Computation of Spread of Tweets on Twitter

Sachin Gupta, Padmaja Joshi
Centre for Development of Advanced Computing
India
sachinep84@gmail.com
padmajanjoshi@gmail.com

Radha Shankarmani
Sardar Patel Institute of Technology
India
radha_shankarmani@spit.ac.in

ABSTRACT: *Twitter is observed to be a very strong platform to reach more audience. The thoughts that are shared on twitter can have positive as well as negative impact. In this paper the focus is to find and provide some metrics to evaluate the spread of any tweet that is made on twitter. This can be considered as the first step towards identifying the impact of any tweet. Metrics are proposed to compute the spread of a tweet on twitter. The assumption while computing the metric is appropriate API's or web services are provided to get required information regarding the tweet. Spread of a certain real time tweet is also demonstrated in the paper. In this work, the speed of spread is also captured. The speed aspect of a spread may be useful in further analysis of certain tweets under investigation.*

Keywords: Social Networks, Twitter, API, Text Mining, Text Metrics

Received: 19 June 2017, Revised 8 July 2017, Accepted 13 July 2-17

1. Introduction

Online social blogging website twitter gives a platform for broadcasting messages to the world. Messages can be your thoughts, views, feelings anything that you desire to share. These messages known as tweets, have constraint on their size and are restricted to 140 characters. Note that only registered users can tweets though unregister users can view it. Twitter members can follow other user's tweets and broadcast tweets by using multiple platforms and devices. Tweets and replies to tweets can be sent by cell phone text message, desktop client or by posting at the Twitter.com website. Twitter has started to play an increasing role in news from being a source for journalists to providing way to comment on and interact with unfolding events. Overall, we found that the majority of people on Twitter follow some form of news account ï£¡ be it a news brand, breaking news account, journalist or their personal interest from person to brand etc. At the same time, it also possesses a threat of the spreading of unreliable information which done for communal violence or which may turn into riots, commonly referred as a sensitive negative message which turned into "negative tweet". The negative tweet may be defined as a statement which contains the sensitive word, a phrase whose truth value is unverifiable or deliberately false. Sometimes, individual, groups use this platforms for negative cause it includes creating and spreading negative news, rumors, creating negativity by posting and replying tweets or use this platform to spread sensitive wrong information by using offensive words which may lead to communal riots or financial loss of country, individual.

Our proposed work To find outspread (count) of any tweet which having an offensive word by considering retweets and quoted tweets which help us to find out a total number of involved persons by taking followers count,retweet followers count etc.

2. Related Work

Twitter provides vast database, which includes different kinds of topics, conversation, interest and personal information about users. The spread of any topic decides its popularity, impact and cause in long term. Twitter Spread analysis and mechanism discussed in paper which gives an overview to proceed our work

Erick Stattner, Reynald Eugenie and Martine Collard [1] described in his paper that how do we spread any information on Twitter, What condition favors user to forward any message in form of retweet.

Praveen Rao [2] In this paper,Author present a simple and flexible unified framework called SocialKB for modeling social media posts and reasoning about them to ascertain their veracity, a first step towards discovering emerging cyber threats.

Velissarios Zamparas, Andreas Kanavos and Christos Makris [3], This paper presents Twitter platform concerning the Influence of a user which depends on the interest of the Followers (via Replies, Mentions,Retweets, Favorites).

Raveena Dayani, Nikita Chhabra [4], This paper presents rumor detection and its spread by taking history data and perform machine learning algorithms like k-nearest neighbor and naive bayes classifier to detect tweets spreading rumors.

3. Proposed Approach

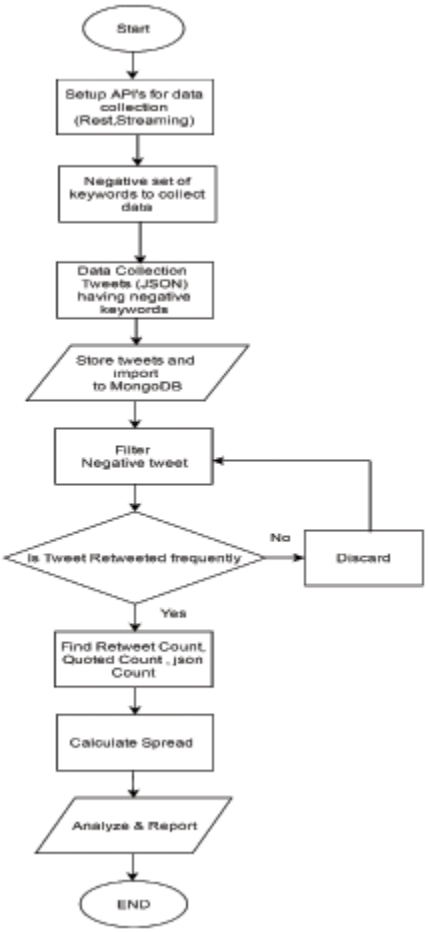

Figure 1. Flowchart of Proposed System

In this section, we introduce some parameters and metrics that can be use to find spread.

1) Spread: A parameter is proposed to get the total count of an engaged user on post via tweet, retweet or quoted tweet. This parameter is required to get how frequently any post is shared and what is the estimated coverage or reachability in social space.we will calculate aggregated count of followers on particular identified negative tweet, retweet or quoted tweet as well as its intensity of spread with respect to time.

2) Retweet count: This count directly refers how audience taking interest in post and feeling to share to others, as much as the tweet is retweeted its reachability and spread will be high.

3) Reply: Reply is a simple message, comment on the post which again helps to make conversation active. It starts with @Twitter-handle of the user. Ex: @sachin how are you, @lavender your post is wrong.

4) Mention: Including any twitter user in conversation even its, not in your follower list just by using @Twitter-handle of the user at any point other than the beginning.

3.1 Spread

Spread defines information extend over a large or increasing area, In our research work it's an important module which decides whether we should continue to dig down about the post or not. If the spread is more it covers more area more audience and chances of more reactions from the audience. Spread decides the popularity of context more popular context spreads faster. The context contains any sensitive information, rumored information or fake information its effects may more dangerous or harmful than our consideration.

Spread starts when anyone interested in post context and starts retweeting, replying or quoting. If person retweet post on his personal wall it spreads to his own followers which increase spread. spread depends on popularity and friend follower ratio too. If the number of followers of any user is more and user publishes any message or shares any tweet ie retweet, it spreads directly to the followers. so we can easily calculate if a negative or sensitive context is spread by the user who has more number of followers it spreads faster and made the negative impact.

Reply to post gives an idea that the user is interacting and its popularity, more replies directly linked with more number of user engaged that is again kind of spread. To get spread count we need to keep eyes on replies having sensitive word as well keep track of who is sharing such type of messages and on how many walls that message is posted by taking count of followers.

4. Architecture

4.1 Data collection Phase

Data collection phase deals with collecting twitter data from different available sources. To collect twitter data we have the wide variety of ways and API's available. The collection of required data and its frequency depends on searching criteria keywords Ex ("riots","bomb","strike" etc). Twitter itself allows you to collect real world tweets by using streaming and rest based API's.

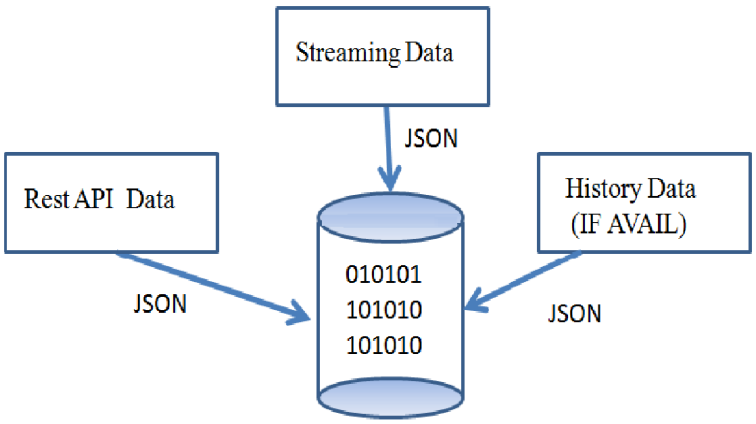

Figure 2. Data Collection Phase

Twitter data collection available in multiple ways: Historic data: Some companies provides paid and unpaid historic data as per need. Streaming API: It allow collecting twitter data Streams of the public data flowing through twitter by pushing data to user Rest API: It allows to read and write Twitter data. The REST API identifies Twitter applications and users using OAuth and return JSON.

4.2 Data Cleaning Phase

This phase deals with cleaning and formatting required data, removing uncompleted JSON and false positive tweets from collected twitter data and make the repository for further processing. In our research work, one level of data filtration done while the collection of data by giving required keyword as search term "bomb", "riots" etc. which helps to filter public tweets and collects only such tweet which is having given words.

4.3 Processing Phase

Processing phase deals with all algorithm and methodology which we are using in research work to find spread, influence user, and futuristic impact.

Spread decides popularity and may help to decide impact of tweet, Twitter spread can be counted by aggregating of various scenarios.

1. Tweet Follower's:

A user tweeted tweet first published to his/her own wall which is counted by considering the number of followers the user has.
Ex: If a user has 4500 followers, and user tweets any message it will directly broadcast to all 4500 followers.

x = Number of followers
$x = 4500$

2. Retweet Follower's:

A retweet is the process of publishing tweeted tweets to your home page which is publically seen by your followers so indirectly, It's spread of others message to your circle by followers.

Ex: Tweet: Hi everyone
Retweet: Hi everyone

Now, this will increment retweet *count* +1 and now it's published to your public wall. so if you have 5000 followers this message is reached to them.

y = Total Number of follower's of retweet ID's

Mention: A Tweet containing another accountï£¡s Twitter username, preceded by the "@" symbol. You can mention other friends and followers to any tweet to include them in conversation.

Without Mention:

Total No. of followers = Tweet source Followers + Retweeted Id's Followers
Total No. of followers = $x + y$

With Mentions:

Total No. of followers = Tweet source Followers + Retweeted Id's Followers + No. of Mentions in conversation.

Total follower without mentions will be calculated by given formula:

$$Totalfollower_{withoutMention} = (TOF + RTF)$$

TOF : Tweet Owner Follower, RTF: Retweet Follower

Spread Calculation

(1) Follower's of Owner who tweeted tweet

$$F_{TO} = \# F_{TOwner} \tag{1}$$

(2) Follower's of Retweeted User's

$$F_{RTO} = \# \sum_{i=1}^{n} F_{RT[i]} \tag{2}$$

Total Spread = $[F_{TO} + F_{RTO}] - [\sum_{i=1}^{n} (\# F_{RTO[i]} \cap \# F_{RTO[i+1]})] + (\# F_{RTO[1]} \cap \# F_{RTO[n]})$

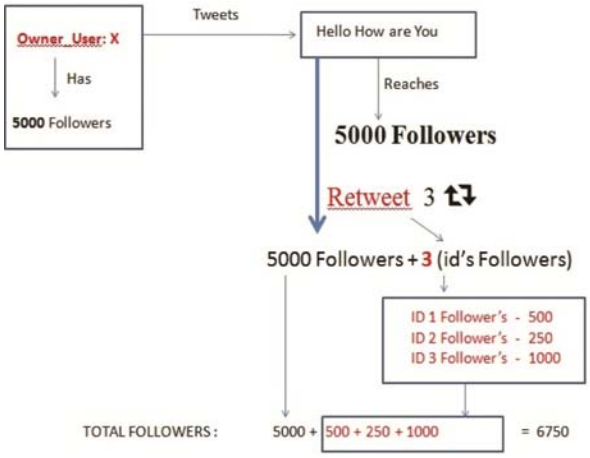

Figure 3. Data Collection Phase

5. Implementation And Results

1. Collection of tweets To collect Twitter data in forms of tweet JSON, we need to register as the Twitter developer on https://dev.twitter.com/ which Twitter used to track your records and activity.

After successful registration as developer, you need to create app which will give you Consumer Key (API Key) and Consumer Secret (API Secret) which will be used as access tokens while collecting data in python, java code. we are using Twitter streaming API to collect the continuous stream of data.

Tweet Id Str	Tweet	Comment
1	well the riots have already begun on my campus they're all chanting fuck Donald Trump	
2	If Saharanpur was in Bihar not UP, had Dalits been victims in Bihar our Indian media would have Surely held the state govt to account.	
3	Terrorist Thakurs attacked Dalits today again in Saharanpur. Total failure of BJP,Government in UP. Implement president rule in UP.	
4	US nuclear submarine traveling through the Panama Canal 10 minutes ago. Are things escalating with North Korea?ï£¡	Account Suspended

Table 1. My Caption

2. Data Collection and Filtration Data collection of negative or sensitive words as per research work need. This phase totally depends on what you want to collect it may any international topics or local active topic which may in trends or sensational news. If you want to collect tweets having negative words like riots, strike, bomb etc. Explicitly it can be mentioned at filter point while collecting tweets JSON. Some time false positive tweets comes in JSON by using any filtration tools remove the false positive tweet and keep only required set of tweets.

3. Processing To process collected tweets in JSON file we need to import in the database, For our research work, we are using MongoDB to save JSON in multiple collections and doing the analysis of tweets. So, First criteria of filtration are sensitive or negative set of keywords should be in the tweet which we did while the collection of tweets. Spread directly depend on the number of Tweet's retweets so first, we are finding the number of retweets of any tweet by checking "retweets status.retweet count" as per sensitivity of topic. If the topic is very sensitive even it's 10 retweets may create the problem. which we will filter and get those tweet's having 10 retweet-count and will give high priority to it. Then, we are getting the information of the user who is tweeted or retweeted and how influential by checking his follower's count. we are taking the intersection of tweeted owner followers and retweeted user follower id's count and frequency of retweeted tweets by considering retweeted status.created-at time and created at time of tweets.

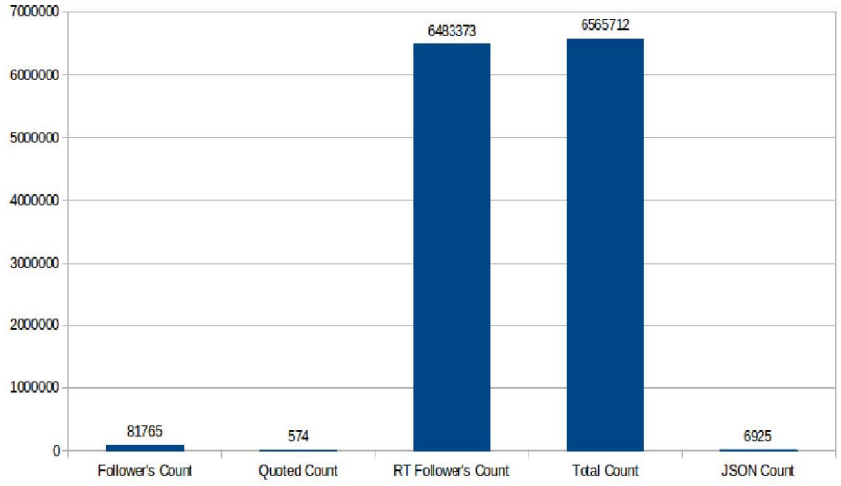

Figure 4. Tweet-1 Details

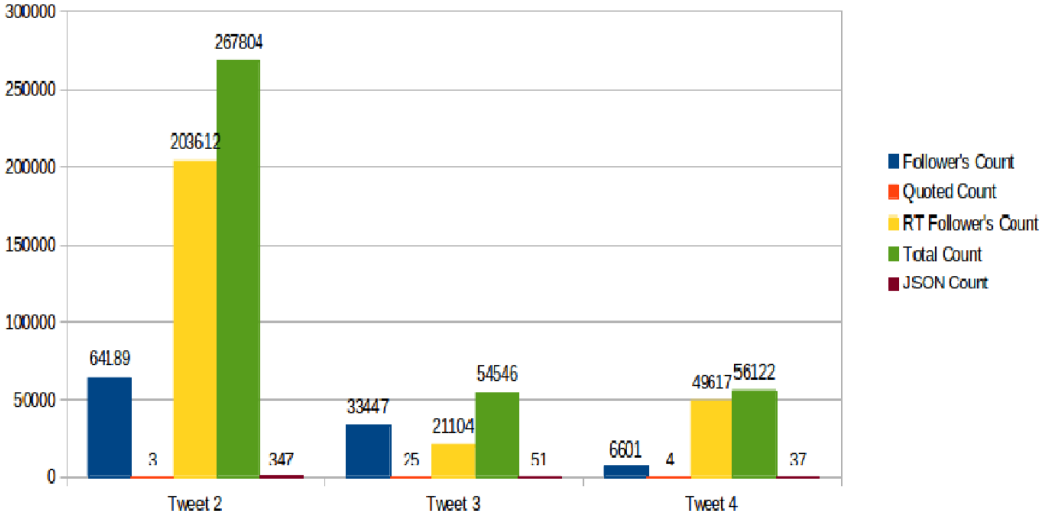

Figure 5. Tweet-2,3,4 Details

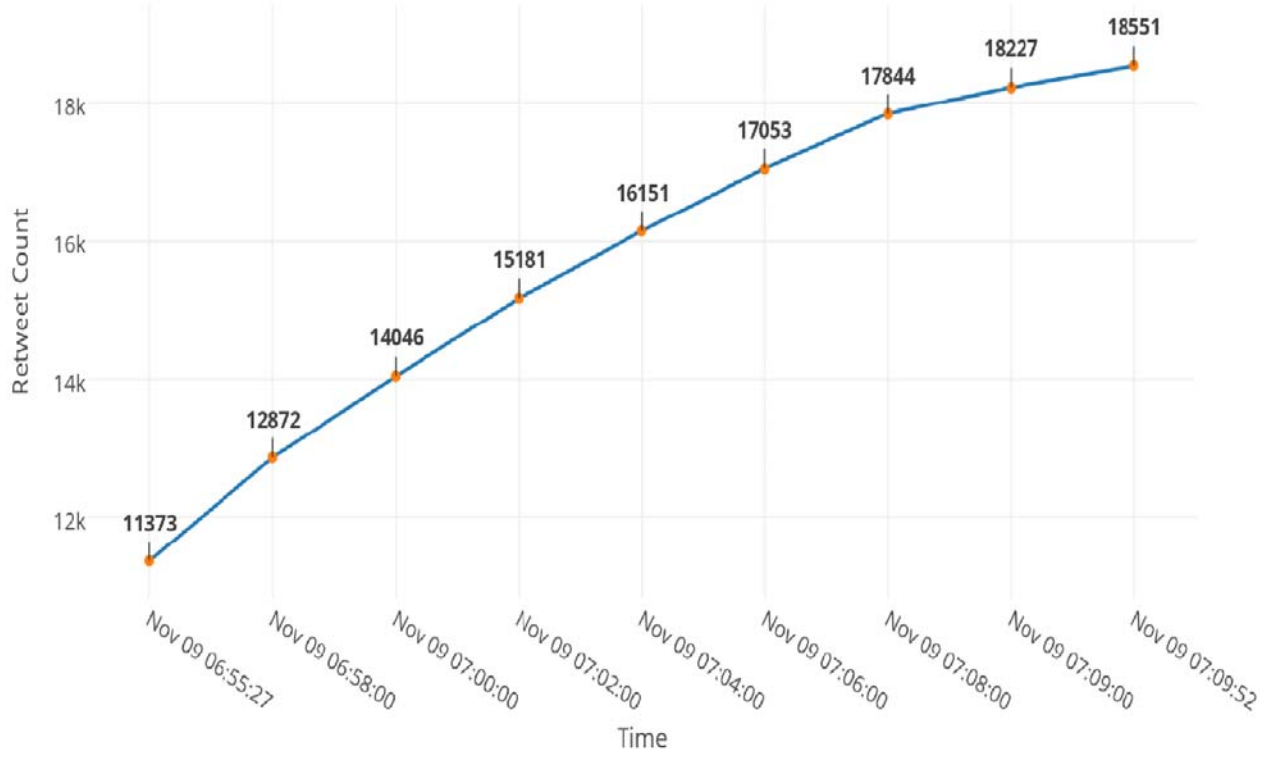

Figure 6. Tweet-1 Graph

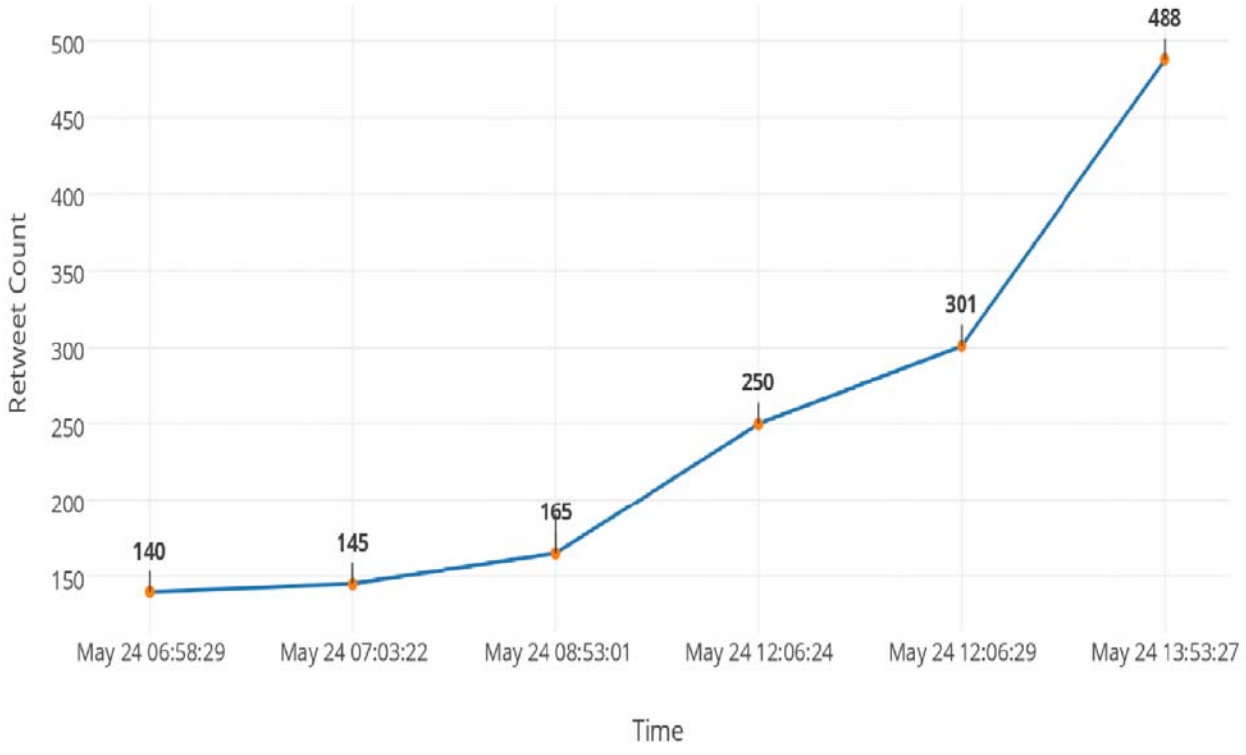

Figure 7. Tweet-2 Graph

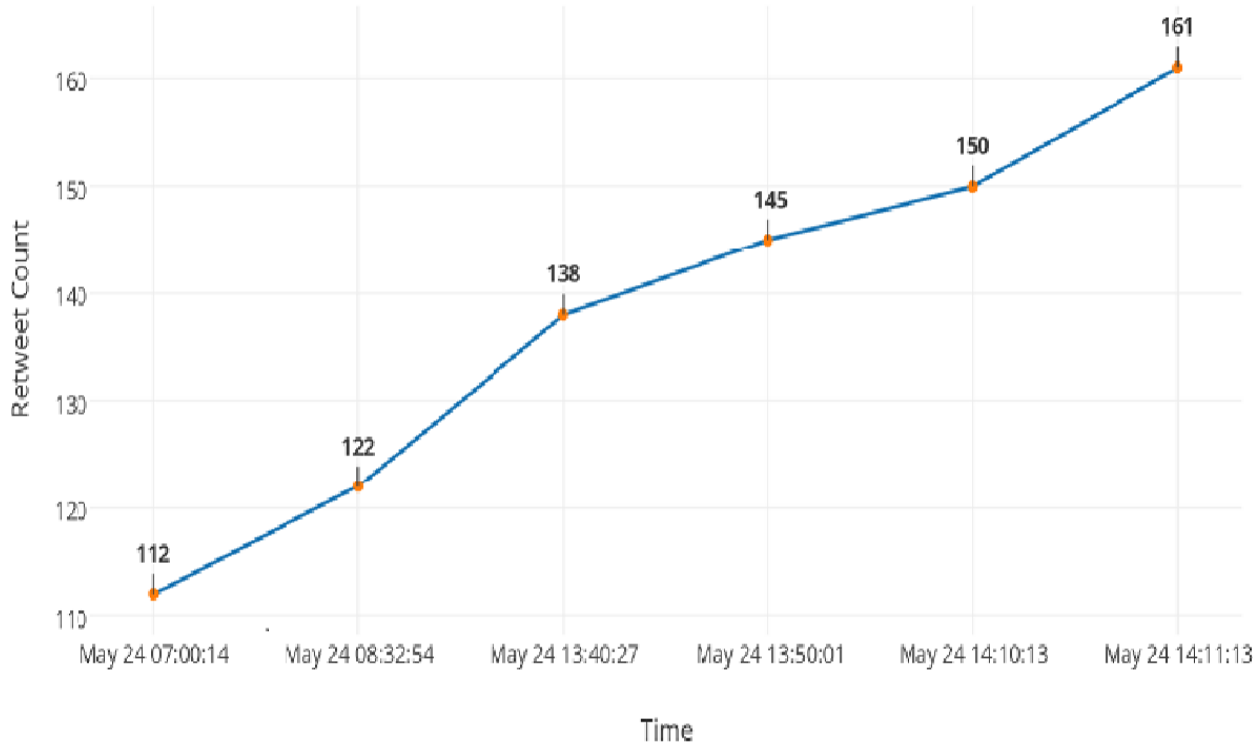

Figure 8. Tweet-3 Graph

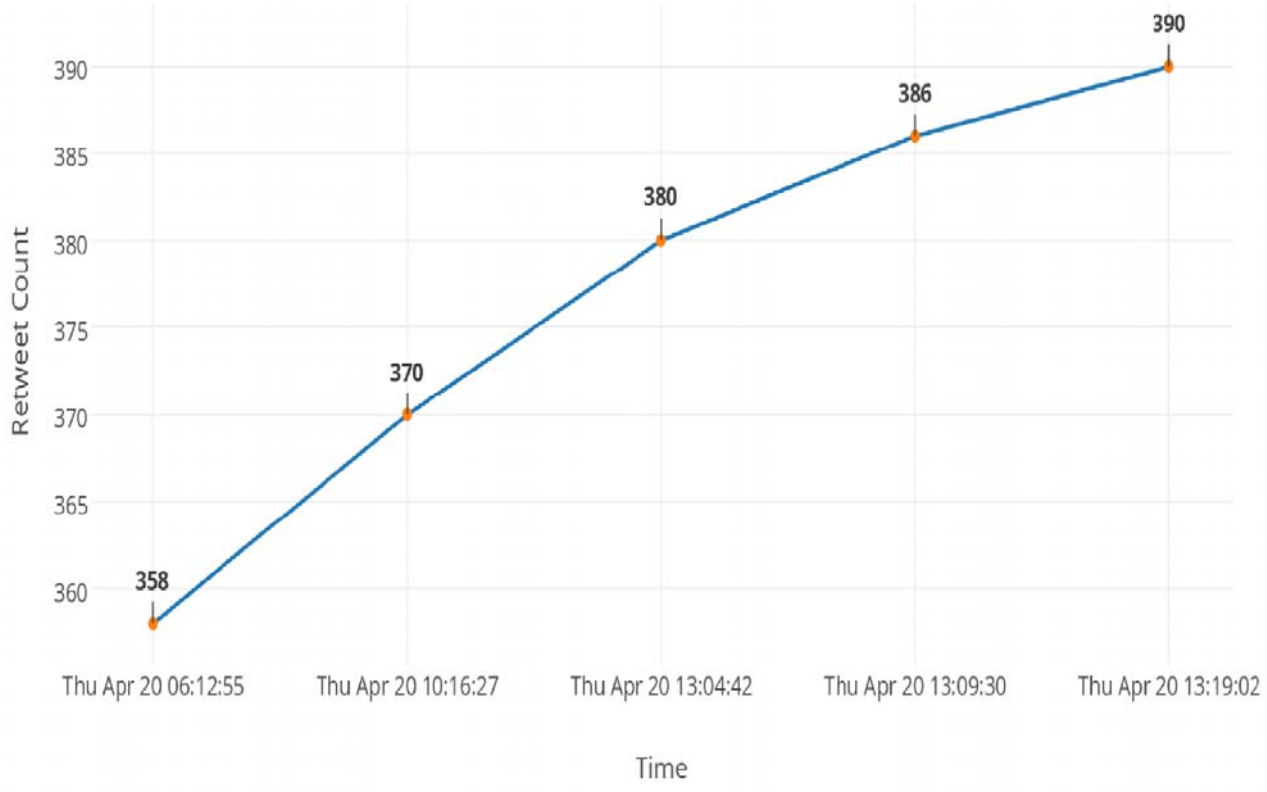

Figure 9. Tweet-4

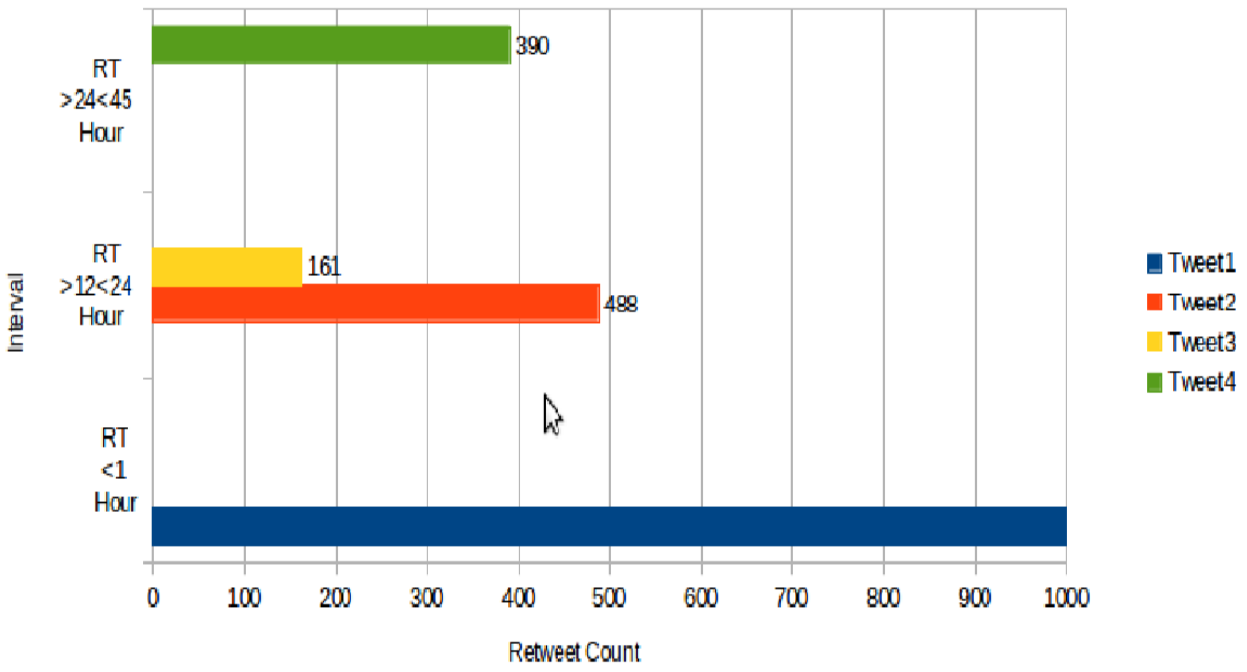

Figure 10. Tweet's Time Based Graph

6. Conclusion

In this work, we looked into the problem of identifying the spread of any negative or sensitive tweet in social space. We proposed metric to find the spread count of identified negative tweet and their retweet count with time. More retweet of tweets and having more quoted and replies directly reflect more number of the target user. We considered own followers, replies, quoted tweet and retweet follower's count.

In future, our work will be focused toward impact metric to find an impact of any sensitive and negative tweets.

References

[1] Stattner, Erick., Eugenie, Reynald., Collard, Martine. (2015). How do we Spread on Twitter? Research Challenges. *In:* Information Science (RCIS), IEEE 9th International Conference.

[2] Rao, Praveen., Katib, Anas., Kamhoua, Charles. (2016). Probabilistic Inference on Twitter Data to Discover Suspicious Users and Malicious Content, *In:* Computer and Information Technology (CIT), IEEE International Conference.

[3] Zamparas, Velissarios., Kanavos, Andreas., Makris, Christos. (2015). Real Time Analytics for Measuring User Influence on Twitter, Tools with Artificial Intelligence (ICTAI), *In:* IEEE 27th International Conference on AI.

[4] Dayani, Raveena., Chhabra, Nikita., Kadian, Taruna., Kaushal, Rishabh. (2015). Rumor detection in twitter: An analysis in retrospect, Robotics and Applications (ISRA), *In:* IEEE International Conference on Advanced Networks and Telecommuncations Systems(ANTS).

[5] Tao, Ke., Hauff, Claudia., Houben, Geert-Jan., Abel, Fabian., Wachsmuth, Guido. (2014). Facilitating Twitter data analytics: Platform, language and functionality, *In:* IEEE International Conference on Big Data.

[6] Thom, Dennis., Kruger, Robert., Ertl, Thomas., Bechstedt, Ulrike., Platz, Axel., Zisgen, Julia., Volland, Bernd. (2015). Can Twitter Really Save Your Life? A Case Study of Visual Social Media Analytics for Situation Awareness, *In:* IEEE Pacific Visualization Symposium (PacificVis).

[7] Balasuriya, Lakshika., Wijeratne, Sanjaya., Doran, Derek. (2016). Finding street gang members on Twitter, Advances in Social Networks Analysis and Mining (ASONAM), *In:*2016 IEEE/ACM International Conference on Social Network Analysis.

[8] Rohan., Perera, D.W., Anand, S., Subbalakshmi, K. P. (2010). Twitter analytics: Architecture, Tools and Analysis, *In:* Robotics and Applications (ISRA), MILITARY COMMUNICATIONS CONFERENCE, 2010 - MILCOM,2010.

[9] https://dev.twitter.com/docs

[10] http://tweepy.readthedocs.io/en/v3.5.0/

[11] https://pythonprogramming.net/twitter-api-streaming-tweets-pythontutorial/

[12] https://en.wikipedia.org/wiki/Twitter