Anomaly Detection in Microblogging Website Trending Topics through Linked Based Method

Ms. Supriya Singh[1], Mr. Sachin Harne [2]
Research Scholar [1], Professor [2]
[12]Dept. of Computer Science and Engineering,
[12]RSR Rungta College of Engineering and Technology Bhilai, Chhattisgarh, India
(E-mail: [1]er.supriyasinghcse@gmail.com, [2]sachin.harne2027@gmail.com)

***Abstract*—** Finding of emerging or trending topics become more interested in the fast development of social networking sites. The data exchanged in the social networking site's post include the text contents, yet in addition images, URL's and video consequently conventional-term-frequency-based approaches may not be suitable in this context. Rise of topics is for the most part centered by the social parts of these networks. In this paper we propose a novel strategy where the user tweets are analyzed by the probabilistic function. We propose joint probabilistic based algorithm count of user tweets, re-tweets and notices. The proposed algorithm outperforms the customary key based strategy as far as start and end time computation of tweets with specific hashtag.

***Keywords*—** *Topic detection, Social Networks, anomaly detection, Burst Detection.*

I. INTRODUCTION

These days, communication over social networking sites, for example, Facebook, Twitter, and so on has been increasing its esteem. The advancement of social networking sites increases the messages or data exchanged between the users are the text contents, yet in addition images, videos and URLs. Especially, we are engaged with the discovery of emerging areas from social media streams which are posted by a large number of users. It tends to be utilized to capture the unedited voice of normal or conventional individuals by means of social media, which creates computerized "breaking news", or underground political developments or find hidden market needs. Henceforth compared to conventional media, the social networking sites are ready to capture the unedited voice of common individuals as ahead of schedule as could reasonably be expected. Therefore, the test is to find the rise of topics soonest conceivable at a moderate number of false positives.

The real difference that makes social networking sites more social and mainstream is the presence of notices. Here, we mean by presence of notices is to link with different users of a similar social networking sites as reply-to, message-to, retweet-of or expressly as text. Single post may contain various notices. There are some user may incorporate notices in their posts rarely; there are different users too which may reference their friends constantly. A few users, for example, superstars may receive makes reference to consistently; and for other being referenced may be a rare event. Therefore, in social media, notice resembles a language with the quantity of words equivalent to the quantity of users.The tremendous prominence of social networking sites, for example, blogs, comments and posts represent critical opportunities. A huge volume of data is created regularly by bloggers and different contents produced over around the world, therefore giving a significant real time view, opinion, feelings, comments, activities, intention and trends of people just as group over the globe. These data may empower right on time for the detection of emerging topics, issues and trends for the significant estimation of interest. Be that as it may, it is troublesome for the user to discover the signature of emerging topics and trend which are covered in the enormous and generally irrelevant data. Therefore, to find helpful single point quickly out of a large number of online data produced day by day by the social networking sites is extremely troublesome.

We are interested to find emerging topics from social networking streams which is based on observing the notice behavior of different users. Here is our fundamental supposition that will be that another emerging point is something in which individuals like commenting, talking about, or sending the valuable data further to their different friends. Therefore, the conventional approaches for the subject detection which have mostly been worried about the frequency of textual words. Frequency of textual words based methodology could experience the ill effects of the ambiguity brought about by the homonyms or synonyms. The objective language may likewise require entangled preprocessing. Moreover, when the content of messages are non-textual data then it isn't connected. Then again, the one of a kind words framed by the notices requires small preprocessing to get the data or data is isolated from the contents and are accessible notwithstanding the nature of the contents [1].

Social networking sites have a few difficulties like discovery of topics, topic pattern from text, bursts, change points and anomaly detection. To beat this question, there are different strategies and different models have been proposed. The difficult assignment is to discover the anomaly [2].

Anomaly detection manage finding the patterns present in a data-sets whose behavior is unexpected which implies not normal. These kind of unexpected behaviors of the patterns in data-sets are likewise called as anomalies or exceptions. Every single anomaly can't be constantly detected or arranged as an assault yet it tends to be sorted as a surprising behavior which is previously not detected or known. That surprising behavior might be or may not be that much harmful.

The anomaly detection gives extremely noteworthy just as critical data in an assortment of applications, for instance personality thefts or the Credit card thefts. At the point when data must be analyzed with respect to discover the relationship or to calculate or predict the known or unknown data mining methods are utilized. What's more, therefore, this incorporates the grouping of data-sets, classification and furthermore machine based learning procedures. So as to achieve larger amount of accuracy, the Hybrid approaches are likewise being created on detection of anomalies. This methodology helps us to endeavor to join the current data mining algorithms which produce better results. Therefore recognizing the abnormal or the unexpected behavior or anomalies will produce to study and this will arrange it as into another kind of attacks or a specific sort of interruption.

II. RELATED WORK

Li C et al. [3], the author proposes a segment-based event detection system for tweets, called Twevent. Twevent first detects bursty tweet segments as event segments and after that clusters the event segments into events considering both their frequency distribution and content similarity. Even more especially, each tweet is part into non-overlapping segments (i.e., communicates maybe allude to named components or semantically critical data units).

Marcus A et al. [4], TwitInfo, a system for visualizing and summarizing events on Twitter. TwitInfo allows users to browse a large collection of tweets using a timeline-based display that highlights peaks of high tweet activity. A novel streaming algorithm automatically discovers these peaks and labels them definitively utilizing text from the tweets. Users can drill down to subevents, and investigate promote by means of geolocation, sentiment, and prevalent URLs.Author contributes a recall-normalized aggregate sentiment visualization to deliver more honest sentiment diagramsAn interview with a Pulitzer Prize-winning journalist suggested that the system would be especially useful for understanding a long - running event and for identifying eyewitnesses. Quantitatively, our system can identify 80-100% of manually labeled peaks, facilitating a relatively complete view of each event studied.

Gaglio S et al. [5], the proposed extended and enhances the Soft Frequent Pattern Mining (SFPM) algorithm by overcoming its limitations in managing dynamic, real-time, detection situations. Specifically, with the end goal to acquire timely outcomes, the flood of tweets is composed in dynamic windows whose size depends both on the volume of tweets and time. In particular, in order to obtain timely results, the stream of tweets is organized in dynamic windows whose size depends both on the volume of tweets and time. Since we aim to highlight the user's point of view, the set of keywords used to query Twitter is progressively refined to include new relevant terms which reflect the emergence of new subtopics or new trends in the main topic. The real-time detection system has been evaluated during the 2014 FIFA World Cup and experimental results show the effectiveness of our solution.

Stilo G et al. [6], Author present a novel system for clustering words in micro blogs, based on the likeness of the related fleeting arrangement. Our technique, named SAX*, uses the Symbolic Aggregate ApproXimation algorithm to discretize the transient arrangement of terms into a little game planof levels, provoking a string for each.

Unankard S et al. [7], the paper propose a methodology for the early detection of rising hotspot events in interpersonal organizations with area affectability. Author considers the message-referenced locations for recognizing the locations of events. In the methodology, author perceives solid connections between user locations and event locations in detecting the developing events. The assessment of methodology based on a true Twitter dataset. The tests exhibit that the proposed methodology can reasonably detect developing events concerning user locations that have assorted granularities.

De Boom C et al. [8], paper spreads out the focal points of the semantics-driven event clustering algorithms created, looks at a novel procedure to help in the generation of a ground truth for event detection purposes, and investigations how well the algorithms upgrade over pattern. Author found that allocating semantic data to every individual tweet results in just a more regrettable act in F1 measure stood out from benchmark.

McMinn AJ et al. [9], Author assess the methodology on an expansive scale corpus of 120 million tweets covering more than 500 events and show that it can detect by and large a greater number of events than current cutting edge approaches while similarly improving exactness moreover, holding low computational multifaceted nature.

Zhou D et al. [10], Author propose a general unsupervised system to examine events from tweets, which contains a pipeline method of sifting, extraction and classification. To sift through loud tweets, the sifting step misuses a lexicon-based way to deal with discrete tweets that are event-related from those that are certainly not. By then, based on these event-related tweets, the organized portrayal of events is extricated and arranged consequently using an unsupervised Bayesian model without the usage of any marked data.

Li J et al. [11], Author propose a scattered and steady worldly subject model for microblogs called Bursty Event dEtection (BEE+). BEE+ can detect bursty events from short-text dataset and model the worldly data. Also, BEE+ forms the post-stream steadily to follow the subject floating of events after some time. As such, the latent semantic records are secured beginning with one time period then onto the next.

Di Wang et al. [12], article looks misuse of Twitter data in the traffic reporting area. A key test is the methods by which to perceive pertinent data from a huge proportion of user-produced data and subsequently research the applicable data for customized geocoded episode detection. The article proposes a minute traffic caution and cautioning system based on a novel latent Dirichlet allocation (LDA) approach ("tweet-LDA"). The system is assessed and seemed to perform better than related methodologies.

Dou W, Wang K, Ribarsky W [13] and Allan J [14], proposed a methodologies for topic detection have essentially been worried about the frequencies of (textual) words.A term frequency based methodology could suffer from the ambiguity caused by synonyms or homonyms. It might likewise require complicated preprocessing (e.g., segmentation) depending on the target language. In addition, it can't be connected when the substance of the messages are mostly non-textual data. Then again, the "words" formed by mentions are unique, requires small preprocessing to get (the data is regularly isolated from the substance), and are accessible regardless to the idea of the contents.

III. METHODOLOGY

In this area we will talk about the technique utilized for breaking down the anomaly of twitter users. The algorithm is utilized here is Join Probabilistic Model. This model is utilized for preparing of twitter dataset.

It is used to capture the notices, re-tweets likelihood of users. It is based on probabilistic theory. For every tweet the likelihood are calculated and based on that the users are related to same tweets.

The equation of Join probabilistic model is presented beneath numbered equation (1)

$$P(\,k\,,V \,|\, \theta\,,\{\pi_{v}\}\,) = P\,(k\,|\,\theta)\prod_{v' \in v} \pi_{v'} \qquad (1)$$

The proposed work stream of anomaly detection framework is presented in figure. fig.1. There are different modules. The modules are depicted in this segment.

A. Social Network Service

It is an administration given by numerous social network sites. The services like posting, commenting, and talking are the services offered by numerous Social networking sites.

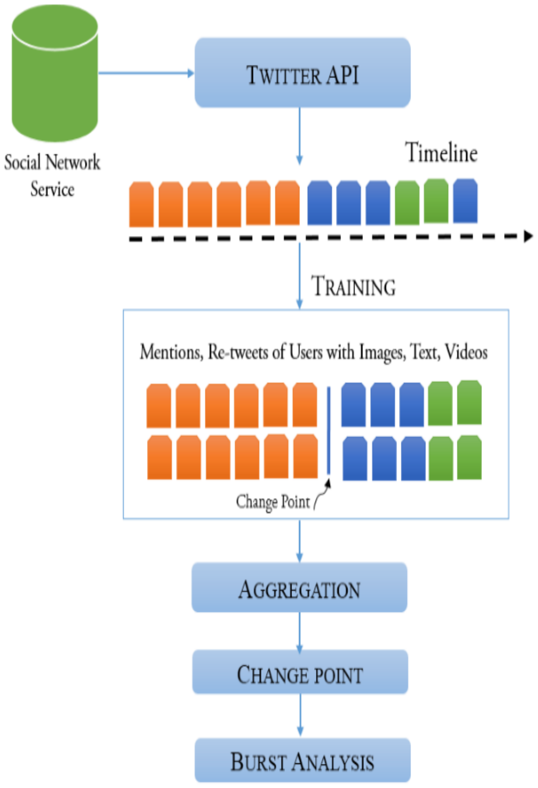
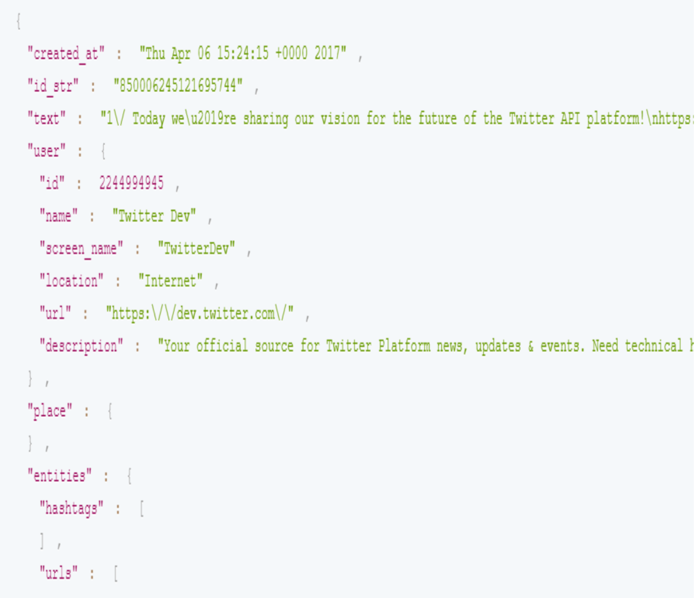


Fig. 1. Flow chart of Proposed Framework

B. Twitter API

The twitter API is utilized for downloading every one of the tweets from the twitter account. The accompanying data are downloaded.



```
{
 "created_at" :  "Thu Apr 06 15:24:15 +0000 2017" ,
 "id_str" :  "850006245121695744" ,
 "text" :  "1\/ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps:
 "user" :  {
  "id" :  2244994945 ,
  "name" :  "Twitter Dev" ,
  "screen_name" :  "TwitterDev" ,
  "location" :  "Internet" ,
  "url" :  "https:\/\/dev.twitter.com\/" ,
  "description" :  "Your official source for Twitter Platform news, updates & events. Need technical
 } ,
 "place" :  {
 } ,
 "entities" :  {
  "hashtags" :  [
  ] ,
  "urls" :  [
```

a. Screen ID

b. Screen Name

c. Tweets

d. Photo

e. Mentions

All Twitter sets, with named characteristics and related qualities. These traits, and their state are utilized to portray objects.

At Twitter we serve numerous objects as JSON, including Tweets and Users. These objects all exemplify center qualities that depict the object. Each Tweet has a creator, a message, a remarkable ID, a timestamp of when it was posted, and here and there geo metadata shared by the client. Every User has a Twitter name, an ID, various supporters, and regularly a record bio.

C. Timeline

Timeline is the gathering APIs that arrival Tweets give that information encoded utilizing JavaScript Object Notation (JSON). JSON depends on key-esteem of the tweets where every one of the tweets is present. It might contain text, images and videos with specific hashtag.

D. Training

The preparation part is for finding the users tweets likelihood. The join probabilistic model is utilized. It considers Images, Text, Video files for training of tweets.We describe a post in an social network stream by the quantity of mentions k it contains, and the set V of names (IDs) of the mentionee (users who are referenced in the post).

There are two kinds of boundlessness we need to consider here. The first is the number k of users referenced in a post. In spite of the fact that, by and by users can't specify many different users in a post, we might want to abstain from putting a fake point of confinement on the quantity of users referenced in a post. Rather, we will expect a geometric conveyance and coordinate out the parameter to keep away from even a certain confinement through the parameter. The second kind of limitlessness is the quantity of users one can make reference to.

Formally, we think about the accompanying joint likelihood distribution in equation (2).

$$\mathbf{P(k ,V \,|\, \theta , \{\pi_{\upsilon}\}) = P (k \,|\, \theta) \prod_{\upsilon' \epsilon v} \pi_{\upsilon'}} \qquad \textbf{(2)}$$

Where,

K = |V| probability of number of mentions

and P(k | θ) defined as the geometric distribution

υ is the user and v is its mentionee probability sites.

E. Aggregation

Aggregation module is utilized for figuring of normal time delay for every tweet. The time delay is calculated in this area and yielded to change point module.sites.

F. Change Point

It is utilized to calculate the start and end time of the tweets. It calculates the time of start of that tweet and calculates the end of the tweets.In measurable examination, change identification or change point recognition attempts to distinguish times when the likelihood appropriation of a stochastic procedure or time arrangement changes. By and large the issue concerns both distinguishing whether a change has happened, or whether a few changes may have happened, and recognizing the seasons of any such changes.

Linguistic change detection refers to the ability to detect word-level changes across multiple presentations of the same sentence. Researchers have found that the amount of semantic overlap (i.e., relatedness) between the changed word and the new word influences the ease with which such detection is made (Sturt, Sanford, Stewart, & Dawydiak, 2004). Additional research has found that focusing one's attention to the word that will be changed during the initial reading of the original sentence can improve detection.

This was shown using italicized text to focus attention, whereby the word that will be changing is italicized in the original sentence (Sanford, Sanford, Molle, & Emmott, 2006), as well as using clefting constructions such as "It was the tree that needed water." (Kennette, Wurm, & Van Havermaet, 2010). These change-detection phenomena appear to be robust, even occurring cross-linguistically when bilinguals read the original sentence in their native language and the changed sentence in their second language (Kennette, Wurm & Van Havermaet, 2010). Recently, researchers have detected word-level changes in semantics across time by computationally analyzing temporal corpora (for example: the word "gay" has acquired a new meaning over time) using change point detection.

G. Burst Analysis

This module compares the time taken by key based and link based anomaly detection algorithm. The key based does not consider the re-tweets and link based considers all the part of tweets, for example, re-tweets and notices.

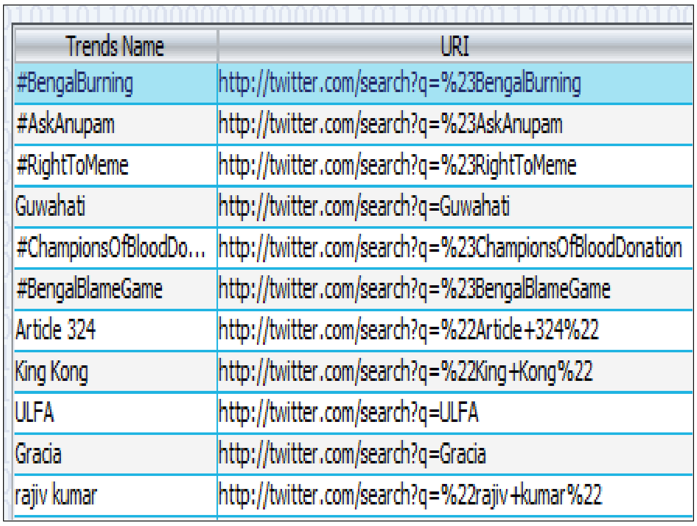
IV. DATA STATISTICS

In this section we will go through some of the dataset which we have used for analysis. We have utilized twitter API for downloading the twitter content. The content of the twitter not only limited to text based tweets but also contains videos and images. The time line is generated for hash tag and based on that all the tweets are downloaded.

Keywords which can be used are based on current events. For eg. At present India vs Bangladesh match is going on we have consider that event for analysis. Like this many events can be considered like IPhone Sale, Samsung S10 Sale Event, Stock Market Up and Down Event, Sports Event, Olympic Events, Birthdays event of famous celebrities etc.Data which are extracted are real time and hence need to be chosen by the application executor. We have utilized around 1000 tweets for analysis, due to processing speed limitation. These 1000 tweets contain text as well as images and videos for analysis. We have trained our classifier using probabilistic model that is statistical approach to detect the change.

V. RESULT AND DISCUSSION

The examination is performed in the Eclipse IDE utilizing Java language. Fig2. Demonstrates the various Twitter Trends extracted from Twitter using Twitter API, from them #BengalBurning hashtag is considered for analysis. From all the extracted Twitter trends, any trend can be selected for analysis. Joint probabilistic based algorithm applied here for the count of user tweets, re-tweets and notices.



Trends Name	URI
#BengalBurning	http://twitter.com/search?q=%23BengalBurning
#AskAnupam	http://twitter.com/search?q=%23AskAnupam
#RightToMeme	http://twitter.com/search?q=%23RightToMeme
Guwahati	http://twitter.com/search?q=Guwahati
#ChampionsOfBloodDo...	http://twitter.com/search?q=%23ChampionsOfBloodDonation
#BengalBlameGame	http://twitter.com/search?q=%23BengalBlameGame
Article 324	http://twitter.com/search?q=%22Article+324%22
King Kong	http://twitter.com/search?q=%22King+Kong%22
ULFA	http://twitter.com/search?q=ULFA
Gracia	http://twitter.com/search?q=Gracia
rajiv kumar	http://twitter.com/search?q=%22rajiv+kumar%22

Fig.2. list of twitter trends extracted from twitter using twitter API

Fig. 3. Demonstrates the correlation of key based and link based anomaly.

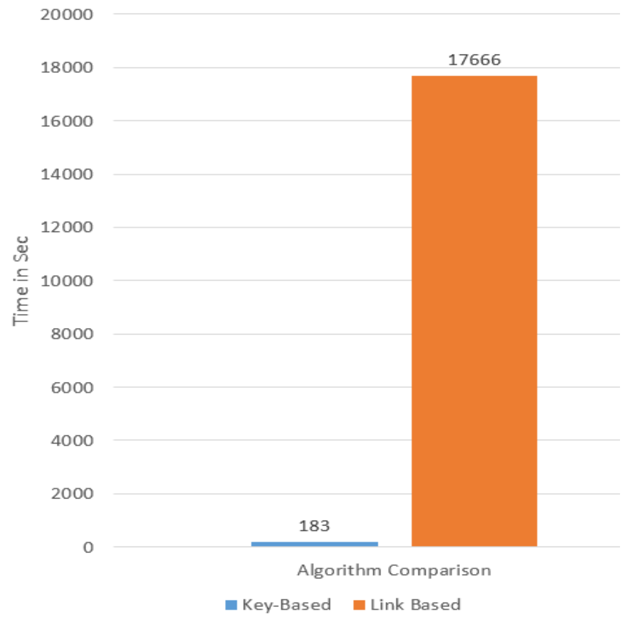


Fig. 3. comparison of algorithms in capturing tweets time

The graph shown in fig.3 is outcome of Burst Analysis. In above figure comparison is performed between key based detection and link based detection. The key based approach method considers only the text content and tweets only. Therefore the timeline creation for this method is for shorter period of time. The time which user tweets based on hashtag. This method cannot able to detect changes effectively because it considers only tweets.

The proposed approach is link based model based on probability. The proposed method detects the changes when it really ends. It considers start and end time from re-tweets, images and videos. Using this approach effective analysis of emerging trend through twitter can be performed. This helps in detecting emerging topics from social system streams dependent on monitoring the mentioning behaviour of users.

IV. CONCLUSION

In this paper, we compare two fundamental strategies for example key based and link based. The key based algorithm can't distinguish the tweets of the users who have re-tweets or notice some other individual into it. The link based algorithm which recognizes based on re-tweets and notice in the tweets outflanks the key based algorithm in term of discovering start and end time. We have considered 1000 tweets which contain text as well as images and videos for analysis. We have trained our classifier using probabilistic model that is statistical approach to detect the change.

Proposed approach is effectively detecting the start and end timing from the tweets and re-tweets. It also consider images and videos posted by users. The start time of key based method is 0 and end time is 183 sec. While the start time of linked based method is 0 and end time is 17666. There is huge gab in emerging trend detection. The linked based method outperforms the key based method with respect to detection of ending time.

VI. FUTURE SCOPE

All the investigation displayed in this paper was led offline, yet the system itself can be connected on the web. We are intending to scale up the proposed way to deal with handle social streams progressively. It would likewise be fascinating to join the proposed connection inconsistency model with content based methodologies, in light of the fact that the proposed link abnormality model does not quickly determine what the peculiarity is. Blend of the word-based methodology with the link abnormality model would profit both from the execution of the mention model and the instinct of the word-based methodology.

REFERENCES

[1] Toshimitsu Takahashi, Ryota Tomioka, Kenji Yamanishi, ”Discovering Emerging Topics in Social Streams via Link Anomaly Detection”, 2012 IEEE Transactions on Knowledge and Data Engineering.

[2]Richard Colbaugh, Kristin Glass, “Detecting Emerging Topics and Trends Via Predictive Analysis of 'Meme' Dynamics”, 2011 IEEE.

[3]Li C, Sun A and Datta A., “Twevent: segment-based event detection from Tweets”, In: Proceedings of the ACM international conference on information and knowledge management (CIKM ’12), Maui, HI, 29 October–2 November 2012, pp. 155–164. New York: ACM.

[4]Marcus A, Bernstein MS, Badar O, “TwitInfo: aggregating and visualizing microblogs for event exploration”, In:Proceedings of the SIGCHI conference on human factors in computing systems (CHI’11), 2011, pp. 227–236. New York, ACM.

[5]Gaglio S, Re GL and Morana M., “A framework for real-time Twitter data analysis”, Comput Commun 2016; 73: 236–242.

[6]Stilo G and Velardi P., “Efficient temporal mining of micro-blog texts and its application to event discovery”, Data Mining Knowledge Disc 2016; 30: 372–402.

[7]Unankard S, Li X and Sharaf MA, “Emerging event detection in social networks with location sensitivity”, World Wide Web 2015; 18(5): 1393–1417.

[8]De Boom C, Van Canneyt S and Dhoedt B, “Semantics-driven event clustering in Twitter feeds”, In: Proceedings of the 5th workshop on making sense of microposts, vol. 1395, 2015, pp. 2–9, CEUR.

[9]McMinn AJ and Jose JM. “Real-time entity-based event detection for Twitter. In: Mothe J, Savoy J, Kamps J (eds)” Experimental IR meets multilinguality, multimodality, and

interaction" 6th international conference of the CLEF association (CLEF '15). Berlin: Springer, 2015, pp. 65–77.

[10]Zhou D, Chen L and He Y,"An unsupervised framework of exploring events on Twitter: filtering, extraction and categorization", In: Proceedings of the AAAI conference on artificial intelligence, 2015, pp. 2468–2475.

[11]Li J, Wen J, Tai Z, Zhang R. and Yu W., "Bursty event detection from microblog: a distributed and incremental approach" Concurrency Computat.: Pract. Exper. (2015); 28: 3115–3130.

[12] Di Wang, Ahmad Al-Rubaie, Sandra Stinčić Clarke, and John Davies. 2017. Real-Time Traffic Event Detection from Social Media. ACM Trans. Internet Technol. 18, 1, Article 9 (November 2017), 23.

[13]Dou W, Wang K, Ribarsky W, "Event detection in social media data", In: Proceedings of the IEEE VisWeek workshop on interactive visual text analytics – task driven analytics of social media content. 2012, pp. 971–980,

[14] Allan J., "Topic detection and tracking: event-based information organization", vol. 12. Berlin: Springer, 2012

[15] Abdelhaq, H.; Sengstock, C.; and Gertz, M. 2013. Even tweet: Online localized event detection from twitter. Proceedings of the VLDB Endowment 1326–1329.

[16] Liu, X.; Zhou, X.; Fu, Z.; Wei, F.; and Zhou, M. 2012. Exacting social events for tweets using a factor graph. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 1692–1698

[17]T.Sakaki, M.Okazaki, Y.Matsuo, Earthquake shakes Twitter users: Real time event detection by social sensors, in: Proceeding s of the19th International Con-ference on World Wide Web, WWW' 10, ACM, NewYork, NY,USA, 2010,pp.851-860.

[18]W.Xie, F.Zhu,J.Jiang, E.-P.Lim,K.Wang, Topic sketch: Real-time bursty topic detection from Twitter ,in : Data Mining(ICDM),2013IEEE 13th International Conference on,2013,pp.837–846.

[19] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang,Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In WWW, pages 331–340,2010.

[20] Hong L, Dom B, Gurumurthy S, Tsioutsioulikis K (2011) Time dependent topic model for multiple text streams. In: ACM conference on knowledge discovery and data mining KDD 2011, San Diego

[21] Huang B, Yang Y, Mahmood A, Wang H (2012) Microblog topic detection based on LDA model and single-pass clustering RSCTC 2012, LNAI 7413, pp. 166–171.

Author Profile:

Ms. Supriya Singh is Research Scholar, Department of Computer Science and Engineering in RSR Rungta College of Engineering and Technology Bhilai, Chhattisgarh, India She is doing research under the guidance of Mr. Sachin Harne , Professor, Department of Computer Science and Engineering in RSR Rungta College of Engineering and Technology Bhilai, Chhattisgarh, India .She received BE degree in Computer Science from Chhattishgarh Swami Vivekanad Technical University, Bhilai, Chhattisgarh, India.Her area of interest is Data Mining.

Mr.Sachin Harne is Professor, Department of Computer Science and Engineering in RSR Rungta College of Engineering and Technology Bhilai, Chhattisgarh, India. He received M.Tech degree in Computer Science and Engineering from Chhattishgarh Swami Vivekanad Technical University, Bhilai, Chhattisgarh, India.