Uncovering Trending Stories from Twitter by Extracting Ground Truth from Datasets

Mr. Vishal Dilip Shinde
Department of Computer Engineering
DGOICOE, Bhigwan, Pune
Savitribai Phule Pune University,Pune
vishalshinde.it@gmail.com

Prof. Tanaji A. Dhaigude
Assistant Professor, Department of Computer Engineering
DGOICOE, Bhigwan, Pune
Savitribai Phule Pune University,Pune
tanajidhaigude@gmail.com

***Abstract*—**Social networking services like Twitter generates contents that reflects series of conversation which shows real-world events. Twitter is social networking site that provide service for a large number of users to communicate with each other simultaneously; it is an asymmetrical relationship between friends and followers that provides an interesting structure among the users of Twitter. Twitter's series of messages called tweets, which are restricted to 140 characters and thus are usually much focused. The basic process is to capture tweets from twitter that extract mostly discussed topic in between users. This tweet dataset can be process for finding trending stories using standard natural language processing. An Uncovering trending stories is therefore a building block is to extract and summarizes the information raised from social networking services. There is verity of methods for finding trending stories that improves quality of result. This paper proposes application for uncovering trending topics from twitter datasets using BNgram topic detection method.

Keywords- *Twitter, Topic Detection, BNgram, Social Networks*

*****

I. INTRODUCTION

The Social Networking has seen widely used in recent years. As social networking services spread rapidly in all age society. The large facts discussion, user interaction and communication happen on social networking sites and this reflects real world events and trending topics; now a day's user is more active for posting message about real-world events on social networking sites, so social networking site now becomes accurate area for drill down for real time information about events, where we can effective detection for real-time stories.

The social networking service has becomes main area, as number of interconnected user increasing rapidly and there rapid and effective conversation about real-time events. Twitter is one of the social networking website that becomes most popular in every one to share his thoughts, ideas and express opinion on real-time events with unique user account. Twitter has number of users, who post their messages called a tweet that contains a maximum of 140 characters. It is estimated that there are 6 to 7 million users who use twitter a total of 134 million times a month [2]. Twitter is openly access for posting information related to breaking news and ongoing events, near about 500 million users and more than 400 million short messages known as tweets [13], These tweets contains conversation, thoughts, impacts of real-time events.

Trending stories uncovered from dataset, which contains series of tweets with respect to time stamp. Uncovering trending stories is most helpful for breaking stories broadcaster, social issues analyzer. [13]And it is very useful for Cyber crime agencies in tracking issues related to certain events.

In this research, we are analyzing different methods for extracting trending topics or stories from Twitter datasets.

The reminder of this paper is as follows. Next, we provide background on the use of Twitter. Then we introduce types of trends, datasets, related work and detailed implementation and methods for uncovering trending stories from twitter.

II. TWITTER

In this section overview of Twitter, details of trends in twitter and which we are using input to methods for uncovering ground truth from trending stories.

- Twitter:

  Twitter is most popular social website, where huge numbers of users share their thought, opinion in the form of short messages called tweets. The success of twitter is due to two reasons first, shortness of tweets, which cannot exceeds 140 characters that create and share minimum period of time .and second is spreading those messages to a large number of user in very little time. The twitter has established syntax for interaction with one another, which syntax adopted by developers .Most major Twitter clients have implemented this as well. The standard in the interaction syntax include [13]:

  - *User mentions*: when a user mentions another user in their tweet, an at-sign is placed before the corresponding username.
  - *Replies*: when a user wants to direct to another user, or reply to an earlier tweet, they place the *@username* mention at the beginning the tweet.
  - *Retweet*: a retweet is a re-share of a tweet posted by another user. Retweets, the new tweet copies the original one in it, then the retweet attaches a *RT* and the @username of the. user who posted the original tweet at the beginning of the retweet.
  - *Hash tags*: it is same as tagging facilities on other social networking service, hash tags included in a tweet to mention other user.

- Trending Stories:

  One of the main feature on the homepage of twitter shows a list of top terms so-called trending

ISSN: 2321-8169
3942 - 3946

topics at all times. These terms reflect that are being discussed most. Twitter focuses on topics that are being discussed much more than usual. Trending topics have attracted big interest not for only user mainly for other information consumers such as journalists, real-time application and social media researcher [13].However, no further evidence is know about the algorithm that extracts trending topics[13].

III. Types of Trending Toics

Next, we treat trending topics in following categories [13]:

- News: On many occasions news break on Twitter before any news agency. We define that a trending topic can be categorized as news when that gives present information.
- Ongoing events: The trending topic is in ongoing event when information is posted by community of users tweeting about an ongoing event.
- Memes: Also trending topic is in memes which is posted by either individual or community with viral ideas. It can be from a funny message that attract user to repost.
- Commemoratives: Last type of trending topic which produced by individual for congratulating celebrity their birthday or anniversary r any memorable day such as Independence Day, Republic Day.

IV. Litreture Survey

This section describes about various technique in uncovering trending topics. In Sensing Trending Topic in twitter [1], Three Twitter datasets are used to extract trending topic detection and it is extracted by BNgram Method. In Emerging Topic detection on Twitter based on social terms evaluation [2], recognize the primary role of twitter and they propose a novel topic detection technique that permits to retrieve in real-time the most emergent topic expressed by the communities of users, They define a directed graph of active authors based on their authority by relying on the well-known page algorithm [2]. In another work TwitterMonitor: Trend Detection over the Twitter Stream [3], they represent TwitterMonitor, a system that provides meaningful analytics that synthesize an accurate description of each topic using Twitter API, Another work using Twitter API is TwitterStand: News in Tweets[4] to build a news processing system. In Detecting and tracking political Abuse in Social media [5], describe a machine learning framework that combines topological, content-based and crowd sourced features using Twitter API.

In Predicting political preference of Twitter users [6], they can predict from their interaction with political parties by building prediction model based on a verity of contextual and behavioral feature training the models by restoring to a distance supervision approach. In another work beyond trending topics: Real-world event identification on twitter [7], explores approach the stream of twitter message to distinguish between message about real world events and non-event messages sing cluster-level event features based on Temporal, social, topical, twitter-centric. In next approach, Taking Topic Detection from Evaluation to Practice [8], avoids generating garbage clusters, they had revert to different approach. In mining Newsworthy Topics from Social media [9], demonstrate by analyzing tweets corresponding to events drawn from the word of politics and sport using BNgram method. Also In breaking news detection and Tracking in twitter [10], propose a method to collect, group, rank and track breaking news in twitter, each group is ranked based on popularity and reliability factors.

In recent work, Real-Time Classification of Twitter Trends [13] uses Twitter API for first obtaining top ten trending topic and second obtaining trending topic with text, timestamp, user, and language for each of the underlying tweets.

V. Problem Satement

Task of uncovering trending stories in real time from social media stream, the stream is pieces of text generated by social media users like post. The Goals and Objectives of our proposed work is first uncovering Trending Stories in each time of slot in Day and extracting the Ground truth from Twitter Datasets.

VI. Proposed System

A. *System Model*

Fig. 1. Describes system architecture of Uncovering Trending Topics from Twitter.

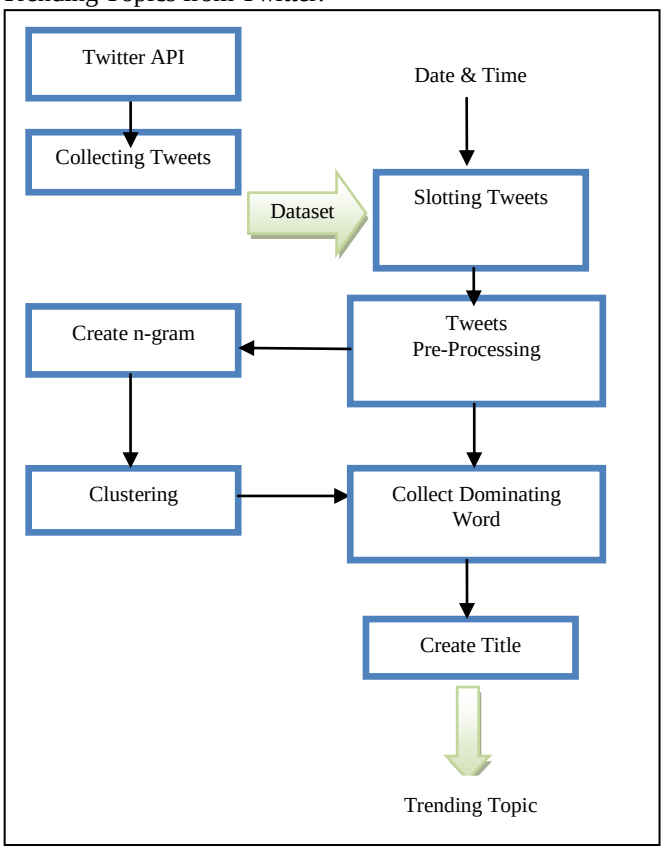


Fig. 1. System Architecture

Here we use twitter dataset, that contains series of tweets, among that we select series of tweets from two instances called Slots, and then on selected tweets we apply text preprocessing operation through natural language processing for removing stop words, non words and stemming. We follow the BNgram Algorithm, to creating n-grams then arranging these n-grams into different clusters using k-means clustering algorithm.

Finally we extract the keyword that having high frequency by using this keyword we create title for trending story.

B. Algorithm

K-Means Algorithm

K -means clustering is a partitioning method. In this terms are classified as one of K-groups. The result of this partitioning method is a set of K clusters. In K- Means Clustering algorithm terms are classified into K clusters; the value of K is user defined. Firstly centroid of each cluster is selected and then according the centroid, the terms having minimum distance from the given cluster, is assigned to that particular cluster. Euclidean Distance is used for calculating the distance of terms from the particular centroid.

C. Mathematical Model

- Input Sets:

Ds={Dsi, i < 0 < n} -Set of Twitter Datasets
Where, n=Number of Twitter Datasets
St={Stj , j < 0 < n} -Set of Text Streams
Where, n=Number of Text Streams

- Processing Sets:

Tw={Twi, i < 0 < n} -Set of Tweets
Where, n=Number of Tweets
Gr={Grj , j < 0 < n} -Set of n-grams
Where, n=Number of n-grams
Cs={Csk, k < 0 < n} -Set of Clusters
Where, n=Number of Clusters

- Output Set:

Rt={Rti, i < 0 < n} -Set of Ranked tweets
Where, n=Number of Ranked tweets
Ur={Urj , j < 0 < n} -Set of Users
Where, n=Number of Users

Let S be the systems we can mathematically represent S using Set of Theory as,
S={Ds, St, Tw, Gr, Cs, Rt, Ur}

D. BNgram

Term frequency-inverse document frequency, or *tf-idf*, has been used for indexing document since it was first introduced [9]. But here we want to find the term which appears more period of time more than others. We select terms with a high temporal document frequency-inverse document frequency [9], by comparing the most recent *m* messages with the previous *m* messages and count how many terms are repeated. We assume most recent *m* messages as one slot. After standard text preprocessing we index all terms from these messages. For each term, we calculate the document frequency for a set of messages using df_{ti} defined as the number of messages in a set of *i* that contain the term *t*.

$$df\text{-}idf_{ti}=(df_{ti}+1).(1/\log (df_{(ti-1)}+1)+1) \quad (1)$$

This produces a list of terms which can be ranked by their $df\text{-}idf_t$ score. To maintain some word order information, we define terms as a *n*-gram, i.e. sequence of *n*-words. Then we arrange this tweets in groups called cluster. Each clusters defines a topic as a list of *n*-grams, we call this process of finding bursty n-grams "BNgrams" [9].

E. Topic Ranking

To maximize usability of result rank topic from very large number of topics. We therefore want to rank the results by relevance. Here we use maximum *n*-gram, in this method rank topics according to the maximum $df\text{-}idf_t$ value of their constituent *n*-grams.

VII. Details of Implementation

The system for uncovering trending topics for extracting ground truth from Twitter datasets, we are implementing system like, and Collecting Datasets from Twitter API that contains series of tweets that are pre processed using natural language processing

A. Browsing datasets and Loading Tweets

From this window we are browsing and selecting dataset which is provided by Twitter API.

B. Selecting Slot and Searching

Here we select Date and time slot for find trending stories, here we have to select appropriate date because in some cases we miss the tweets that contains trending stories.

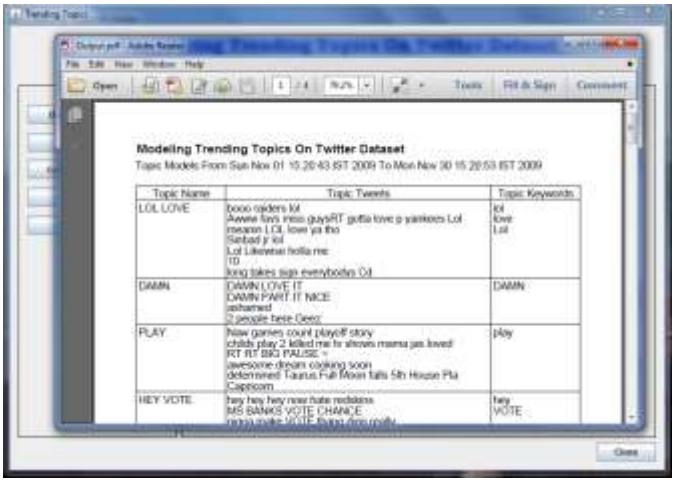
 ISSN: 2321-8169
3942 - 3946

C. Clustering and Modeling Topics

This module will genrate B-Ngram and It's Clusters

D. Showing Trending Topic

It shows the Trending Topic along with its keywords which is mentioned in tweets.



VIII. Experimental Result

Table 1.1

Slot From To	#Tweets	#Topics	#Retrieved topics	#Correctly Identified	Precision	Recall
01 to 30/06/2009	382	58	56	53	0.94	0.96
01 to 30/07/2009	33	05	04	03	0.75	0.80
01 to 30/08/2009	98	18	16	13	0.82	0.88
01 to 30/09/2009	55	09	08	07	0.87	0.97
01 to 30/11/2009	161	26	24	21	0.87	0.92

In experimental we have used a twitter Dataset that contains series of tweets; we are extracting trending stories or topics in various time slots obtained result shown in above table plotted graph with respect to Precision and recall by examining number of tweets, number of Topics, retrieved topics, and correctly identified topic.

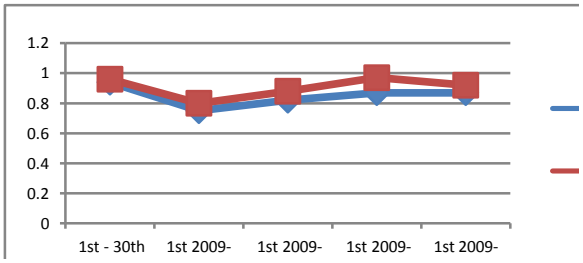


Figure 1.2

IX. Conclusion

The objective of our work is to provide a solution for uncovering trending stories using BNgram and taking overview on various methods which is emerged in recent years. This analysis shows that different technologies used in all the paper with taking different way for detecting trending topic for various purpose. Although applying this method along with preprocessing to uncovering trending stories from Twitter by extracting ground truth. And this System is showing trending story discussed in particular period of time using efficient way.

X. Future Work

In Future work, we extract trending topic and its Initiator i.e. it will identify which user has post first tweet. It helps us to find myth stories and its initiator.

Acknowledgment

I wish to express my sincere thanks and deep gratitude towards my guide Prof. Dhaigude T.A. for his guidance, valuable suggestions and constant encouragement in all phases. I am highly indebted to his help in solving my difficulties which came across whole Paper work. Finally I extend my sincere thanks to respected Head of the department Prof. Bere S.S. and all the staff members for their kind support and encouragement for this paper. Last but not the least, I wish to thank my Mother for her unconditional love and support.

References

[1] LucaMaria Aiello,Sensing Trending Topics in Twitter, IEEE Transaction on Multimedia, Vol. 15 No.6, Oct. 2013 pp. 1268-1282.

[2] Complete.com- http://siteanalytics.compete.com/twitter.com Site Profile for twitter.com Retr. July 1,2009

[3] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in Proc. MDMKDD: 10th Int. Workshop Multimedia Data Mining, New York, NY, USA, 2010, pp. 4:1–4:10, ACM.. 271–350.

[4] M. Mathioudakis and N. Koudas, "Twitter monitor: Trend detection over the Twitter stream," in Proc. SIGMOD: Int. Conf. Management of Data, New York, NY, USA, 2010, pp. 1155–1158, ACM.

[5] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, andJ. Sperling, "Twitter stand: News in tweets," in Proc. GIS: 17th ACM Int. Conf. Advances in Geographic Information Systems, New York, NY, USA, 2009, pp. 42–51.

[6] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, A. Flammini,and F. Menczer, "Detecting and tracking political abuse in socialmedia," in Proc. ICSWM: 5th Int. AAAI Conf. Weblogs and Social. Media, 2011..

[7] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in Proc.SocialCom: 3rd IEEE Int. Conf. Social Computing, Boston, MA, USA,Oct. 2011.

[8] H. Becker, M. Naaman, and L. Gravano, "Beyond trending topics:Real-world event identification on Twitter," in Proc. ICWSM: 5th Int.AAAI Conf. Weblogs and Social Media, 2011.

[9] James Allan,Stephen Harding and Devid Fisher, Taking Topic Detection From Evaluation to practce,

[10] Carlos Martin, David Corney, Ayse Goker and Andrew MacFarlane, Miniing Newsworthy Topic from Social Media.

[11] JjS. Phuvipadawat and T.Murata, "Breaking news detection and tracking in Twitter," in Proc.Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM Int. Conf., 2010, vol. 3, pp. 120–114.

[12] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to Twitter," in Proc. HLT: Annual Conf. North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 181–189.

[13] Charu C. Aggarwal and Karthik Subbain, Event Detection in Social Stream

[14] Arkatiz Zubiaga, Damiano Spaina, and Victor Fresno, Real-Time Classification of Twitter , trends in American Socity for Infrmation Science and Technology,

[15] S. Petrović, M. Osborne, and V. Lavrenko, "Streaming first story detection with application to Twitter," in Proc. HLT: Annual Conf. North American Chapter of the Association for Computational Linguistics, Stroudsburg, PA, USA, 2010, pp. 181–189.

[16] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval.New York, NY, USA: McGraw-Hill, 1986.

[17] X.Xu, N. Yuruk, Z. Feng, and T. A. J. Schweiger, "SCAN: A structural clustering algorithmfor networks," in Proc. KDD: 13th ACM Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, 2007, pp.824–833.