I.J. Education and Management Engineering, **2018, 2, 31-40**
Published Online March 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2018.02.04

Available online at http://www.mecs-press.net/ijeme

TweetRush: A Tool for Analysis of Twitter Data

Avnish Dawar[a,*], Archana Purwar[b], Nikhil Anand[a,], Chirag Singla[a]

[a]*Student, JIIT Noida, India*
[b]*Assistant Professor, JIIT Noida, India*

Received: 28 August 2017; Accepted: 19 December 2017; Published: 08 March 2018

Abstract

Twitter network has millions of users spreading information in the form of 140 character messages called tweets. And each user expresses his or her opinion with the tweet, these tweets have been used to know a person's state of mind, get recommendations and also predict the pattern. But a research is an effective one only if its results can be easily understood and a clear understanding requires visualization of the inferences. There isn't any data-graph, pie chart or a tree depicting the results of twitter analysis. Hence this paper suggests "TweetRush" tool to analyze twitter data. It is able to find the influence of a particular user in his network. Graphs help us determine a user's outreach. This tool will help advertisers to target the exact audience; budding entrepreneurs can make use of the influential factor to market their start-ups.

Index Terms: Twitter, internet, social media, tweet, analysis, and graph.

1. Introduction

Twitter is one of the most important platforms for social interaction having around 310 million monthly active users. It allows the users to express their views/opinions and also spread important information in the form of text, pictures or videos. It is easily accessible on gadgets with latest technologies and electronic devices that constitute the major part of our lives. Twitter is such a medium where a common man can reach the highest authorities and even big celebrities with just one single tweet. Users from all around the world can show their views by retweeting a tweet. Twitter has the relationship of followers and following.

Twitter data till now has been used to get statistics on a particular topic, know the recent trends, predict information, and get live polls etc. But the statistics only work on drawing conclusions based on the content of a tweet. But, all the tweets do not get much popularity. Hence, it is very important to determine whose tweets

* Corresponding author.
E-mail address: avnishd380@gmail.com

are popular and influential in a network or to how many people, a piece of information reaches. Therefore, this paper aims at providing the analysis, which user has more influence and is able to reach a greater network.

Rest of the paper is organized into 4 sections. Second section gives the related work of the work under study. Proposed Methodology in detail is described in section 2. Section 3 propounds experimental framework and results and lastly section 5 concludes and tells about the future scope of our research. At the end, refer to the Acknowledgements, References and Appendix.

2. Related Work

There are different conventions being followed by different people, some tweet, some retweet or some use '@' or '#'. Every user has a different reason for choosing a different style. And this leads to alteration of the original content of a tweet.

Twitter data has been used to analyze the kind of tweets, a user does and how it leads to ambiguity and fake messages being spread.

Various alterations to retweets are,

Ego retweets: are the ones when people retweet only those tweets that refer to them. The marketers study ego tweets to know the positive and negative view on a brand. [2]

Broken Telephone: When the original message of the tweet is lost. While retweeting, people modify the content of the tweet according to their own understanding which alters the original information being spread. [2]

A model similar to Tweetrush, based on predicting the flow of tweets is called Matchbox. It was originally developed to predict the movie preferences of the users based on the meta-data about movies. If the model has very strong predictive power, then pi = yi and the negative log-score is zero. If the model is not perfect, then the negative log-score will increase. Therefore, a smaller negative log- score means better model performance. [5]

Automatic detection of tweets that provide Location-specific information will be extremely useful in conveying geo-location based knowledge about the users. Location Centric Word Co-occurrence is a weighting scheme, which uses mutual information score of tweet bigrams; the tweet's inverse document frequency (IDF); the term frequency (TF) of tweets, and the user's network score to determine the location-specific tweets. Bigram sequences are the different combination of keywords, numbers, hashtags used by users while tweeting. [8]

Twitter data is known for its diverse uses, it is also used to predict real time events. One such model is Temporal model, which describes the time series of data. Each tweet has its own post time, depending upon the quantity of tweets of a particular location probability density function is calculated and then plotted exponentially. This can be used to predict the occurrence of events like earthquake, etc. in a particular area. [6]

As termed by the scientist Bollen, "A tweet is a microscopic, temporally-authentic in- sanitation of sentiment". Since tweets are crispy and brief, the public sentiment can be easily explored. Sentiment analysis, also known as opinion mining refers to the use of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. One of the approaches of sentimental analysis is the A-NEW based approach, which uses its dictionary to provide pre- existing, normative emotional ratings for 1034 words along the three dimensions of valence, arousal and dominance. These values were then plotted in a 2D emotional circumplex model. The tweet emotion was determined its position within the model. [9]

For a better insight about the sentiment analysis, it is important to know three the classes of sentiments,

1. **Positive:** if the positive sentiments are increased, it is referred to be good. In case of product reviews, if the positive reviews about the product are more, many customers buy it.
2. **Negative:** if the negative sentiments are raised, it is avoided from the preference list.

3. **Neutral:** these are neither good nor bad words about the target. Hence it is neither preferred neglected. [10]

But still there is no model depicting the flow of tweets and how information flows from one user to the other. TweetRush aims at depicting the relationship between different users and how are they influencing their followers.

3. Proposed Methodology

3.1. Functionalities of TweetRush

3.1.1. User Login to Twitter

TweetRush enables you to perform analysis of your own timeline. As you login with your twitter username and password, you will receive a unique PIN, which authenticates TweetRush to do the analysis.

3.1.2. Admin Login

After the first implementation, data from your Home Timeline gets stored in the database and can be used to compare the results when analyzing live twitter data. In admin login, a home page is displayed next, showing all the tweets stored in database. This is helpful for the future scope when we would require large number of tweets for mining.

3.2. Algorithm

Timeline of a user shows all the tweets and retweets by different users in chronological order. Now to find the source of a retweet we compare the followers and following of two users, if user1 is following user2 only then user1 can retweet user1’s tweet.

Another constraint is as follows, if user 3 follows user1 and user2 follows both user1 and user3, so user2 will first see the retweet of the user3 because tweets are displayed in descending order.

Hence tweet flow is as such, user1->user3->user2

```
1. for ( i=0; i<no. of users; i++) {
       for (j=position of user-1; j>=0;j--){
        for (k=0;k<size of user-followers; k++){
          if( followers[j][k] == target-user)
              return current-user-index;//from
              multidimensional array } }
              return main-posting-user;
2. get source[ ] and target[ ];
3. Display path graph using vis.js library.
```

3.3. Influential Factor

There are two ways to calculate the IF for any user,

1. IF for an individual user is measured as the number of direct branches a node has. If two people retweet from user A’s tweet then IF for user A is 2. Refer to the example shown in Fig. 1.

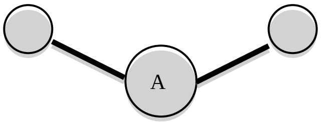


Fig.1. IF for A=2

2. IF for a user in a network, referring to fig. 1., IF of user A was 2, because directly attached nodes are only two. Here we are considering user B, who's retweet was further retweeted two times, giving it individual IF, 2; hence IF for A now becomes 4 (2+2).

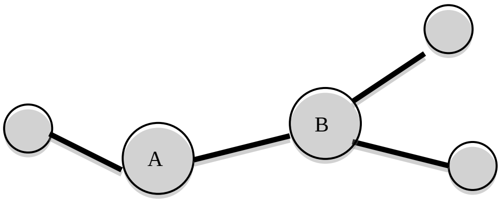


Fig.2. IF for A=4, IF for B=2

4. Experimental Setup and Results

4.1. Twitter data generation

There are different tools and ways to extract data from twitter database, namely Twitter Archive, Bird Song Analytics, Cyfe, NodeXL, TW chat [1] and Rest & Streaming APIs.

API: is Application Programming Interface. It basically works on a formula of give and take; it takes an instruction and then performs an action for the user. It includes a list of commands and can also return various commands.

In reference to twitter,

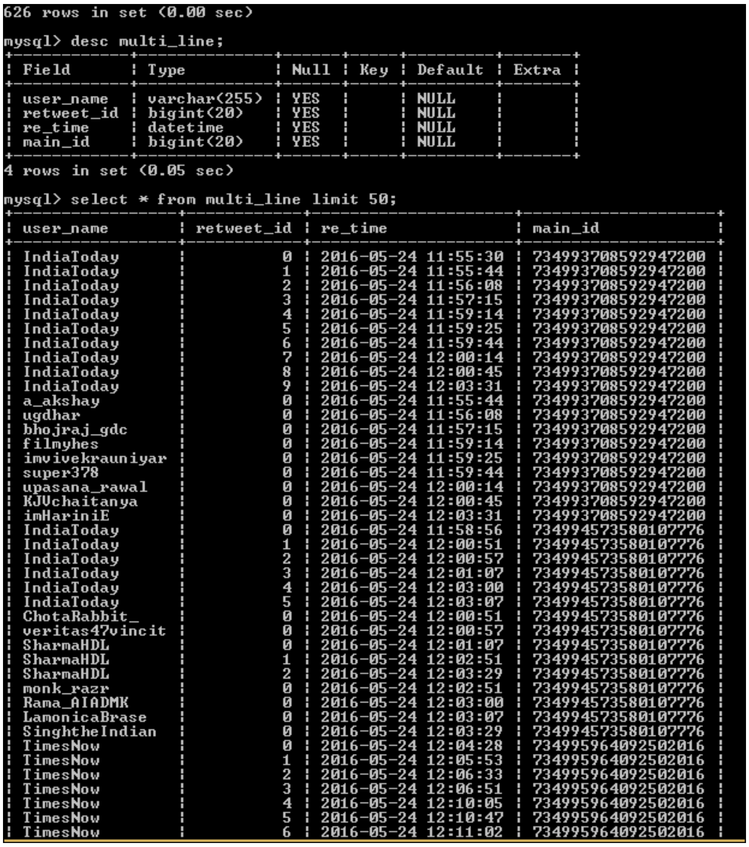
Rest API: Twitter rest API makes use of the authorization account to perform actions on any user's account. It provides the real-time data and allows us to perform various queries on it.

Streaming API: Twitter's streaming API is that it is a long running request, which extracts the data when it is available.

In TweetRush, we have used twitter Rest API to get the real-time data, http://twitter4j.org. (Fig. 3.)

Following are the functions we used to extract data,

- getfollowerslist(): Extracts the list of followers of the logged in user.
- getusertimeline(): Extract the live data of current tweets on the Home page of the logged user.
- getuser(): Extracts user-id of a particular user.
- getscreenname(): Provides screen name/username corresponding to user-ids.
- getretweetcount(): Number of time a particular tweet has been retweeted till now.
- getcreatedat(): Provides the time when a particular retweet was done.



```
626 rows in set (0.00 sec)

mysql> desc multi_line;
+------------+--------------+------+-----+---------+-------+
¦ Field      ¦ Type         ¦ Null ¦ Key ¦ Default ¦ Extra ¦
+------------+--------------+------+-----+---------+-------+
¦ user_name  ¦ varchar(255) ¦ YES  ¦     ¦ NULL    ¦       ¦
¦ retweet_id ¦ bigint(20)   ¦ YES  ¦     ¦ NULL    ¦       ¦
¦ re_time    ¦ datetime     ¦ YES  ¦     ¦ NULL    ¦       ¦
¦ main_id    ¦ bigint(20)   ¦ YES  ¦     ¦ NULL    ¦       ¦
+------------+--------------+------+-----+---------+-------+
4 rows in set (0.05 sec)

mysql> select * from multi_line limit 50;
+-----------------+------------+---------------------+--------------------+
¦ user_name       ¦ retweet_id ¦ re_time             ¦ main_id            ¦
+-----------------+------------+---------------------+--------------------+
¦ IndiaToday      ¦          0 ¦ 2016-05-24 11:55:30 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          1 ¦ 2016-05-24 11:55:44 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          2 ¦ 2016-05-24 11:56:08 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          3 ¦ 2016-05-24 11:57:15 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          4 ¦ 2016-05-24 11:59:14 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          5 ¦ 2016-05-24 11:59:25 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          6 ¦ 2016-05-24 11:59:44 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          7 ¦ 2016-05-24 12:00:14 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          8 ¦ 2016-05-24 12:00:45 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          9 ¦ 2016-05-24 12:03:31 ¦ 734993708592947200 ¦
¦ a_akshay        ¦          0 ¦ 2016-05-24 11:55:44 ¦ 734993708592947200 ¦
¦ ugdhar          ¦          0 ¦ 2016-05-24 11:56:08 ¦ 734993708592947200 ¦
¦ bhojraj_gdc     ¦          0 ¦ 2016-05-24 11:57:15 ¦ 734993708592947200 ¦
¦ filmyhes        ¦          0 ¦ 2016-05-24 11:59:14 ¦ 734993708592947200 ¦
¦ imvivekrauniyar ¦          0 ¦ 2016-05-24 11:59:25 ¦ 734993708592947200 ¦
¦ super378        ¦          0 ¦ 2016-05-24 11:59:44 ¦ 734993708592947200 ¦
¦ upasana_rawal   ¦          0 ¦ 2016-05-24 12:00:14 ¦ 734993708592947200 ¦
¦ KJVchaitanya    ¦          0 ¦ 2016-05-24 12:00:45 ¦ 734993708592947200 ¦
¦ imHariniE       ¦          0 ¦ 2016-05-24 12:03:31 ¦ 734993708592947200 ¦
¦ IndiaToday      ¦          0 ¦ 2016-05-24 11:58:56 ¦ 734994573580107776 ¦
¦ IndiaToday      ¦          1 ¦ 2016-05-24 12:00:51 ¦ 734994573580107776 ¦
¦ IndiaToday      ¦          2 ¦ 2016-05-24 12:00:57 ¦ 734994573580107776 ¦
¦ IndiaToday      ¦          3 ¦ 2016-05-24 12:01:07 ¦ 734994573580107776 ¦
¦ IndiaToday      ¦          4 ¦ 2016-05-24 12:03:00 ¦ 734994573580107776 ¦
¦ IndiaToday      ¦          5 ¦ 2016-05-24 12:03:07 ¦ 734994573580107776 ¦
¦ ChotaRabbit_    ¦          0 ¦ 2016-05-24 12:00:51 ¦ 734994573580107776 ¦
¦ veritas47vincit ¦          0 ¦ 2016-05-24 12:00:57 ¦ 734994573580107776 ¦
¦ SharmaHDL       ¦          0 ¦ 2016-05-24 12:01:07 ¦ 734994573580107776 ¦
¦ SharmaHDL       ¦          1 ¦ 2016-05-24 12:02:51 ¦ 734994573580107776 ¦
¦ SharmaHDL       ¦          2 ¦ 2016-05-24 12:03:29 ¦ 734994573580107776 ¦
¦ monk_razr       ¦          0 ¦ 2016-05-24 12:02:51 ¦ 734994573580107776 ¦
¦ Rama_AIADMK     ¦          0 ¦ 2016-05-24 12:03:00 ¦ 734994573580107776 ¦
¦ LamonicaBrase   ¦          0 ¦ 2016-05-24 12:03:07 ¦ 734994573580107776 ¦
¦ SinghtheIndian  ¦          0 ¦ 2016-05-24 12:03:29 ¦ 734994573580107776 ¦
¦ TimesNow        ¦          0 ¦ 2016-05-24 12:04:28 ¦ 734995964092502016 ¦
¦ TimesNow        ¦          1 ¦ 2016-05-24 12:05:53 ¦ 734995964092502016 ¦
¦ TimesNow        ¦          2 ¦ 2016-05-24 12:06:33 ¦ 734995964092502016 ¦
¦ TimesNow        ¦          3 ¦ 2016-05-24 12:06:51 ¦ 734995964092502016 ¦
¦ TimesNow        ¦          4 ¦ 2016-05-24 12:10:05 ¦ 734995964092502016 ¦
¦ TimesNow        ¦          5 ¦ 2016-05-24 12:10:47 ¦ 734995964092502016 ¦
¦ TimesNow        ¦          6 ¦ 2016-05-24 12:11:02 ¦ 734995964092502016 ¦
```

Fig.3. Data Generation

4.2. Results and Analysis

Home Page

- A timeline displaying all the tweets, one of the important parts of our project. It shows real time tweets, which can be closely monitored and can be used for various purposes.
- By doing sentiment analysis we can get specific tweets, we can evaluate the response of the general population on a new law being enforced or policy implemented or crime being committed or to get reviews of a newly launched product.

User-Tweet flow graph

- Your twitter home page is comprised of the tweets/retweets of all the users you are following; ‘View Graph’ option displays the flow of a particular tweet from its source to all the users who retweeted it.
- Source node is taken in the center; a line is drawn between the source node and every different user node that directly retweeted the source tweet.

If user B retweets the retweet of a user A, in this case user A is considered the source node for user B and a line is drawn between them.

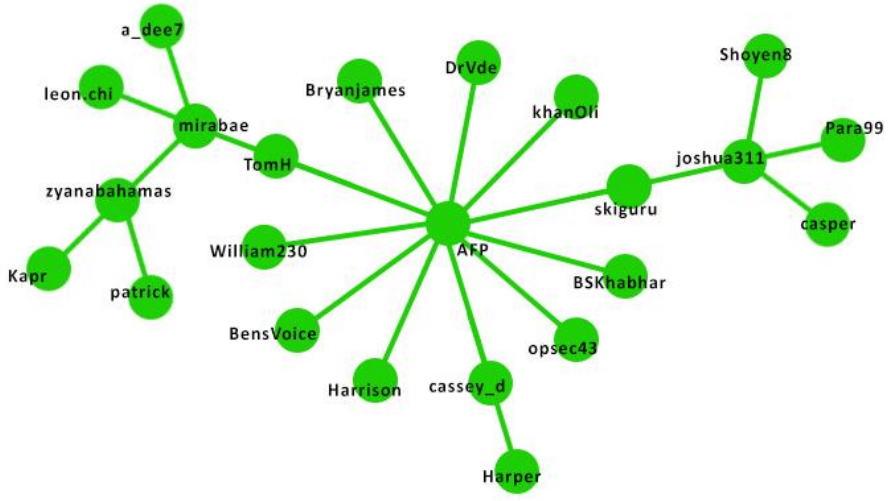


Fig.4. User-Tweet Flow Graph or Path Graph

Individual Analysis

Below is the explanation of making of the Individual-user Analysis graph (Fig. 5.) in the form of its pseudo code.

- Take the target user
- Match with source array
- If source array contains that user

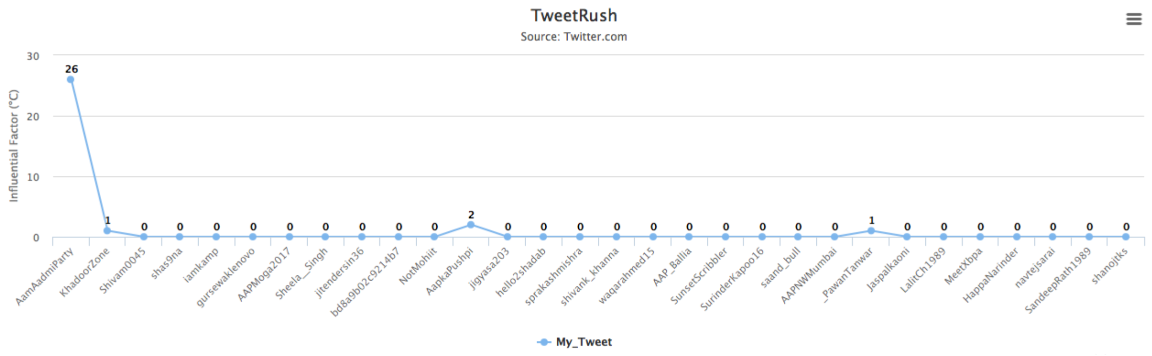


Fig.5. Influential Factor vs. User-name

Multiple-user Analysis

Variation of Influential factor with each retweet is calculated and displayed with this graph (Fig. 6.). A different colour is user for each user. Below is the explanation of making of the Multiple-user Analysis graph in the form of its pseudo code.

- Create Multi-dimensional array of
  a. User-name
  b. Tweet-time
  c. Influential-factor

- for (i=0; i<source-array size(); i++){
  if (i==0)
  add data of source in user-time.

- Append data of user-tweet-time and influential factor in user-time element.}

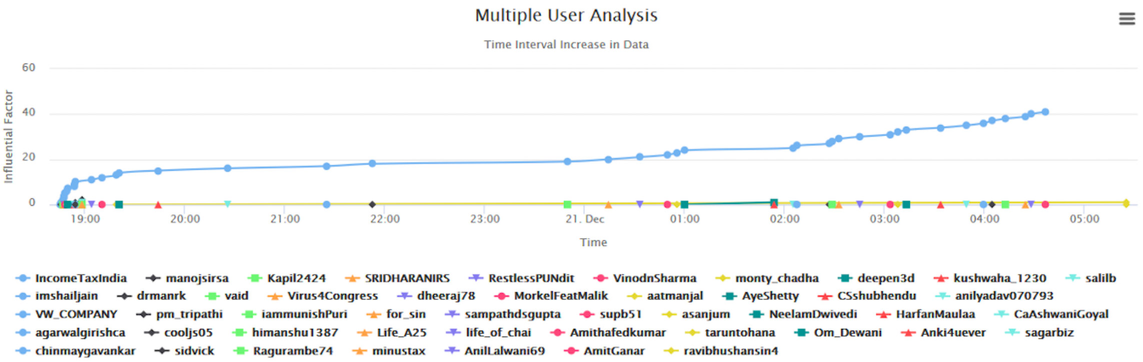


Fig.6. Multiple User Analysis

5. Conclusion and Future Scope

Twitter being the most used social media platform is used for different kind of analysis. TweetRush analyzes the twitter data and provide results useful to marketers and advertisers. Line graphs User Tweet Flow and Influential factor vs Username can be used to give recommendations, hence maximum and minimum out reach of a tweet is calculated. How beneficial a user will be in spreading the tweet? Constraint to this tool is only of limited data; the methodology works fine and provides genuine results.

Limitations

Security and Privacy of users is of utmost concern for Twitter hence it restricts the extraction of data up to 100 recent tweets; therefore, not allowing us to show the analysis of more than 100 users.

Twitter 4j API requires authorized access token to extract any kind of data and it is restricted to the information of only 15users per API call. Although limit value of a single token is reset but it is done only after 15 minutes.

TweetRush can be improved by applying it on large size of data.

Future Scope

Targeting the most influential user across all social media is termed as Influencer marketing. It is easier to look for an influencer on YouTube based on the number of views and subscriptions, but there is no quantitative measure to look for the same on Twitter, so this where TweetRush can be implemented.

TweetRush can also be combined with sentimental analysis to observe the mindset of user with maximum influential factor. In reference to a research conducted on Punjab Legislative Assembly Election 2017, sentiment analysis was used to predict the outcome of the elections, the programming language python was used for extraction of tweets and tweets were categorized based on NLP. Now, the authenticity of this research can be proved, by looking into the Source user of each tweet and each user's network, using TweetRush. [11]

Acknowledgements

We are greatly indebted to Dr. Sanjay Goel, Professor and Head, Department of Computer Science and IT, Jaypee Institute of Information Technology, Sector 62, Noida, under whose guidance this work has been completed. We wish to express our most profound feeling of appreciation to all the professors for their direction, guidance, and consistent support all through the course for this venture work. Without the support of different educators and our companions, help consolation in times of stress, we would not have possessed the capacity to finish this work so easily.

References

[1] Ann Smarty. 5 Tools for Downloading and Analyzing Twitter Data, 2015

[2] Danah boyd, Scott Golder, Gilad Lotan. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. HICSS, 2010, 412.

[3] Georgieva M. How to Use Hashtags on Twitter: A Simple Guide for Marketers?

[4] Doctor V. What Do Twitter Trends Mean?

[5] Zaman TR, Herbrich R, Gael VJ, Stern D. Predicting Information Spreading in Twitter.

[6] Devi DR,and Puduru Devi. A Probabilistic Model of Real Time Event Detection and Reporting. International Journal of Computer Science and Information Technologies 2014, 7614-7617.

[7] Kwak H, Lee C, Park H, and Moon S. What is Twitter, a Social Network or a News Media? Proceedings of the 19th international conference on World Wide Web 2010; 5(6): 591-600.

[8] Rakesh V, Reddy CK, Singh D, Ramachandran MS. Location-Specific Tweet Detection and Topic Summarization in Twitter. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2013.

[9] Bolla RA. Crime pattern detection using online social media. Student Research & Creative Works, Scholar's Mine 2014.

[10] Rani M, Arora J. A Review of Data Analysis of Twitter. International Journal of Advanced Research in Computer Science and Software Engineering 2016; 6:5.

[11] Akhilesh Kumar Singh, Deepak Kumar Gupta, Raj Mohan Singh. Sentiment Analysis of Twitter User Data on Punjab Legislative Assembly Election. I J. Modern Education and Computer Science 2017, vol.9(9), 60-68.

[12] Vinay K. Jain, Shishir Kumar. Towards Prediction of Election Outcome Using Social Media. I J. Modern Education and Computer Science 2017, 12, 20-28.

[13] Mehmood, A., Palli, A. S., & Khan, M. N. A. A study of sentiment and trend analysis techniques for social media content. International Journal of Modern Education and Computer Science, 2014, 6(12), 47.

[14] Deebha Mumtaz, Bindiya Ahuja. A Lexical Approach for Opinion Mining in Twitter. I J. Modern Education and Computer Science 2016, vol.6(4), 20-29.

[15] Jeff Zabin and Alex Jefferies. Social media monitoring and analysis: Generating consumer insights from the online conversation. Aberdeen Group Benchmark Report, January 2008.

[16] Asad Mehmood, Abdul S., Palli, M.N.A. Khan. A Study of Sentiment and Trend Analysis Techniques for Social Media Content. I J. Modern Education and Computer Science 2014, vol.6(12), 47-54.

Authors' Profiles

Avnish Dawar (born January 12, 1995) is an Indian engineering graduate from Jaypee Institute of Information Technology (JIIT), Noida. He is pursuing his post graduate from MICA, Ahmedabad in the field of Communication and Marketing Strategy.

Nikhil Anand (born October 20, 1994) is an Indian engineering graduate from Jaypee Institute of Information Technology (JIIT), Noida. He is pursuing his career in the field of Computer science at Infosys ltd.

Chirag Singla (born October 25, 1996) is an Indian engineering graduate from Jaypee Institute of Information Technology (JIIT), Noida. He is pursuing his career in the field of Technical Consulting at XL Catlin.

Archana Purwar (born March 5, 1979) has been working as an Assistant professor in department of CS/IT at Jaypee Institute of Information Technology (JIIT), Noida for more than 10 years.

Appendix A

A.1. Influential Factor

It is measured in terms of the number of directly connected nodes, in terms of the tweet of a source user, all the users retweeting directly the source-user's tweet are considered while evaluating influential factor. It is denoted by IF.

A.2. NLP

Natural Language Processing (NLP) is the ability of a computer program to understand human language, it is a component of artificial intelligence.

How to cite this paper: Avnish Dawar, Archana Purwar, Nikhil Anand, Chirag Singla,"TweetRush: A Tool for Analysis of Twitter Data", International Journal of Education and Management Engineering(IJEME), Vol.8, No.2, pp.31-40, 2018.DOI: 10.5815/ijeme.2018.02.04