content was first used to describe the matter social media audience create [10]. Twitter came into existence in 2005. Addition of audio and video format played a significant role in transforming text-based internet communication into an interactive and robust format of communication [10]. The interaction happening within social media has created a global village of collective dialogue [11]. To examine the inner sights known as emotions with the dialogues resulted due to events such as political, social, marketing campaigns has led emergence of computing and sentiment analysis [12]. Recently, sentiment analysis or opinion mining is employed to identify the public opinion about people's thought [10]. The purpose of the sentimental analysis is to distinguish between positive, negative or neutral sentiment words [9].

Unavailability of linguistic features of Urdu makes sentiment analysis of social media difficult. However, due to advancement in mobile technology, Urdu-enabled keyboards are becoming common. Following are a few common challenges researchers face:

- Handing Urdu language skill to be upgraded
- Roman Urdu is using English Character set – it is difficult to recognize the language.
- The majority of Tweets consist of English language.

This study also faced major challenges, such as:

a) Dialect Spectrum. Urdu language using four possible dialects, other than Roman-Urdu (using English alphabets to write Urdu). These four dialects are Dakhini, Hyderabadi Urdu, Rekhta, and Urdu. All these four dialects are the same in writing style and script as shown in Table I below.

Table I: Different dialects of Urdu with Respective spellings

Text in Dakhini Hyderabadi Urdu Rekhta Urdu	Roman Urdu	English
پاکستان ایک امن پسند ملک ہے	Pakistan ek aman pasand mulk hay	Pakistan is a peace-loving country

b) Idioms. Interestingly many of the idioms have no place in lexicon dictionary, but are very common words used in the Urdu language; they are occasion/event based and exists within the common community and make their place among them. For example, "دے گھماکے" (De Ghoma-ky) – has no meaning, but commonly referred in Cricket for scoring (hitting).

c) Sentimental Analysis and Subjective
   a. Sometimes names in the Urdu language are adjectives; this issue is very critical to distinguished; for example, "شیر" (sheer or tiger) is an animal, whereas sheer "شیر" in Urdu used to indicate bravery.
   b. Negation handling in the Urdu language is also difficult to distinguish.

d) Twitter specific challenge
   a. Using Roman Urdu on Twitter
      i. Main Qamyab Ho Gaya
   b. English along with Roman Urdu
      i. Main Pass Ho Gaya
   c. Infirmity of language
      i. Hashtags, Uniform Resource Locators, etc.

e) Hashtags. The term tags refer to the characterized topic of interest mainly used to indicate in the social media tools including Twitter, while hashtags allow, for example, for easy analysis of trending topics. Hashtags in Twitter plays a vital role in recognizing the sentiments of people and trends. It also helps specify the interest of the community and showing what is going on by finding the top 10 hashtags [14].

f) Links. These are the pointers to link with other pages and/or applications. Technically speaking tweets are

made of limited lengths, so it may contain compressed hyperlinks.

g) Mentions. They are represented in Twitter by using "@" followed by the username. This pattern is the same as the Facebook tagging pattern.

h) Conversation: It allows direct interaction with the user as a reply to other user's tweets. These replies are only seen by the users; who either follow them or the user to whom reply is sent.

i) Known methods. Twitter with more than 261 million users and a penetration rate of 31.9% is considered to be a high speed growing social media network [15]. More than 300 billion tweets are already shared online till today. Twitter has been more likely being used by mobile users with data package than on any other device. The research found that the English language accounted for more than 50% of social media use on the Internet [16]. Other than English language support was the huge milestone announced by Twitter; they announce 34 language support platform to accommodate non-English language speakers; however, due to the advancement of operating systems, Twitter allows their users to write tweets in their native language along with the English language [17].

3. Material and Methods

Urdu is the commonly spoken language of the sub-continent and national language of Pakistan. This study will explore use of Urdu in the digital world.

Research Instrument

1) Python. *It is no denying fact that Python is mainly used for data analysis and it is a very powerful programming language. Python language constitutes a rich library and used for this study to monitor the status of Urdu language tweets on Twitter. In this study python version, 3.5 has been used.*

2) *Tweepy API. It is an open-sourced API, freely available at GitHub. Tweepy API enables Python to communicate easily with the Twitter platform. Tweepy version 3.6.0 was used at the time of this study.*

The Urdu language sentimental analysis using Twitter is also part of this study. This analysis yields three results as:

- Positive: Completed my graduation.
- Negative: Didn't perform well in Exams
- Neutral: Schools are closed due to heavy rain

As of 2018 World Bank, the population of Pakistan was 220 million and as per Internet World Statistic (2018), it has 44.6 million Internet users and the Internet penetration rate is 22%. There are 35 million active social media users with penetration rate is 18%. There are 109.5 unique mobile users and penetration rate is 55%, while there are 32 million Active mobile social media users and the penetration rate is 16%. Gender wise 77% of Pakistani users are male and 23% are female. According to tribune 72% of users 'log on' on a daily basis. In Pakistan there are about 3.1 million Twitter users [18].

Table II: Keywords selected for testing in twitter streaming API

English word	Urdu word
Pakistan	پاکستان
Imran Khan	عمران خان
Nawaz Sharif	نواز شریف
Bhutto	بھٹو
Justice	انصاف
Dam	ڈیم
Pakistan	پاکستان
Imran Khan	عمران خان

4. Results and Findings

If compared with international languages, Urdu shares a very minimal share in the digital world.

A. Research Instrument

Twitter was chosen as the source of information and three tools were developed for this research. The first tool in Python was developed to download filtered tweets. The second tool was developed for sentiment analysis and displaying bar chart in percentage for the use of common words. Finally, a third tool was developed to view the location of active Urdu users around the world. Common words used during data collection are shown in Table II.

Table III below depicts the result of sentimental analysis of English words commonly used in Pakistan. Here in this study keyword "Pakistan" strength was highest i.e. 42%, whereas keyword "Justice" strength was lowest i.e. 14%. People were more positive about "Dam" with 25%, whereas, the keyword "Imran Khan" shows sentiments of 9:7; here 9 means positive and 7 means negative.

Table III: Sentimental analysis of English tweets

S. No	Search Term	strength	sentiment	passion	reach	Sentiment Analysis		
						Positiv	Neutral	Negativ
1	Pakistan	42 %	9:10	6	43	9	182	10
2	Imran Khan	17 %	5:2	12	18	15	203	6
3	Nawaz Sharif	22 %	9:7	9	23	9	218	7
4	Bhutto	20 %	11:8	26	20	11	228	8
5	Justice	14 %	1:1	9	15	14	157	14
6	Dam	22 %	25:8	0	23	25	227	8

Table IV:Sentimental Analysis of Urdu Tweets

S. No	Search Term	strength	sentiment	passion	reach	Sentiment Analysis		
						Positiv	Neutral	Negativ
1	باکستان	13%	11:3	0	14	11	202	3
2	عمران خان	2%	5:8	3	3	5	267	8
3	نواز شریف	0%	2:1	0	0	4	77	2
4	بھٹو	11%	1:0	0	12	7	138	0
5	انصاف	17%	4:3	26	18	20	316	15
6	ڈیم	20%	1:1	3	20	7	173	7

Table IV depicts the result of sentimental analysis of Urdu words commonly used in Pakistan. Here in this study keyword "ڈیم" strength was highest i.e. 20%, whereas keyword " نواز شریف" strength was lowest i.e. 0%. People were more positive about "انصاف", whereas, keyword "Imran Khan" shows sentiments of 11:3; here 11 means positive and 3 means negative.

Figure 1, below depicts the most popular hashtags, which reflects the thinking of Pakistani people. This figure depicts that people are twitting 'Pakistan' and more concerned than anything else.

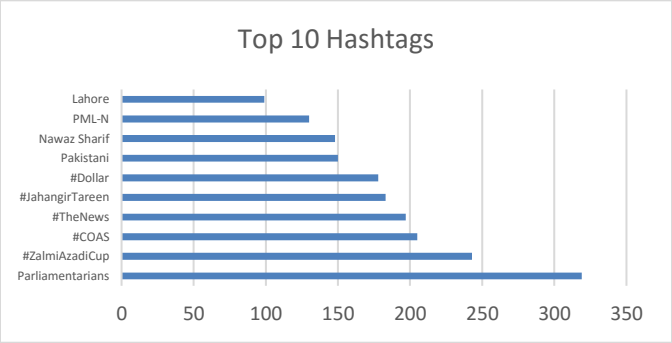

Figure 1: Top 10 keywords from 2000 tweets

Figure 2 depicts the top 10 keywords collected, where Urdu stands tops of the list for three hours, from 6:00 am to 9:00 am. The Urdu word (دہشت گردی) "terrorism" shows a strange of 7 positives; on the other hand, the English word "Terrorism" for the same meaning given 7 negatives.

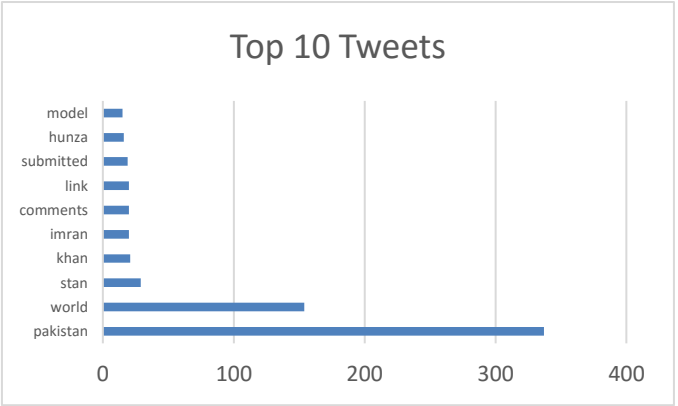

Figure 2: Top 10 keywords from 1200 tweets

Figure 3 below, depicts the running filter only on Urdu tweets mentioning Pakistan in Urdu and English without location restriction; revealed 93.27% users use English text and 6.73% use Urdu.

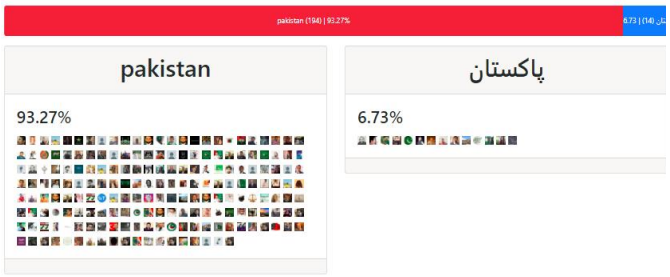

Figure 3: Shows the percentage of mentions of two similar words in different language between 21:00 and 00:00 UAE ST

The Figure 4 below depicts the running filter only on Urdu tweets mentioning Pakistan in Urdu and English, with location restriction (set as Pakistan), revealed that less 34.89% users use Urdu text and 65.11% use English.

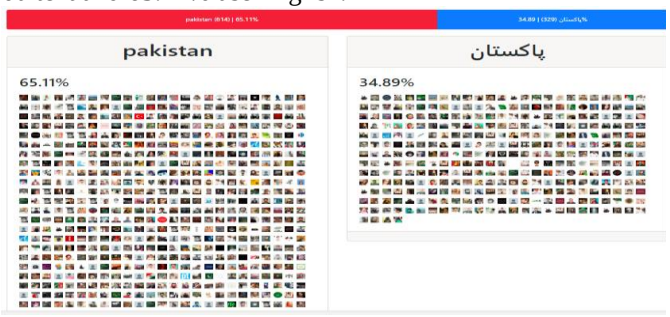

Figure 4: Shows the percentage of mentions of two similar words in different language (21:00 UAE ST)

The Figure 5 depicts the running filter only on democracy in Urdu and English language, with location restriction (set as Pakistan); found that English users scored 96.7% while Urdu users scored 3.3% (14:00 UAE ST).

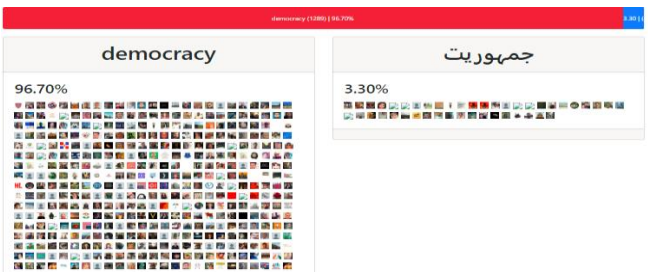

Figure 5: Shows the percentage of mentions of hashtags of two similar in different language (21:00 UAE ST)

5. Discussion & Future Work

People all around the world speak in their native language and with the popular use of social media tools, people tend to communicate in English as English is the native language of social media. Urdu is one of the popular languages, which deserves to be used on social media along with the English language. This study was conducted to identify where Urdu stands along with social media language English? Twitter was used as a social media tool to confirm findings. The study was not able to identify the exact number of tweets generated in Pakistan and was restricted by privacy control policy of Twitter. Therefore, this study assumes that there are possibly more users than the number Twitter claims. Information on the Internet in the Urdu language must be verified using a content credibility measuring system, which makes sure that the information posted is in the Urdu language. This can be accomplished using a rank or scoring system on information, which is being collected and forwarded. Currently, research on Urdu sentiment analysis is limited and the language is still trying to intercept itself in this growing digital field.

This study depicts the initial step towards Urdu sentiment analysis using Twitter. The retrieved tweets during this study were analyzed to provide their sentiments polarity (positive or negative). Urdu has also issues with Lexicon based sentimental analysis and this is the possible future area, where researchers can start their work. Thus this study concludes based on findings that the Urdu language is not strange to social media tools, such as Twitter. The Urdu language used infrequently on social media along with English. Urdu language users use Urdu on certain normal or abnormal events occurrence while happening in their lives, but commonly these users use English more than Urdu.

ACKNOWLEDGMENT

An earlier version of this paper was presented at the International Conference on Computing, Mathematics and Engineering Technologies (iCoMET 2018) and was published in its Proceedings available at IEEE Explorer. https://ieeexplore.ieee.org/document/8346370

REFERENCES

[1] "Company," Twitter, [Online]. Available: https://about.twitter.com/company. [Accessed 20 01 2016].

[2] N. Jabbari, A. Boriack, E. Barahona, Y. Padron and H. Waxman, "The Benefits of Using Social Media Environments with English Language Learners," in *Proceedings of Society for Information Technology & Teacher Education International Conference 2015*, 2015.

[3] P. Singh, R.S. Sawhney and K. S. Kahlon, Sentiment Analysis of demonetization of 500 & 1000 rupee banknotes by Indian governemtn, 2017, The Korean Institute of Communication and Information Sciences, https://doi.org/10.1016/j.icte.2017.03.001

[4] E. Bruce , F. Rob and P. John , "Mapping the Arabic Blogosphere: Politics, Culture and Dissent," *New Media and Society*, vol. 12, no. 1225-1243, pp. 1225-1243, 16 06 2010.

[5] S. Hussain and D. N, "Urdu Word Segmentation.," in *1th Annual Conference of the North American Chapter of the Association for Computational Linguistic*, Los Angeles,, 2010.

[6] K. Riaz, "Stop Word Identification in Urdu.," in *onference of Language and Technology*, Bara Gali, 2007.

[7] R. PAREKH, "Dawn News," Dawn , 17 12 2011, . [Online]. Available: http://www.dawn.com/news/681263/urdus-origin-its-not-a-camp-language. [Accessed 03 11 2015].

[8] Dr.Balasaravanan.K. (2018). TWITTER SENTIMENT ANALYSIS. International Journal of Pure and Applied Mathematics, 1785-1791.

[9] H. Al Suwaidi, Tariq Rahim Soomro and K. Shaalan, "Sentiment Analysis for Emiriti Dialects in Twitter", Sindh University Research Journal (Science Series), vol. 48, no. 4, 2016.

10 A. Lenhart, M. Simon and M. Graziano, "https://www.pewinternet.org/," 2001. [Online].

11 "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210-230, October 2007.

12 M. Berns, in *Concise Encyclopedia of Applied Linguistics*, Oxford, UK, Elsevier Ltd, 2010, p. 557.

13 E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE*, vol. 31, no. 2, pp. 102-107, 2016.

[14] H.-C. Chang, "A new perspective on Twitter hashtag use: Diffusion of innovation theory," *Proceedings of the American Society for Information Science and Technology*, vol. 47, no. 1, pp. 1-4, 2010.

[15] "Statista," Statista, 01 02 2016. [Online]. Available: http://www.statista.com/statistics/303684/regional-twitter-user-distribution/. [Accessed 01 02 2016].

[16] C. Honeycutt, "Beyond Microblogging: Conversation and Collaboration via Twitter," in *47th Hawaii International Conference on System Sciences*, Hawaii , 2014.

[17] [Online]. Available: https://dev.twitter.com/web/overview/languages. [Accessed 03 03 2019].

[18] "Pakistan Social Media Stats 2018", [Online]Available: http://alphapro.pk/pakistan-social-media-stats-2018/ [Accessed 03 03 2019].