An Analysis of Twitter users of Pakistan

Rehan Khan, Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan, Email: rehan_mlk@yahoo.com

Hikmat Ullah Khan[1]*, Department of Computer Science, COMSATS Institute of Information Technology, Wah, Pakistan. Email: Hikmat.Ullah@ciitwah.edu.pk

Muhammad Shehzad Faisal, Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan. Email: Hikmat.Ullah@ciitwah.edu.pk

Khalid Iqbal, Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan. Email: khalidiqbal@ciit-attock.edu.pk

Muhammad Shahid Iqbal Malik, Department of Computer Science, COMSATS Institute of Information Technology, Attock, Pakistan. Email: m.shahidiqbalmalik@ciit-attock.edu.pk

***Abstract*- Web mining analyzes web content, its usage and structure. The users' behavior, interaction and generated content analysis have vast applications such as market analysis, social issues examination and study of human behavior. Twitter is a widely used social network that provides short message facility to its users to generate their own content. There are a number of research works about twitter data analysis, but the relevant literature lacks to analyze the data of Twitter users of Pakistan. In this work, we have prepared the Twitter dataset of users of Pakistan. Then, we have performed user profile analysis and statistical analysis. The user profile analysis covers users' account verification, region as well as province-wise analysis. The statistical analysis focuses to analyze certain Twitter features such as hashtags, user mentions. In addition, the top active users and influential users have also been identified. This research discusses the users' issues and is helpful to understand society, people and their behaviors. We find that Twitter users in Pakistan are different as they interact with real life relations, share religion and sports related content, whereas the rest of the world users follow celebrities such as actors, musicians, and political personalities. In the future, we intend to analyze the content and target to find the top trending topics discussed by Twitter users in Pakistan.**

I. INTRODUCTION

Internet users are shifting from traditional web to the social web. The social web provides various platforms to its users to form virtual communities, such as Twitter, Facebook, YouTube, Myspace, LinkedIn, and Google+. A social web platform allows its users to post content of their interest and view the content of their community where they can express their personal views related to any topic. Social network websites provide a platform for users to discuss the political conversation [1]. People use social web platforms to express their feelings about events, share their opinions about different topics. Therefore, the social web is a useful resource to analyze the public views. Users create reviews about different products and services to provide guidance for others before purchasing a product. The created reviews by different users are useful for marketing and advertisement [2]. The promotion of products at low cost is an advantage of the social web. Similarly, users exchange information and ideas about different topics or highlight the issues that are happening around them. The finding of top users helps the business person to proactive [3]. The smarter marketing campaigns are created after discovering theses hot topics and it is

[1] * Corresponding author

also predictable about consumers' opinion about the products. Twitter is one of the most widely used social web platforms and has become the de-facto micro-blog that supports short messaging. These messages are called tweets. The Tweets are shown at each of the user's profile. Using Twitter, users usually follow each other in a graph patter, where each follower gets alerts about any new tweet. However, tweets remain available for everyone to explore, browse, read and reply. There are a number of research works that provides analysis of Twitter users, but we did not find any work that aims to analyze the Twitter users of Pakistan.

In this paper, we perform the profile analysis of Twitter users. At first, after, Twitter data extraction and preprocessing, a new dataset of Pakistani users is created. The verified twitter user accounts are used to gather information regarding their provinces and regions to perform demographic analysis. The statistical analysis is performed to explore significant twitter features. Similarly found out the top ten influential users and active users. The behavior of Pakistani users is compared with the rest of the world. This analysis can help to understand the people and society.

The rest of the paper is organized as follows: Section 2 review existing related works of twitter analysis, Section 3 provides preparation of our dataset, Section 4 presents the proposed research methodology. In Section 5, results are discussions before concluding the paper.

II. RELATED WORK

Twitter is one of the widely used social networks in the world after its launching in in 2006. We find a number of research works that analyze the Twitter data. The observation of micro-blogging features is carried out by observing geographical characteristics of twitter [4] .The users do micro-blogging by talking about their daily routines and sharing the information and having similar interest can interact with each other [5]. The objective of using twitter has been categorized into conversations and content sharing [6]. At the same time some other researchers draw a distinction among users around a number of evidences and these evidences are the usage of hashtags, @mention and retweeting [7]. Temporal and spatial attributes are very significant for orthogonal techniques. The main stream of social media is what people are thinking about a product monitored. The objective was to know about users' review for a particular product. Method of identifying significant emerging phrases associated with products, companies, or masses of interest were demonstrated [8].

According to a research analysis [9], 80% tweets are related to news and on average one tweet was retweeted by almost 1000 users regardless the parameter of their popularity. The follower – followee topology has been used to identify influential users by two ranking standards, i.e., the number of followers and their in-degree analysis. Sina Weibo is one of the most popular social sites in China. The key topics at Sina Weibo and twitter [10] are compared and following points are resulted. The difference existed between the sharing content of China users and the rest of the world. The Jokes, images and links of the videos are retweeted and later they become the hot topics in china on the basis of retweeting factor. 16 % accounts are verified by top trend setters that means unverified users' accounts average is high.

The hundreds and thousands of users in the world use twitter for real time information source. One of the most significant characteristics of news on twitter is the spatial nature of factors that are indulged in discussion or the sources of information. About more than 4000 topics which are popular and less significant were studied and temporal and spatial analysis on the crawled dataset is also performed [11]. A study was conducted about the analysis of users having a high number of followers, their impact in society and effective initiators with popularity of the topics. It was revealed that trendy topics were imitated by the users having in degree.

Many techniques have been applied to observe the geological location of users. A novel technique was introduced using map API with respect to position in the user's profile. Content of political conversation that tweeted on twitter in 2011 in Spain and also in the 2012 US Presidential election are analyzed [12]. They found that the male user ratio is higher in political discussions because males are more interested in politics as compared to females. These results have significant implications for future research on the relationship between social media and politics. But, no work in the relevant literature covers the analysis of Pakistani users.

III. DATASET PREPARATION

There are various approaches for crawling data from twitter which are designed on the basis of particular reasons such as to crawl only tweets content, or tweets content along date and time. We take a novel approach to crawl the data. We have crawled twitter data by setting longitude and latitude for Pakistan geographical area, by using hashtags "Pak" and "Pakistan" as target topics so that not only Pakistani data but also the tweets about Pakisan can also be extracted and thirdly by applying statistics from social bakers. Initially the crawled data saved in JOSN format, then transformed and saved in MySQL for further considerations. The preparation architecture of dataset is presented in Fig. 1.

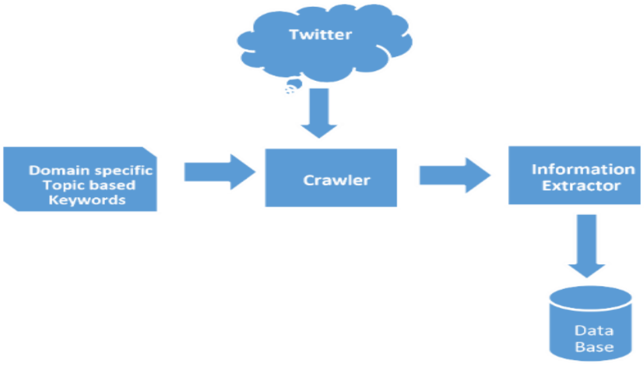

Figure 1: The Architecture of the Twitter Data Crawling Procedure

We encountered many issues with crawling data like having garbage values, improper data combination, missing values and multilingual problems. In our case, the data we extracted was found as noisy, contains redundant and irrelevant information. As per standard pre-processing procedure, the dataset is prepared in the proper format before using it. Each tweet in the dataset contains other attributes as well as that tweet itself. Though there was a flag between each tweets and it is in the form "{}" means tweet started from the symbol"{" and ended at symbol "}". We split each tweet by applying regular expressions as each tweet may or may not contain retweets, second each tweet data is further split and retweets data is extracted from it. Finally, each tweet data contains many attributes, such as Twitter ID, language, tweet text, etc. The characteristics of dataset is presented in Table 1.

TABLE I.

THE STATISTICS OF THE TWITTER DATASET

Feature	Value
Number of Tweets	13065
Number of unique hashtags	2057
Number of unique Users Mentions	1963
Number Of URLs	1161

IV. RESEARCH METHODOLOGY

In this Section, a brief overview of User Profile and Statistical analysis is presented. These analyses further help to better understand how Profile and Statistical analysis is carried out.

A. *Users' Profile Analysis*

In Twitter, user profile shows the detailed information about each user, such as user name, the record about the followers, the followee along with tweets record such as the number of tweets tweeted by the user and further favorite tweet and retweet record. In our dataset, we have Boolean values of the verification process of the user, i.e., verified or not.

B. *Country Province Analysis*

In province-wise demographic analysis, the current location of a user is being shown by an attribute i.e. "location" we have chosen four provinces i.e., Punjab, Sindh, Khyber Pakhtunkhwa and Baluchistan. In addition, the capital terrirory of the Islamabad is also considered.

C. *World Region Analysis*

The United Nations have grouped the countries in the world regions. In our dataset, tweets which are having hashtag Pakistan around the globe, we grouped them into one label of "Pakistan". We consider other four regions as follows South-Central Asia, Middle East, Northern Europe and North America. This analysis is unique in this sense that in this analysis, we are considering the users belonging from various regions of the world who discuss about Pakistan.

D. *Twitter Features Analysis*

A hashtag is an important and widely used feature in social media. In our data set, we analysis the hashtags used by the users in the dataset. Similarly, User mention is also a feature of social media and in twitter as well. It depicts the mentioning of a user by another user, it can be a reply of his.her tweet. We find out the total number of users mentioned in our data set and share the most frequently used hashtags. The third feature is the existence of URLs in tweet content. The link provided in the URL contains the link of web pages, multimedia content such as pictures, videos, etc. We have analyzed the URLs mentioned in tweet content as well.

E. Finding Top Influential User

The number of followers is as the significant feature to measure a users' influence within the community of the social network. In our dataset, we find top influential users by finding the number of followers for each user and then ranking them to share the top influential users.

F. Finding Top Active Users

The active users mean those users of Twitter who share the higher number of tweets. These are the users who generate their own content. We find top active users by in our data set by counting the number of tweets for each user and ranking them and analyzing them against the top influential users.

V. RESULTS AND DISCUSSIONS

In this section, we present the discussion about the users' profile analysis and statistical analysis. Each section is discussed separately. In user profile analysis, the verification of accounts is discussed. Then, country province wise as well as globally region wise is carried out. The Twitter features are statistically analyzed. The top influential users and the top active users are discovered. s

A. Users' Profile Analysis

The users' profiles of Twitter users show the user name, brief information, number of followers, number of followee and his/her friends. The profiles also display the tweets and retweets record of every user. Twitter has two types of accounts, i.e., verified user and regular user. The prepared dataset contains a new Boolean value attribute about accounts of twitter. The true value predicts verified account and the false value represents the account is being used by some fake person which leads to fake information and rumours, as it is happened many times. Twitter only verifies account of popular personalities. These users may be actors, musicians, politicians, sportsmen or businessman. Verification process doesn't consider the in-degree measure. In Table 2, around 26% users of specific category have verified accounts and 74% have unverified accounts. It means that the information (tweet) tweeted with their names may be rumour or the account may be fake. On the other side only 4% of total users have verified accounts in public user category where as 96% are unverified accounts. It is not a surprising result because according to the twitter, it concentrates on the celebrity users. The specific user category means the well-known users who are celebrities whereas normal users mean the common persons. It is notable that the common public does not verify their accounts.

TABLE II.
A COMPARISON OF VERIFIED AND NON-VERIFIED ACCOUNTS

Categories	Verified	Unverified
Specific	26%	74%
Normal	4%	96%

B. *Country Province Analysis*

In our dataset, “location” is an attribute which shows the user current location. User can tweet his/her tweets and can update his/her location at any time. We categorized it province-wise analysis as shwon in Fig.2. Four provinces, i.e., Punjab, Sindh, Khyber Pakhtunkhwa, Baluchistan and Islamabad, the capital of Pakistan, are highlighted. It is predicted from Fig. 2 that mostly people living in Punjab use the twitter. Sindh ranked as second. Khyber Pakhtunkhwa and Baluchistan both have same percentage i.e. 1% which is very low. We find that people who are living in urban area, use twitter in mostly whereas twitter users from Khyber Pakhtunkhwa and Baluchistan are only few. As mentioned in Fig.2.

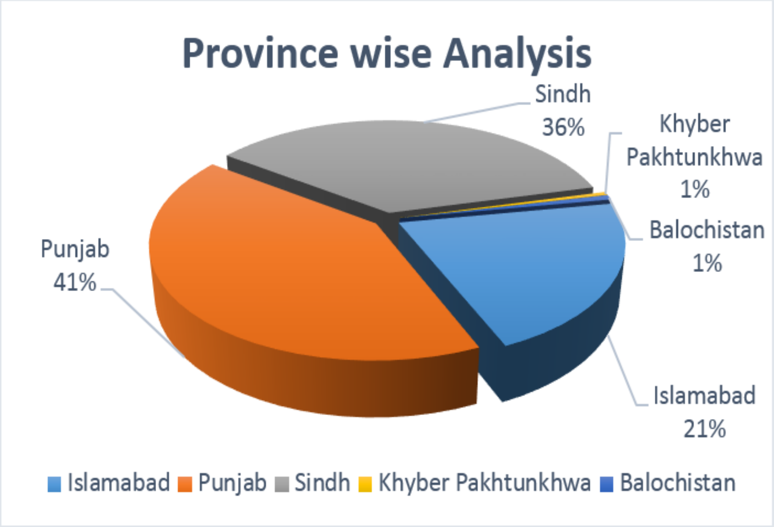

Fig No. 2 Province wise Analysis

C. *World Region Analysis*

The world regions are based on the United Nations country grouping. In our dataset, we consider tweets tweeted with hashtag Pakistan not only from Pakistan but also tweeted from other countries i.e., United Kingdom, United States of America, and Afghanistan etc., to reveal the interest of people from other countries about Pakistan. We group these countries on the basis of regions, later we compared them with Pakistan as shown in Fig. 3. The pie chart is about the world regions and is divided into five categories, i.e., Pakistan, South-Central Asia, Middle East, Northern Europe and North America. Pakistan has the biggest Pie, which is 81%. It means mostly tweets are tweeted from Pakistan. North America has 7% ratio and it is ranked at second position. South Central Asia has 6%, Northern Europe has 4% and Middle East has 2%, which can easily see in Fig No. 3:

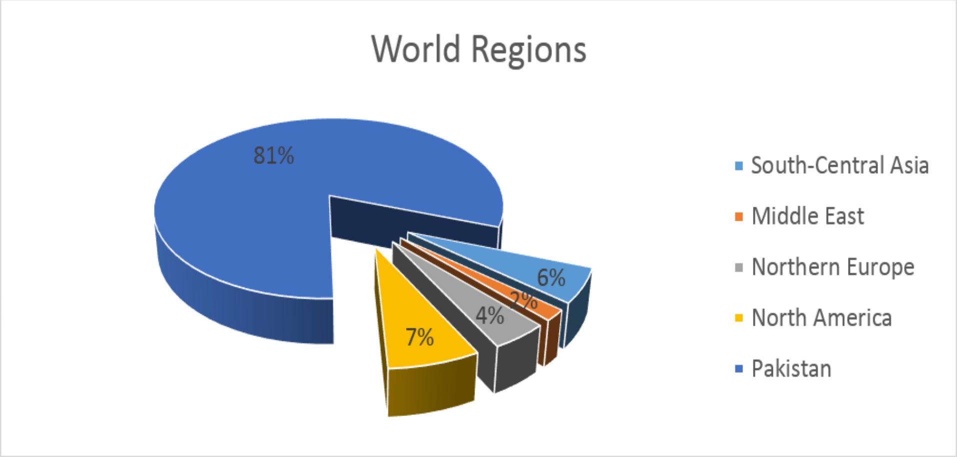

Fig 3: Region-Wise Analysis

D. *Twitter Features Analysis*

In statistical analysis, we conduct twitter feature analysis, compute top influential users, top active users and top ten hashtags. The details about each analysis are given in separate sections. Hashtag is a widely used on many social media sites such as Twitter, Google+, and Instagram. It is basically a topic which is created by the user. These topics are about events or social media product campaign. Table III shows the importance of hashtags as the number of hashtags is higher as compared to other features. Similarly, User mention is a feature of twitter that is reply to other user. It is shown that 1993 unique users were mentioned in the month of February 2014 as the data set contains crawled information for the month of February 2014.The third feature is URLs which contains the link of pictures, videos and web addresses. It is demonstrated that 1161 unique URLs are tweeted in the specified time.

TABLE III.
TWITTER FEATURE LIST AND FREQUENCIES

No.	Twitter Features	Occurrence
1	Hashtags	2057
2	User Mentions	1993

E. *Finding Top Influential Users*

The users who have more number of followers on the social networks are considered as the top influential users. It is taken as a significant measure for fast information dissemination. As shown from Table IV, ESPNcricinfo covers cricket that is major sports in Pakistan has more followers and therefore ranked at top. While Imran Khan, a famous former cricketer and politician ranked at second, have 890141 followers. Another former cricketer, Wasim Akram is ranked at third, has 420003 followers. However Wasim has 50% less followers as compare to Imran Khan.

Ali Zafar, a singer and actor, is ranked at fourth position. From the results of ranking, we can summarize that cricketers and politicians have more followers as compare to other type of users.

TABLE .IV
TOP INFULENTAIL USERS BASED ON NUMBER OF FOLLOWERS

ID	Name	Screen Name	Followers
16542390	ESPNcricinfo	ESPNcricinfo	1023876
122453931	Imran Khan	ImranKhanPTI	890141
210792232	Wasim Akram	wasimakramlive	420003
69229721	Ali Zafar	AliZafarsays	380909
453507741	Maryam Nawaz Sharif	MaryamNSharif	291750
127483019	PTI	PTIofficial	285778
223080656	Sana Bucha	Sanabucha	261232
56650317	Marvi Memon	marvi_memon	244677
233070139	Atif Aslam	Itsaadee	224473
452830010	Asad Umar	Asad_Umar	216346

F. *Finding Top Active Users*

The users who have share more tweets, generate more content, in other words, having higher status count are considered more active. Table V provides the list of top active users. Shahid Afridi, a cricketer, has the highest status count and his twitter id is also verified. Mobilink, a telecom company, is ranked at second in the list. The Events Pakistan website is ranked at third in the list. In this top-10 list, telecom organizations politicians and sportsmen are in majority. Only two cricketers are listed i.e. Shahid Afridi and Azhar Mahmood.

TABLE. V.
TOP ACTIVE USERS BASED ON STATUS COUNT

ID	Name	Screen Name	Status Count
113392039	Shahid Afridi	SAfridiOfficial	5843
57617762	Mobilink	Mobilink	4309
136626769	EventsPakistan.Com	EventsPakistan	3543
484545844	Official Warid	OfficialWarid	3436
97537762	Bakhtawar Bhutto Z	BakhtawarBZ	3372
68352914	Ufone	Ufone	3248
64951214	djuice Pakistan	djuicepakistan	3216
137418466	Azhar Mahmood	AzharMahmood11	2747
98319522	Zong	Zongers	2113
122453931	Imran Khan	ImranKhanPTI	1825

TABLE. VI.

TOP 10 HASHTAGS AND THEIR FREQUENCIES

No.	Hashtags	Occurrences
1	#Pakistan	221
2	# PAK	99
3	# news	33
4	# Karachi	30
5	# IncompetentRulers	27
6	# MQM	27
7	# BanBollywood	26
8	# Bollywood	19
9	# cricket	14
10	# TTP	13

CONCLUSION

In this research work, the dataset of Twitter users of Pakistan is prepared and analyzed. Most frequent users, hashtags, and other Twitter related information have been explored. We perform profile analysis and statistical analysis which can be applied in the fields of marketing, resolving social issues like detection of movement in a society. It is summarized that less number of people use twitter in Pakistan and the accounts of most are not verified. Pakistani users follow sportsman and religious scholars and marketing companies. Whereas the world users follow celebrities such as musician, actors and politicians. In the future, we intend to extend this work by analysis the content using semantic approach [14, 15], create a model to find influential users in more effective manner [16] and taking more features into consideration, apply the sentiment analysis of the tweets content [17, 18].

REFRENCES

[1] Bennett, W. L., & Iyengar, S. "A new era of minimal effects? The changing foundations of political communication", *Journal of Communication*, vol 58, no 4 , pp.707-731, 2008

[2] Bollen, J., Mao, H., & Zeng, X. "Twitter mood predicts the stock market',*Journal of Computational Science*, vol.2, no.1, pp.1-8,2011

[3] Abel, F., Gao, Q., Houben, G. J., & Tao, K., "Analysing user modeling on twitter for personalized news recommendations", *In User Modeling, Adaption and Personalization*, pp. 1-12, 2011

[4] Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J, "Who says what to whom on twitter", In *Proc. of the 20th international conference on World wide web*, pp.705-714, 2011

[5] Bakshy, E., Hofman, J. M., Mason, W. A., & Watts, D. J., "Identifying 'influencers' on twitter", *In Fourth ACM International Conference on Web Seach and Data Mining (WSDM)*,2011

[6] Java, A., Song, X., Finin, T., & Tseng, B.,"Why we twitter: understanding microblogging usage and communities". In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pp.56-65, 2007

[7] Krishnamurthy, B., Gill, P., & Arlitt, M.,"A few chirps about twitter', In *Proc. of the first workshop on Online social networks* pp.19-24, 2008

[8] Goorha, S., & Ungar, L.,"Discovery of significant emerging trends", In *Proc. of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp.57-64, 2010

[9] Allan, J. (Ed.). (2002) "Topic detection and tracking: event-based information organization", Vol. (12)

[10] Asur, S., Huberman, B. A., Szabo, G., & Wang, C.,"Trends in social media: persistence and decay", In *Proc. International Conference on Weblogs and Social Media*, pp.434, 2011

[11] Yin, Z., Cao, L., Han, J., Zhai, C., & Huang, T.,"Geographical topic discovery and comparison'. In *Proc. of the 20th international conference on World Wide Web*, pp.247-256, 2011

[12] HerdaĞdelen, A., Zuo, W., Gard-Murray, A., & Bar-Yam, Y.,"An exploration of social identity: The geography and politics of news-sharing communities in twitter", *Complexity*, Vol.19, no. (2), pp.10-20, 2013

[13] Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A., "Sentiment strength detection in short informal text". *Journal of the American Society for Information Science and Technology*, vol.61, no.12, pp.2544-2558, 2010.

[14] H.U. Khan, S.M. Saqlain, M.Shoaib, M.Sher, "Ontology based semantic search in Holy Quran", *International Journal of Future Computer and Communication*, Vol. 2, no 2, pages 1-6, 2013.

[15] H.U.Khan, T.A.Malik, "Finding resources from middle of RDF graph and at Sub-query level in suffix array based RDF indexing using RDQL queries", *International Journal of Computer Theory and Engineering*, Vol.4, no.3, pages 369-375, 2012.

[16] H.U. Khan and A. Daud, T.A.Malik, "MIIB: A metric to identify top influential bloggers in a community", *Plos One*, vol. 10, no. 9, 2015.

[17] U. Ishfaq, H.U. Khan, K. Iqbal, "Modeling to find the top bloggers using sentiment features", in *Proc. International Conference on Computing, Electronic and Electrical Engineering* , pp. 227-233, 2016.

[18] H.U.Khan, A.Daud, "Using machine learning techniques for subjectivity analysis based on lexical and non-lexical features", *International Arab Journal of Information Technology,* Vol. 14, no. 4. 2017.