Use of Twitter Stream Data for Trend Detection of Various Social Media Sites in Real Time

Sapumal Ahangama

MillenniumIT, Sri Lanka
sahangama@gmail.com

Abstract. Emergence of social networks such as Twitter has enhanced communication among large proportions of participants sharing enormous volumes of data. Categorization and analysis of the data in-depth will enable to generate valuable insights and information. In this paper, a new method is presented to find the trending content and trending topics of various social media networks using real time data shared on Twitter. The insights on current trending content generated by the proposed system will be of high importance as majority of the external social media networks doesn't directly publish any real time data related to trends or most interesting content within itself.

Keywords: Trend Detection, Social Media, Social Computing.

1 Introduction

Social networks are a framework used worldwide to share personal views, ideas, debate on various topics and exchange personal experiences. The use and penetration of popular social networks in societies have grown rapidly during the past few years.

Among popular social media networks, Twitter can be considered as a great reservoir of information with nearly 200million Twitter users worldwide. The users engage with nearly 400million micro blogging posts ('Tweets') published daily [1]. Within such information, trends or popular topics are driven by emerging events, breaking news and general topics that attract attention of many Twitter users. Since the data originates mainly from human users scattered worldwide and as the tweets would display the direct personal opinion on a topic, the data can be of immense value in decision making and generating foresights. Understanding and identifying the important topics in these social networks will give insights on how the society gives value to various issues. In addition, structured analysis could lead to arriving at definitions on how people in various segments would respond to an issue in the social media domain. Such an analysis could lead to understand how a community would come into a conclusion or a decision based upon a news outbreak. Also, such intelligence would be important in commercial as well as in political aspects. Having an understanding on the reaction and impact of a political decision taken by a main stream political party would enable to improve the decisions such that the positive sentiment among the public can be increased. Similarly, commercial enterprises would be able

G. Meiselwitz (Ed.): SCSM 2014, LNCS 8531, pp. 151–159, 2014.

to assess the success of a commercial campaign by determining the popularity strength of the keywords the campaign created. Availability of a proper model would enable fine tuning of such a commercial campaign to attain higher popularity with a greater positive sentiment.

With the increase in popularity of social media networks, various new social media sites emerge with sharing limited to a specific type of content such as social media networks specializing on sharing solely the images, videos or location based information. Since these social media networks specialize in a specific type of content being shared, understanding the patterns of the data will add more value. For example in the most simplest form, identifying the trends in a video sharing social media network in real time will enable to determine the videos creating the highest social impact at the time indicating the success of such videos. Further, the identification of user behavior patterns would enable to target and produce the videos in the specific user interest areas.

Various attempts have been previously made to derive intelligence, conclusions and behavior models using Twitter real time data in different fields with satisfactory results. In this paper, we look in to an indirect method, where the trend analysis in external social media networks will be carried out using the twitter real time stream data. Since many of the current social media networks do not provide at least the basic real time content analytics data such as trending or most talked about content, it is intended to bridge this gap with the indirect method proposed.

In this paper, Section 2 describes the related work where Twitter data has been purely used to generate various conclusions. Section 3 describes the content categories in Twitter as well as the methodology used to derive trends of external social media sites. An analysis of the data and results are done in Section 4 and Section 5 deals with the concluding remarks with an explanation of limitations of the system.

2 Related Work

Basic location based trend identification on Twitter would provide an insight into the most talked about topics in a specific geo location. Further in-depth analysis has proved that Twitter data can be used in a variety of other fields as well. Various studies on Twitter data has been carried out beyond mere identification of worldwide trends or location based trends of overall Twitter data. Becker et al. [2] demonstrated on improving such generic real time analysis of 'trending topics' to identify real world events using Twitter data.

Further, studies have successfully displayed how Twitter data can be used to generate foresights in a specific domain such as entertainment, medical, disaster recovery etc. Sitaram et al. [3] displayed how user chatter on Twitter could be used in a real world scenario. It was shown how the chatter can be used to generate forecasts of box office performances of movies. Sakaki et al. [4] demonstrated how real time algorithmic analysis of Twitter data can be used to generate location estimations of earthquakes in Japan. Vieweg et al. [5] analyzed the Twitter data published during natural disasters and proposes a framework such that the important information could be

extracted where the emergency responders could use it. In the medical field, Achrekar et al. [6] have gone to the level of predicting flu trends by analyzing real time Tweets which can be used to identify flu outbreaks and contain them.

The above approaches are of great use as the information generated directly from people can be obtained in real time which is not possible in traditional methods of data collection via manual or online surveys. In addition, no previous attempts were found where Twitter data was used to derive content analytics of external social media networks which is the intention of this paper.

3 Methodology

A variety of studies have been carried out on the content categories shared on the micro-blogging platform. Naaman et al. [7] analyzed Tweets originating from a random set of users other than re-Tweets. In this study, data was collected from the Twitter API and categorized into various segments based on the content of the Tweet. In the categorization, a vast majority of the Tweets fell in categories such as the current state of the user and self-promotion of the user (Categories Current State and Self-Promotion in Figure 1). Content shared on external social media sites such as for image or video sharing consist mainly of content that is about the user's current state or content with intention of self-promotion. Hence there is a high content category correlation among overall Twitter content and the content on external social media sites.

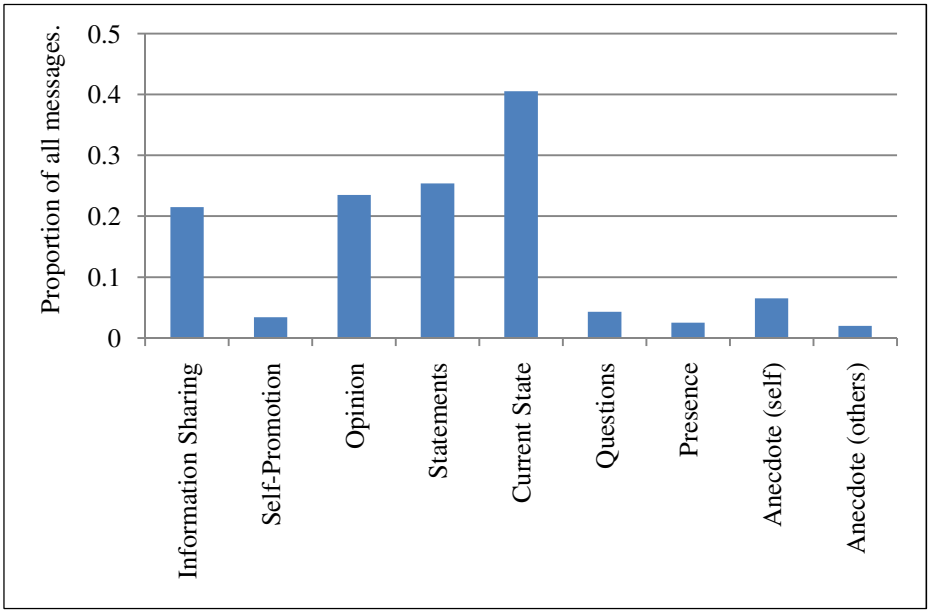

Fig. 1. Twitter Content Category Frequency Source Naaman et al. (2010)

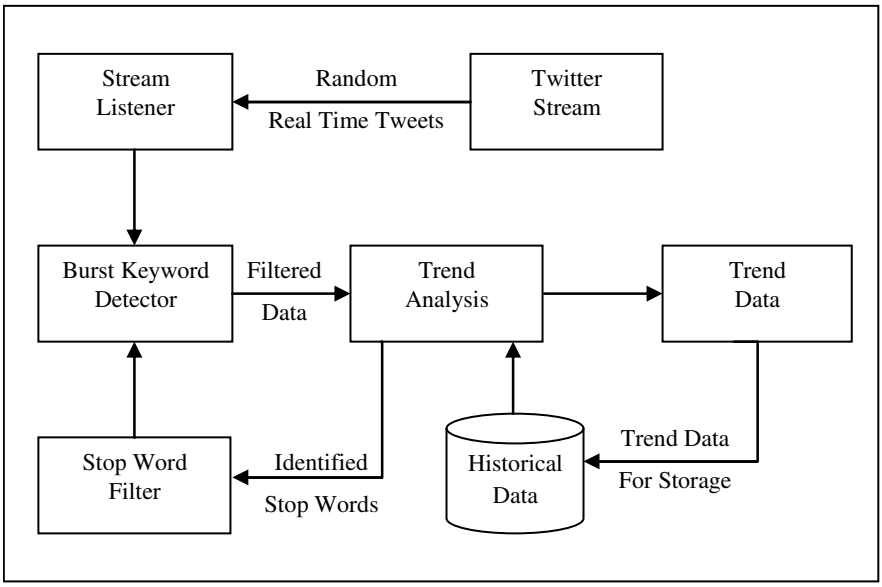

Fig. 2. System architecture to collect and analyze data

Majority of the external social media sites are provided with the feature to share the content in Twitter alongside the original post. The Tweet would provide a link to the original post with a short description (e.g. When an image is posted on Instagram, a summary and URL of this post can be shared on twitter automatically). It could be assumed that a random sample of the content posted on these external social media networks get shared on Twitter automatically. Thus, due to the properties of random distributions, the content samples shared on Twitter and the original social media network will have the same equivalent statistical properties as the sample size is large.

Further, based on the interest, these tweets get re-shared (re-Tweeted) on Twitter itself. The amount at which a specific image is re-Tweeted will provide a measure of the interest users, show to the respective content. Such high interest content can be taken as trending content at a time frame.

For the testing of the concept, Instagram images shared on Twitter were considered. Twitter data was obtained by subscribing to the real time public data streams made available by Twitter. The data obtained from the Twitter stream is guaranteed to be a random sample by Twitter, hence the statistical properties will be the same as the complete data set on Twitter [8]. Each Instagram image shared on Twitter takes a specific format. For example, in case of Instagram an image URL is with the format "http://instagram.com/p/xxxxxxx/" where "xxxxxxx" is the unique image id. This URL format as appearing on Tweets was used to filter data when the data subscription with the Twitter data stream was made. Trend estimation was limited to hash tags rather than all keywords such that the analysis would be simpler.

Figure 2 shows the system architecture used to collect the data from Twitter and analyze trends. The stream listener will continuously collect Tweets on the subscribed topic from the public Twitter stream. The listener further filters for Tweets

with Instagram images. Each of the hash tags reported of the collected tweets is then analyzed for a burst within the time frame. The burst keyword detector uses a stop word filter. The stop word filter will filter any common hash tags that report large data counts, but the count is within the average count reported in a time frame. The filtered data is then analyzed to detect any trends. For trend analysis a database stores the historical data such that the system can maintain a state.

4 Analysis of Data

Table 1 summarizes the data collected from Twitter. Since the sample obtained from Twitter is guaranteed to be random [8] and can be considered as a large sample, it can be assumed that the analyzed data set represents the original data set in statistical properties.

Table 1. Summary and analysis of data collected

Average number of images posted on Instagram daily	55,000,000 [9]
Average rate of Tweets collected (per second)	50
Average daily Tweets processed	4,320,000
Average number of Instagram images processed from Tweets (daily)	3,850,000
Average unique number of Instagram images processed from Tweets (daily)	3,210,000

Table 2 shows the distribution and breakdown of all collected hash tags during a 1 hour time window. Hash tags collected had a positively skewed distribution with vast majority of the hash tags reporting a very few occurrences within the time frame while a few displaying relatively larger numbers of occurrences. These hash tags were filtered to identify any tags that report in bursts which may turn out as trends.

Upon filtering for stop words and trend detection, many of the hash tags identified as trending, coincided with top news stories and events around the world that are currently taking place. Sports events, award ceremonies, deaths of popular people and protests are few examples. Hash tags of these events became trends with people sharing images related to the event in bursts during the event. Figure 3 shows the general life time pattern of a hash tag trend over time. In addition the figure shows an example hash tag trend where counts are plotted every ten minutes. In the graph pattern, the number of occurrences in a time interval and the time taken for the trend to fade away

is a direct measure of the strength of the trend. Highly talked about and highly interesting tags tend to show higher bursts. Stories which tend to develop further with new sub events showed to last longer in the time axis.

Table 2. Number of occurrences of a hash tag within a 1 hour time frame

Number of occurrences within time frame	Number of hash tags
1	51,443
2	8,352
3 – 10	7,766
11 – 20	1,101
21 – 1000	948
> 1000	7
Total	**69,617**

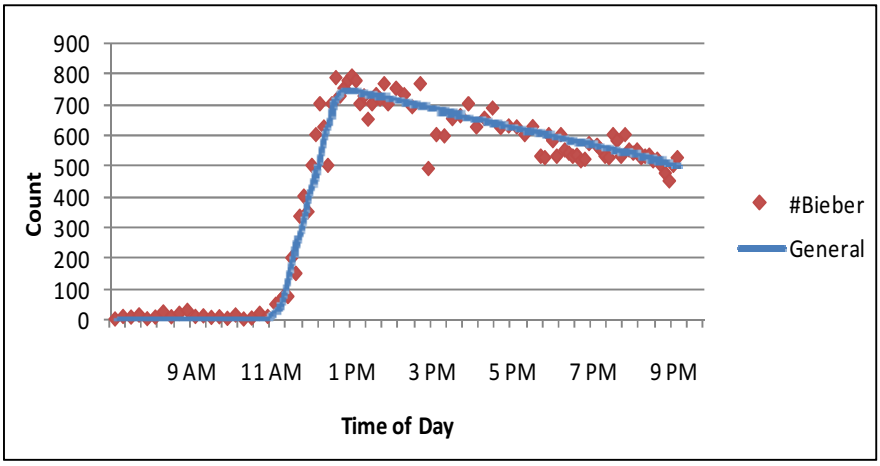

Fig. 3. Generalized life time pattern of a trending hash tag with an example (#Bieber)

An example set of such tags trending related to an event identified by the system are given in Table 3.

Table 3. Example trending events and popular event hash tags

Hash Tags	Event
#Egypt, #TharirSquare, #ByeByeMorsi	During 2013 Egypt uprising.
#NelsonMandela, #Madiba	During the mourning period of Nelson Mandela.
#4thofjuly, #America, #July4, #independencedayusa	During July 4th Independence celebrations in USA.
#princesskate, #royalbaby, #boyorgirl	During the period where Royal Baby was expected in UK.
#wimbledon, #murray, #AndyMurray, #tennis, #champion,#77years	During 2013 Wimbledon Championship.
#runningofthebulls	During Running of the Bulls event in Spain.
#UnitedforMarriage,#Equality, #EqualityForAll, #prop8	During 2013 Equal Marriage Supreme Court hearing in Washington, USA

During the analysis of Instagram images collected, nearly 5% of the images collected during a day accounted to reporting multiple times as a result of re-Tweets or re-sharing of the same image. Table 4 demonstrates the distribution of images reported more than once in the time period. The images which recorded highest number of re-Tweets were shared by popular personalities on Instagram such as celebrities and sports stars. These users had large number of followers as well as the images received large number of likes. As a result of their popularity, the images recorded a large number of re-Tweets resulting in certain hash tags of the images identified as trending.

Table 4. Repeat count of images within a time frame of 1 day

Repeat Count	No of Images
2	121,522
3 – 10	42,930
11 – 20	3460
21 – 100	2956
101 – 1000	588
> 1000	32
Total	171,488

5 Conclusion

In the analysis of results for Instagram, it was found that trends directly coincided with major news and other popular events taking place in the time frame. Reasons for image trends fall into 2 main categories. Firstly, major social events such as protests, civil unrests, and festivals create image trends with many people sharing images related to the event. The second category is when a popular person posts a controversial or a unique image, tags of the image became a trend with many re-sharing the same image in quick succession due to the high interest. It is assumed that a random sample of the content originally shared on the external social media site is shared on Twitter. It could be said that Twitter stream can be successfully used to generate insights on Twitter connected external social media sites if the assumption holds true. This assumption is a limitation of the system when expanding to monitor trends of external social media networks which are not highly integrated with Twitter. In addition, due to text content length limitations in Twitter, certain parts of the original post may get truncated leading to incomplete data. Also the method cannot be verified accurately as various social media sites use proprietary algorithms and custom variables to calculate their own trends and most information is not published openly.

References

1. The Official Twitter Blog, Celebrating #Twitter7, https://blog.twitter.com/2013/celebrating-twitter7
2. Hila, B., Naaman, M., Gravano, L.: Beyond Trending Topics: Real-World Event Identification on Twitter. In: ICWSM (2011)

3. Sitaram, A., Huberman, B.A.: Predicting the future with social media. In: 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), vol. 1, pp. 492–499. IEEE (2010)
4. Takeshi, S., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: Proceedings of the 19th International Conference on World Wide Web, pp. 851–860. ACM (2010)
5. Sarah, V., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1079–1088. ACM (2010)
6. Harshavardhan, A., Gandhe, A., Lazarus, R., Yu, S., Liu, B.: Predicting flu trends using twitter data. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 702–707. IEEE (2011)
7. Naaman, M., Boase, J., Lai, C.: Is it really about me?: message content in social awareness streams. In: Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work, Savannah, Georgia, USA, February 6-10 (2010)
8. Twitter Developers, Frequently Asked Questions, How are rate limits determined on the Streaming API?, `https://dev.twitter.com/docs/faq#6861`
9. Instgram Press Page, `http://instagram.com/press/`