International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4557 - 4563

Detection of Topic Trend and Removal of Vulgar Words from User Data Streams

Roshani M. Shete
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Wardha, India
roshshete@gmail.com

Prof. S. W. Mohod
Department of Computer Engineering
Bapurao Deshmukh College of Engineering
Wardha, India
sudhir_mohod@rediffmail.com

Abstract:- Nowadays, social media is becoming very much popular. More than 170 million people are using it to being connected to the world. Trend detection is nothing but to summarize the event which the world is discussing about. This paper explains about the system of detecting current events from user stream. Here, implemented the hybrid algorithm which will extract the subset of current event. The system will tell us which or about whom the crowd is discussing. Natural language processing is used for preprocessing and filtering. And bisect K-means is used for clustering.

Keywords:- *Trend Detection, User data stream, Filtering, Clustering, Control spamming.*

*****

I. INTRODUCTION

As we all know, Communication is for exchanging and sharing the views. And for communication, social media has become powerful way. These short texting and messages reflects on real time dynamics.

People talk about their personal lives, about celebrities, politics, sports, education, and many more on such sites. So, we can give direction or can make use of these positive as well as negative comments/posts. Also people discuss topics which are not really necessary and increase the volume of data. So, here the system is designed which will extract the summery or you can say subset of events. Because we can't list each and every event and one can't access whole events

Here, designed one website for showing the results. One admin will be there who will manage the account. Whatever user post or comment, from that data steams the topic trend will be decided.

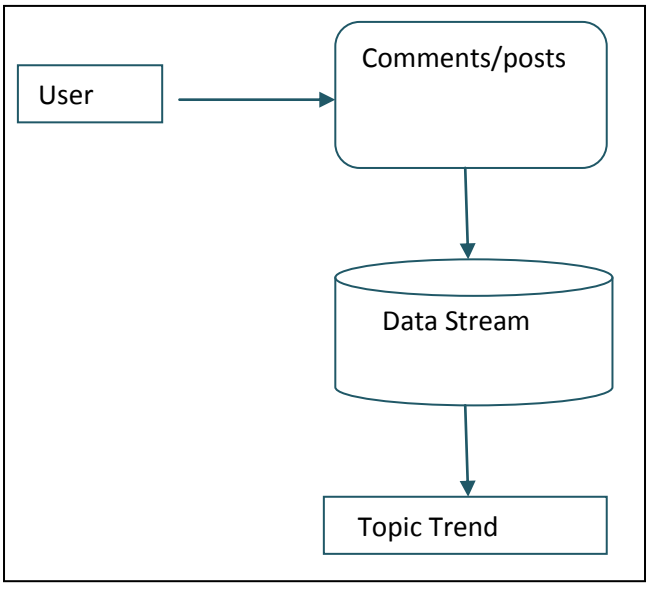

Figure 1. Overview

The hybrid algorithm is designed for extract the subset of current events. Also, nowadays the social sites are infected by spammer. As they use vulgar words which are not really necessary and also reflects on real time values. So here proposed work used the dictionary approach for removal of slang words called control spamming.

First, preprocessing and filtering is done by using the NLP. Then, for categorizing the events, Bisect K-means is used. The topic trend detection is done by using the feature extraction i.e. keyword based mining.

II. RELATED WORK

Topic Detection and Tracking extract event from public generated data on social sources and identify the trend in term of time [6]. In this the public generated data means posts uploaded by them.

Feature pivot method means the term or keyword will be considered while clustering but the drawback of that is it capture misleading term. For example if anyone wants to search like "definition of class", it will shows extra result like subclass, superclass etc. So the accuracy of result is very less, also redundancy and ambiguity gets formed in this.

M. Cataldi, C. Schifanella and L. Di Caro [7] proposed two measures, term frequency to calculate nutrition for each word and a page rank measure. After that Bursty keywords are obtained using nutrition trend. Then by using graph based approach for bursty keywords generates the topic boundary. Sayyadi, Maykov and Hurst [8] used graph approach in which clustering of keywords is done by matching pairs. They used community detection algorithm in which made a graph whose nodes are clustered. Also the topic extraction is carried out by identifying document with similar term. Lehmann, Kleinberg and Backstorm [9] have used the graph for short phrases. Phrases are connected by edges.

One of the method modeled called Latent Dirichlet Allocation (LDA) [10], the idea of knowing the most breaking news by calculating the bursty terms in document [9]. This avoids the other topics by capturing the high peak [11]. So first find bursty term then cluster them for event detection. In some graph based approach, the first step is to tag the terms, then group it and then find the interest in social media [12].

ISSN: 2321-8169
4557 - 4563

III. BLOCK DIAGRAM

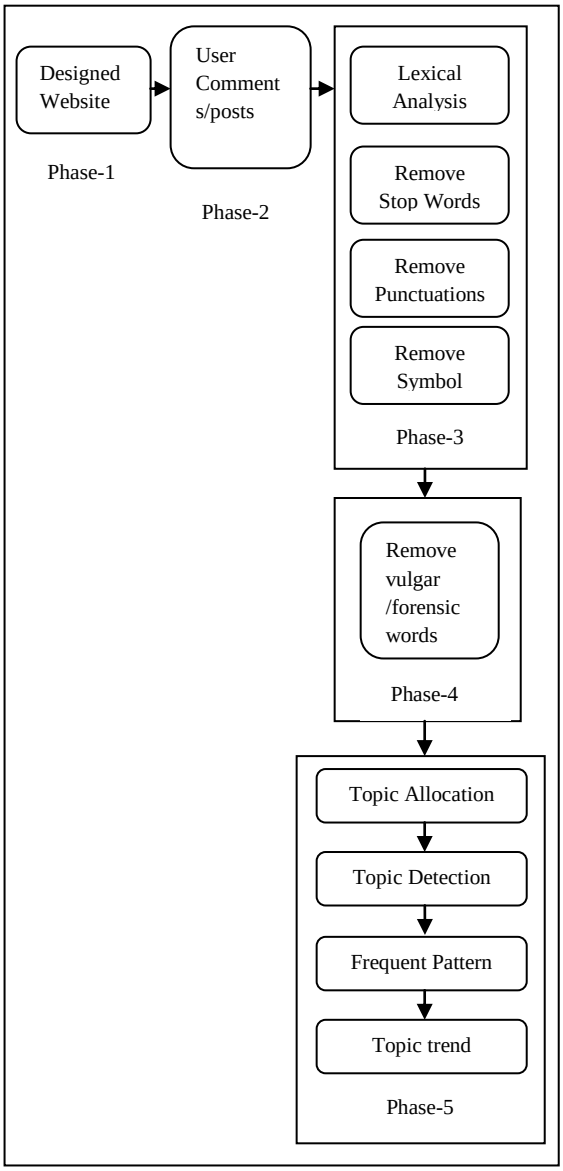

Figure 2. Work flow

In proposed work, address the task of detecting topics in real time from social media streams. To keep our approach general, consider the stream is made of pieces of text generated by users i.e. posts, messages generated by user. The flow of proposed work will be phase-1, phase-2, phase-3, phase-4 and phase-5. In phase-1, one account is created for showing results. In phase-2, whatever user will comment/post that data stream will be collected, in phase-3 filtering and preprocessing will be done using NLP, phase-4 for spam control and phase-5 for identifying topic trend.

IV. IMPLEMENTATION

In this section, the implementation of work is defined. The work flow is narrated in above section-III, now their implementation is explained below,

A. Phase-1: Designed Website

One website is designed, where account is created for showing the results. So, in that designed master page for login.

Figure 3. Master Page

There are two options in master page login and registration. For Login user need to register by filling the registration form given. After registration, user can register using registered id.

Figure 4. Registration form

This is dataset of information of users, who registered for profile,

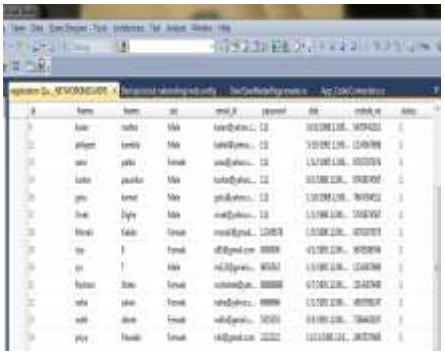

Figure 5. Registration data

After login the created account will be opened with login id,

Here, user needs to type ID and Password for login,

 ISSN: 2321-8169
4557 - 4563

Figure 6. Login with created ID

This is the created profile page,

Figure 7. Profile

B. Phase-2: User Comment/post

After having the profile, user will comment/post about real world that will be saved in database.SQL commands are used.

These are comments and posts of user and their friends,

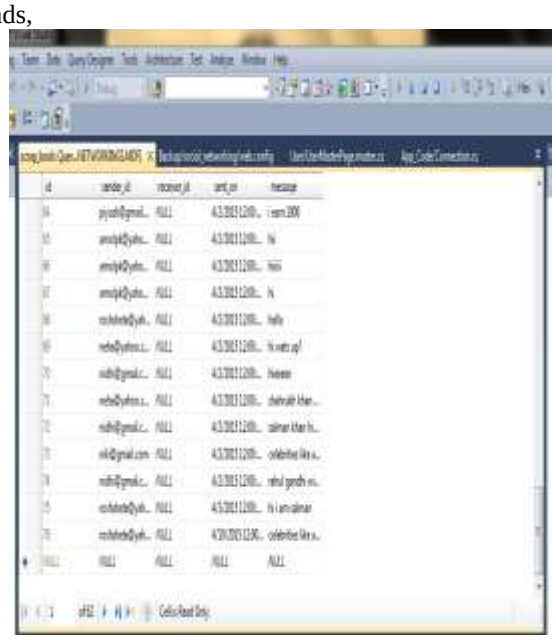

Figure 8. Comments/Posts

C. Phase-3: Data Preprocessing

In preprocessing, the useless data will be removed that is filtering of user generated data.

Removing stop words, punctuation, symbols is necessary. Natural Language Processing is used for filtering.

1) Lexical analysis: The lexical analyzer covert sentences into words then words convert into characters.

2) *Elimination of punctuations:* Remove punctuations like comma, full stop etc.

3) Elimination of symbols: Remove symbols like @, # etc.

4) Elimination of stop words: Remove words like in, of, the, is, and, for etc.

Here, dictionary of stop words is created which will grow every time. The comments will be spitted and each word will be compared with stop word dictionary. If it matches, then that word or character will be removed.

stopwords: Quer...NETWORKINGS.MDF)

id	stopword
1	a
2	about
3	above
4	after
5	again
6	against
7	all
8	am
9	an
10	and
11	any
12	are
13	aren't
14	as

Figure 9. Stop word dictionary

D. Phase-4: Control Spamming

Spamming is nothing but unwanted behavior of spammer. It is nothing but posting useless and vulgar words in comment/post. Nowadays, some people deliberately use vulgar words. Those words really not related or we can say no need of using such words in comments as it creates spam. It is necessary to remove such words from user comments. Because, nowadays children started creating account on social media by hiding their real age. So the aim is to replace vulgar words with '****'.

For spam control, dictionary of slang words is created. So, whenever user use any slang word in the post or comment that word matches with the words available in the dictionary and it replaces with the stars (****).

For example if user posted something and his/her friend commented "you dog", so this comment will be replaced by "you ****". Because the word dog is slang word and it is defined in the dictionary.

 ISSN: 2321-8169
4557 - 4563

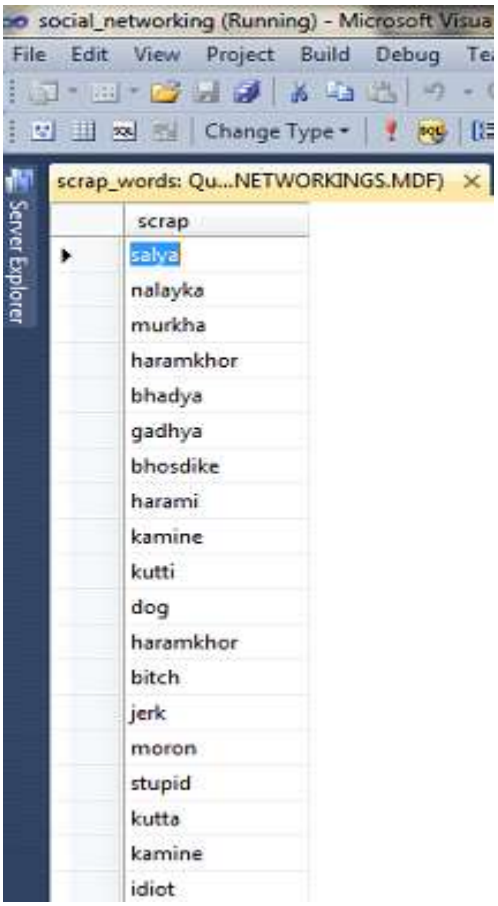

Figure 10. Slang word dictionary

Figure 11. Example of Spam

E. Phase-5: Event Detection

1) Topic allocation and Topic detection:

This is the first part of event detection where event will be detected by field wise. It means whether the given comment related to bollywood, politics, sports, education or business.

If comment does not exist in anyone of it, then it will be shown in 'other'. First the dictionaries of bollywood related words, business related words, and politics related words are created. So the process is that, each comment/post will be split word by word. Then each word will be compared with the dictionary words. Then if any word of comment/post is match with one of dictionary after that comment\post will be shown in the respective field. For clustering Bisect K-means algorithm is used.

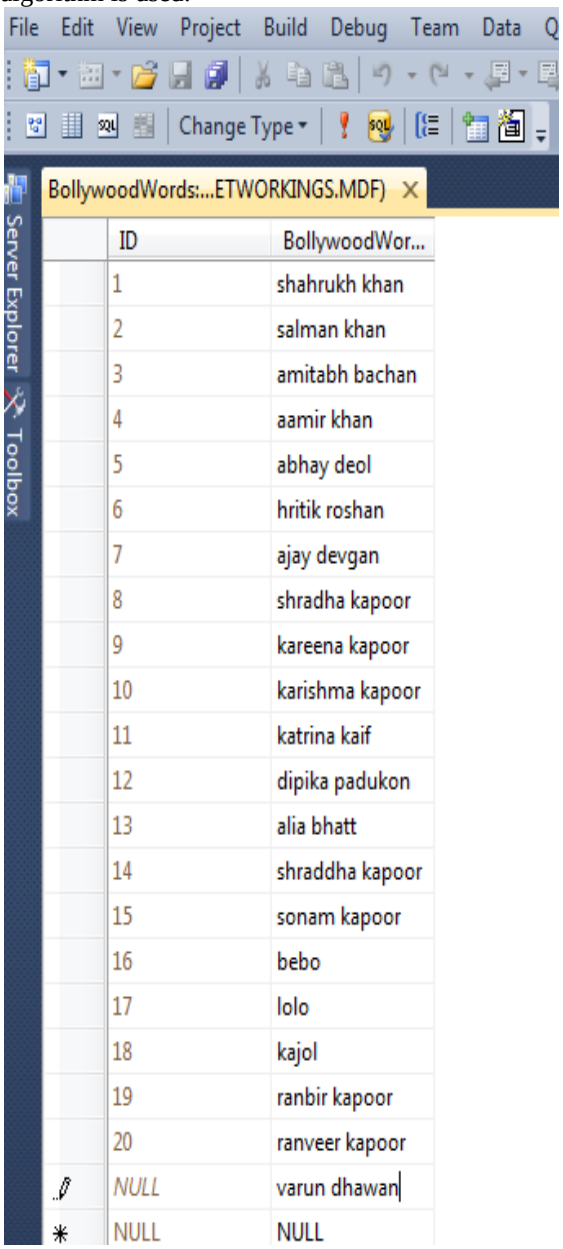

Figure 12. Bollywood word dictionary

 ISSN: 2321-8169
4557 - 4563

PolyticsWords: Q...NETWORKINGS.MDF)

ID	PolyticsWords
1	Narendra modi
2	Indira gandhi
3	Soniya gandhi
4	Atal bihari
5	Rahul gandhi
6	Jay lalita
7	mamta banerjee
8	lalu yadav
9	nitish kumar
10	prithviraj chauh...
11	devendra phad...
12	salman khurshid

Figure 13. Politics dictionary

Id	sportwords
1	yuvraj singh
2	sachin tendulkar
3	suresh raina
4	virendar sehwag
5	rahul dravid
6	ajinkya rahane
7	shikhar dhawan
8	rohit sharma
9	umesh yadav
10	ravichandran as...
11	bhuvneshwar k...
12	mohammed sh...
13	ishant sharma
14	ambati rayudu

Figure 14. Sports dictionary

Figure 15. Categorized events

2) Topic Trend:

In topic trend, the term which is appearing frequently that will be counted using machine learning concept. The word which is appearing more than 5 times in a conversation will be consider as current event. For topic trend, the hybrid algorithm is designed,

Hybrid Algorithm:

Topic = Ø

For each term t in T do

S=t;

$D_s = D_t$

ContinueExpanding=true;

Repeat

1. $\hat{t} = \text{GetBestMatchingTerm}(D_s, S, T)$

if $t = \hat{t}$ then

$S = S \; U \; \hat{t}$

$D_s = D_s + D_{\hat{t}}$

if $D_s > 5$ *then*

Topic = Topic U S

else goto 1;

else

ContinueExpanding = false;

end

Where,

 ISSN: 2321-8169
4557 - 4563

T = the set of candidate terms,
t = a candidate term,
S = set of new terms,
D_s = vector for S,
D_t = vector for D.

So first the topic is empty, then t the term of candidate will be added to S. After that each other term will be compare to term t in S. It it matches the vector will be incremented. And the term which have count greater than 5 that will be consider the term related to the topic term. And, then the subset of topics which having that frequent term will be extracted as trend.

Figure 16. Example of topic trend

V. RESULT ANALYSIS

Here narrated graphs of bollywood, business, education, politics and sports. The graph line is posts per 2 days, days on X-axis and posts on Y-axis. Values are predefined for posts if posts are 25 in 2 days it will show 1500 and like that.

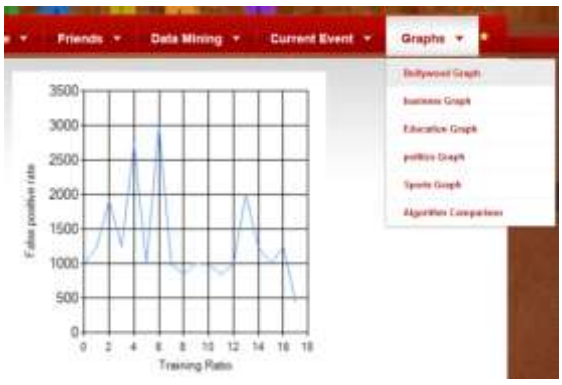

Figure 17. Bollywood graph

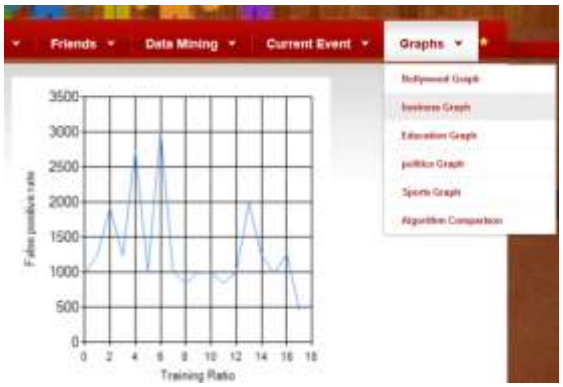

Figure 18. Business graph

Figure 19. Education graph

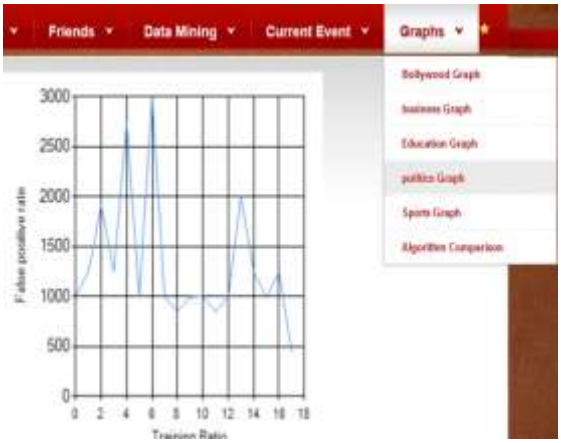

Figure 20. Politics graph

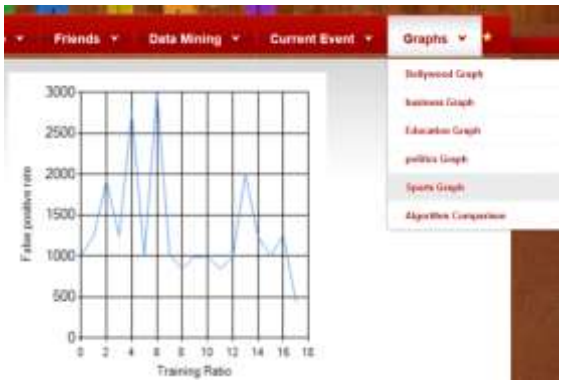

Figure 21.. Sports graph

 ISSN: 2321-8169
4557 - 4563

The below graph is efficiency graph, first the efficiency of previous algorithm is calculated. As the proposed algorithm is hybrid i.e. idea of previous algorithm is combined with proposed algorithm. But accuracy of proposed algorithm is more because in that the filtering using NLP, clustering using Bisect K-means and control spamming concept is used and it is not used in previous algorithm.

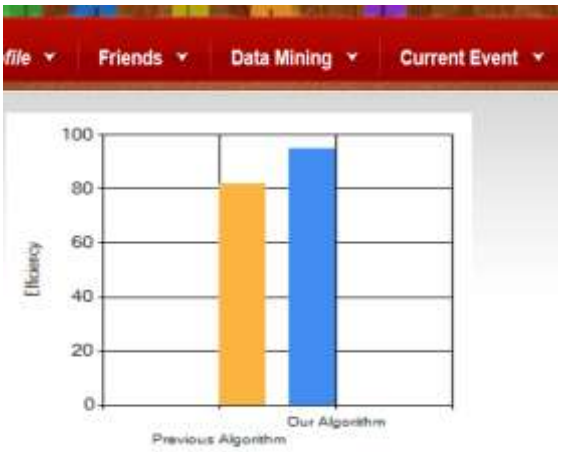

Figure 22. Efficiency graph

At last, it is analyze that people more talk about celebrity's death, celebrity's moves, about sports and elections.

VI. CONCLUSION

Here, the task is to identify real time current event from user generated data in account. User generated data is posts/comments given by people on any news or story of real world. Posts/comments are formed by sequence of words, keywords or term. So to perform this task there is need of user data stream. Also, the removal of slang words from user data is another approach covered here. For showing results one website is created. Data mining technique clustering is used and one hybrid algorithm is designed for current event detection. Without user generated data it is not possible to implement this approach. But even if it needs user data stream, this approach is specific and generic to any type of event or topic. For keeping the approach updated and current, the time frame is allotted. That means, after every two days the topic trend will be changed. The previous detected topic will be removed with the time. So, the update rate is of 2 days.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my guide Prof. S. W. Mohod, coordinator of M.Tech. in Computer Science and Engineering branch, for his constant guidance. I am extremely grateful to him for his sincere, expert and valuable guidance extended to me.

REFERENCES

[1] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon Papadopoulos, Ryan Skraba, Ayse Göker and Ioannis Kompatsiaris, "Sensing Trending Topics in Twitter", IEEE Transactions on Multimedia, Vol.15, No.6, October 2013.

[2] Roshani M. Shete, and Prof. S. W. Mohod, "Identification of Current Events and Control Spamming from Social Networking Site-A Review", in IJERT International Journal of Engineering Research & Technology, Volume. 4, Issue. 01, January-2015.

[3] Roshani M. Shete, and Prof. S. W. Mohod, "Using Natural Language Processing for Detection of Events and Spam Control from User Data Stream in Social Sites", in IJERT International Journal of Engineering Research & Technology, Volume. 4, Issue. 04, April-2015.

[4] Jurafsky D. and Martin, J., "Speech and Language Processing", Prentice Hall, Upper Sale River, NJ 2000.

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

[6] L. M. Aiello, M. Deplano, R. Schifanella, and G. Ruffo, "People are strange when you're a stranger: Impact and influence of bots on social networks," in Proc. ICWSM: 6th AAAI Int. Conf. Weblogs and Social Media. AAAI, 2012, pp. 10–17.

[7] 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.

[8] Sayyadi, M. Hurst and A.Maykov, "Event detection and tracking in social streams," in ICWSM, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds. Palo Alto, CA, USA: AAAI Press, 2009.

[9] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in Proc. KDD: 15th ACM Int. Conf. Knowledge Discovery and Data Mining, New York, NY, USA, 2009, pp. 497–506.

[10] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003

[11] D. A. Shamma, L. Kennedy, and E. F. Churchill, "Peaks and persistence: Modeling the shape of microblog conversations," in Proc. CSCW: ACM Conf. Computer Supported Cooperative Work, New York, NY, USA, 2011, pp. 355–358

[12] J. Yang and J. Leskovec, "Patterns of temporal variation in online media," in Proc. WSDM: 4th ACM Int. Conf. Web Search and Data Mining, New York, NY, USA, 2011, pp. 177–186.

[13] S. Papadopoulos, Y. Kompatsiaris, and A. Vakali, "A graph-based clustering scheme for identifying related tags in folksonomies," in Proc. DaWaK: 12th Int. Conf. Data Warehousing and Knowledge Discovery. Berlin, Germany: Springer-Verlag, 2010, pp. 65–76.

[14] Manning C. and Schutze H. "Foundations of Statistical Natural Language Processing", MIT Press, Cambridge, MA, 1999.