Framework for Real-World Event Detection Through Online Social Networking Sites

Ritesh Srivastava, M. P. S. Bhatia, Veena Tayal and J. K. Verma

Abstract In recent few years, due to the exponential growth of users on online social networking sites (OSNs), mainly over micro-blogging sites like Twitter, the OSNs now resemble the real world very cohesively. The excess of continuously user-generated online textual data by OSNs that encapsulates almost all verticals of the real world has attracted many researchers who are working in the area of text mining, natural language processing (NLP), machine learning, and data mining. This paper discusses the feasibility of OSNs in detecting real-world events from the horizon of the virtual world formed over OSNs. Moreover, this paper also describes the framework for real-world event detection through online social networking sites.

Keywords Online social network · Event detection · Social network analysis
Data mining · Text mining

1 Introduction

The evolution of Web 2.0 [1] allows users to interact and collaborate with each other on OSNs platform [2]. In recent years, an exponential growth in the users of OSNs has been witnessed. The numbers of users on OSNs are getting double in every five years with about 2.5 billion users this year. A projected result shows that about 39.9%

R. Srivastava (✉) · V. Tayal
CSE Department, FET, MRIIRS, Faridabad, India
e-mail: ritesh21july@gmail.com

V. Tayal
e-mail: veena.mittal06@gmail.com

M. P. S. Bhatia
NSIT, University of Delhi, New Delhi, India
e-mail: mpsbhatia@nsit.ac.in

J. K. Verma
Galgotias University, Greater Noida 201310, India
e-mail: jitendra.verma.in@ieee.org

L. C. Jain et al. (eds.), *Data and Communication Networks*, Advances in Intelligent
Systems and Computing 847, https://doi.org/10.1007/978-981-13-2254-9_17

of world population will become OSN users with the end of 2020. The users of micro-blogging sites (like Twitter) actively participate in sharing their opinions on various hot topics (e.g., personalities, products, and events) by posting their comments about the topics. Usually, the comments written by users of micro-blogging sites are short snippets of text that are limited to only a few characters. These short snippets of textual comments over OSN are also termed as online micro-texts [3]. The consistent posting of comments by millions of users of micro-blogging sites and the exponential increase in the number of users on micro-blogging sites have flourished two interesting aspects of social-networking-enabled micro-blogging sites:

(i) The WWW has now become a massive source of online micro-texts. Micro-texts are extremely subjective in nature as they generally contain the positive or negative opinion of users about an entity. Data mining for searching some interesting information from such data has gained the affection of many researchers in the previous years. Sentiment analysis (SA) is one of the most prominent ways for the analysis of this valuable collection of subjective data for making predictions in various events.

(ii) Another interesting facet of OSN is the creation of virtual communities. A virtual community is a set of social entities, especially human beings that are connected to each other on the basis of some common interests on any topics and events over OSN [2].

With the exponential growth of users on OSNs, the OSNs now resemble the corresponding real-world community very cohesively. As a conscience, any event that may be initiated in any of these communities has a significant reaction in both the communities and vice versa. This cohesiveness among the virtual and its corresponding real community has motivated many researchers and data analyst to sense the happenings of the real-world events through the contents of OSN.

Knowing about future has always been fascinating. Predictive analytics is a way by which one can predict the unknown future events. In the predictive analysis, historic and the current data are analyzed to make a prediction, which utilizes many methods such as statistics, data mining, and machine learning. The problem of the predictive analysis can be abstractly classified into two different set of problems.

(i) Making predictions of future by utilizing the current data.

(ii) Making predictions of some attributes of one observation space from another observation space at the same time.

Nowadays, the OSNs offer great opportunities for making predictions about the real-world happenings in both cases. From Fig. 1, it can be easily understood that the degree of cohesion (d) between a real-world community and its corresponding online virtual community formed over OSN is directly proportional to the number of active users on the online virtual community and the number of users commenting about an event occurred in the real world as given in (1) and (2).

$$d \propto \text{Number of active users on Online Virtual Communitiy} \tag{1}$$

$$d \propto \text{Number of users commenting about an event} \tag{2}$$

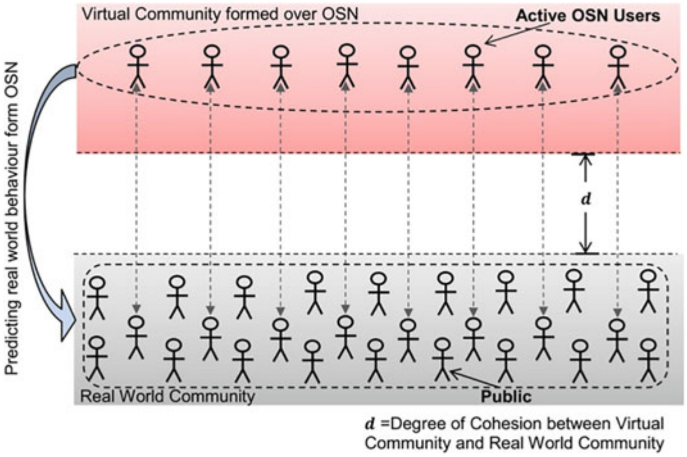
Virtual Community formed over OSN
Active OSN Users
Predicting real world behaviour form OSN
d
Real World Community
Public
d =Degree of Cohesion between Virtual Community and Real World Community

Fig. 1 Degree of cohesion between real world and virtual world of OSN

The consistent posting of comments by millions of users on micro-blogging sites and the exponential increase in the number of users on micro-blogging sites have made the WWW a massive collection of online micro-texts and increased the degree of cohesiveness between the real-world communities and their corresponding online virtual communities, respectively. The availability of massive data on OSN that represents the social and behavioral aspects of a big section of the population provides the immense possibilities to us to observe the public behaviors from the virtual world formed over OSN. Furthermore, it also opens up new opportunities to conduct predictive analysis for the future events. Consequently, an emerging area of research is to make the prediction of real-world happenings and behaviors of people in the real world by analyzing their behaviors in the virtual worlds of OSNs.

The social networking sites have proven their huge power of prediction in predicting the results of the events of real world. Recently, Twitter is utilized in various tasks such as monitoring, predicting, and analyzing the real-world events and activities like breaking news tracking [4], election prediction [5], natural disasters like earthquake [6], and crime, radicalization, and terrorism [7]. Certainly, the text stream mining of Twitter data can substitute the traditional polling [8].

We believe that the real-world activities could be monitored in real time by performing the real-time analysis of contents of micro-blogging site like Twitter. Such type of real-time analysis can enhance the real-time decision-making and an alternative for both types of predictive analysis as mentioned above.

2 Events and Event Detection Through OSN

The Oxford Dictionary defines the word ***event*** as a thing that happens or takes place, especially one of importance. An event is usually associated with time and location. The process of event detection from temporally ordered textual data can be explained as an automatic process for identifying novel events evolved meanwhile in that textual stream [9]. In early years of 2000s, with the evolution of Web 2.0, the enormous use of computer-mediated communications motivated researchers for automatic event detection from user-generated text stream. The event detection from textual data has long been discussed as topic detection and tracking (TDT) [9–11]. Most of the previous works related to event detection are implemented on the conventional news based textual contents from various news broadcasting media [12–14].

The task of event detection in the Twitter data stream can be broadly categorized into two: (i) specified or targeted events and (ii) unspecified (untargeted) events [15]. The targeted event detection is a kind of supervised process in which a sufficient amount of prior information is known in advanced; based on this information, the events are detected. For specified event detection, the prior information may include place, time, domain, description, and features about the event. For example, election in any country is kind of specified event as the date of the election, names of contesting parties as well as the names of the candidates are known in advance for performing the analysis. Conversely, in the case of unspecified event detection process, no clues about the event are known in advance. Moreover, an event can undergo with some sub-events. Sub-events can be defined as those events that occur in between the discussion duration of any event and cause significant impact over the points of discussion. Sub-events generally change the sentiment of the main event significantly. We state such event as sentimental events or sub-events. The task of event detection in the Twitter data stream can be broadly categorized in two: (i) specified or targeted events and (ii) unspecified (untargeted) events (Fig. 2) [15]. The targeted event detection is a kind of supervised process in which a sufficient amount of prior information is known in advance; based on this information, the events are detected. For specified event detection, the prior information may include place, time, domain, description, and features about the event. For example, election in any country is a kind of specified event as the date of the election, names of contesting parties as well as the names of the candidates are known in advance for performing the analysis. Conversely, in the case of unspecified event detection process, no clues about the event are known in advance. Moreover, an event can undergo with some sub-events. Sub-events can be defined as those events that occur in between the discussion duration of any event and cause significant impact over the points of discussion. Sub-events generally change the sentiment of the main event significantly.

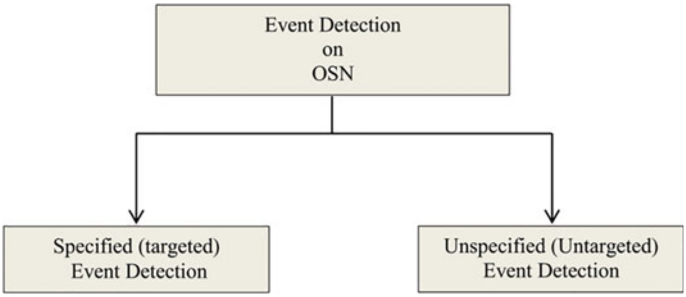

Fig. 2 Types of event detection task on OSN

2.1 Types of Events on OSN

Based on the ways, in which the events evolved in real world and discussed on OSN, we can classify the events on OSN into the following three types:

(i) Periodic events.
(ii) Sudden events.
(iii) Long-duration gradual events.

As depicted in the graphs of Fig. 3, the periodic events are those events that reoccurred periodically on certain pre-decided date and time. Over OSNs, such type of events hugely attracts comments of users on the date or time of event occurrence periodically. The periodic events are generally designated by certain prior information such as a list of frequent keywords and time of occurrence. For example, #FollowFriday is a periodic event that involves discussions on new movie release on every Friday of a week. The periodic events can further be classified on the basis of their recurrence intervals; for example, daily event may be (#GoodMorning), the weekly event may be follow Friday (#FollowFriday), and the yearly event may be any festival (#HappyChrismas). The next type of events is sudden events; such type of events gains the sudden interest of users of OSN in their posts after the occurrence of the events. For example, after the occurrence of an earthquake, the user's posts mentioning the word earthquake increase suddenly. Unlike the periodic events, such events do not offer any prior information concerning the time and the place of events; moreover, in many cases, they do not designate any predefined keywords or hashtags keywords also. Terrorist attacks and riots generally belong to such category of events.

Another essential category of events discussed on OSN is long-duration gradual events. Such kinds of events are specified and designated by the date, place, and other related information such as keywords and terminologies in advance. The discussions about a topic in a long-duration event persist for a long period. The users on OSN have prior knowledge of such kind of events. An example of such kind of events includes an election in any country. Usually, the date of an election is declared in advance. The users of OSN start discussing the election few days prior to the election (e.g., one

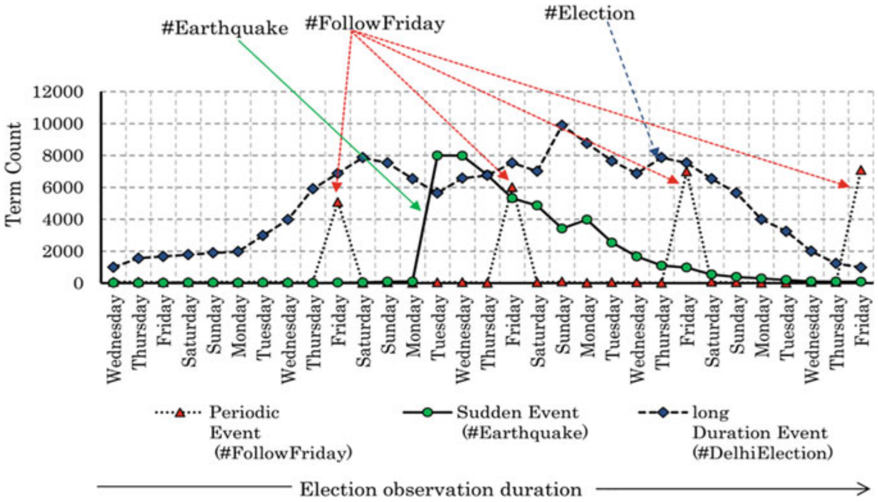

Fig. 3 Types of events on OSN

month) and persist it throughout the election campaign. Some major keywords for an election are comprised of contesting party names, contesting candidates' names, etc.

3 Framework for New Event Detection in OSN

A generic framework for new event detection is text data stream process which is a two-step process as depicted in Fig. 4. The first step is responsible for feature-based signal generation from the input text stream. The second step is responsible for detecting the burst in the signals.

(i) **Signal Generation**: The signal generation depends on features of the incoming text streams. Identification of the best features of the incoming text stream is very crucial for accurate event detection.

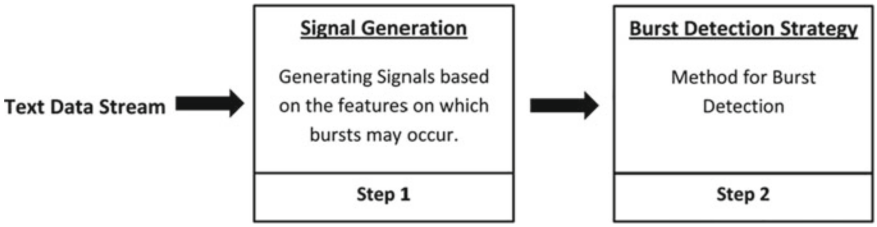
Text Data Stream
Signal Generation
Generating Signals based on the features on which bursts may occur.
Step 1
Burst Detection Strategy
Method for Burst Detection
Step 2

Fig. 4 Model for event detection in text data stream

(ii) **Burst Detection Strategy**: Any unprecedented change in the observed signal generated by signal generator the incoming text streams can be considered as a burst in the feature-based signals. Bursts in the generated signal are the potential indicators of the occurrence of events.

The rest parts of this section describe various strategies widely used for signal generation and burst detection in textual data stream.

3.1 Features for Signal Generation in Text Data Stream

1. **Frequency of words**: The most prominent burst detection methods in text data streams relied on the burst detection based on the examination of the frequencies of words present in the text stream [16–18]. The term frequency and inverse document frequency (TF-IDF) is the most popular method for generating term-frequency-based signal. The TF-IDF can be explained as follows: Let d be a document in a corpus of N documents and let t be a term in documents, then TF-IDF can be calculated by using (3).

   $$\mathrm{TF} - IDFt, d = 1 + \log TFt, d \times ID(t) \tag{3}$$

   where $\mathrm{TF}(t, d)$ is the term frequency of t in d, and $\mathrm{IDF}(t)$ is the inverse document frequency, i.e.,$N/n(t)$, where $n(t)$ is the number of documents containing the term t.

   In Twitter data streams, instead of calculating the TF-IDF, the *Term Frequency and Inverse Tweet Frequency* (TF-ITF) has been calculated for a given temporal window. For the real-time event detection, the TF-ITF is monitored periodically and any unprecedented changes in the TF-ITF are recorded as the bursts in the signal.
2. **Platform-Specific Features**: Online social networking sites like Twitter generally offer some specific notations to emphasize the topic of discussion in order to grasp the attention of other users. The notion of hashtags (#) is widely used by almost all OSN these days. Hashtags are generally created by users to indicate event or issue and floated over OSNs for drawing comments of other users over the events. Hashtags are the most prominent features for detecting events via OSNs. Instead of monitoring frequency of all words belonging to the text stream, the monitoring of the frequency of hashtags is more beneficial.
3. **Signal Generation based on online analysis of raw features**: There are certain signal generators which take raw features of the text data streams and process them online for generating the signals, for example, change in the sentiment score and change in domain of discussion.

   a. **Sentiment-Analysis-Based Signal Generation**: The sentiment analysis is defined as the automatic process of determining the sentiment of digitally stored textual documents [19–21]. While performing online sentiment analysis [18, 20, 22, 23] for a specified event, a significant change in the sentiment score with respect to time is a strong indicator of the occurrence of sub-event. Such significant changes can be utilized for tracking the occurrence of sub-events.

b. **Domain-Specific Features**: In text streams, the features represent the domain changes very frequently; for instance, the discussion of users on OSNs may change from the topic politics to sports after any crucial sport result. Observing the domain-specific features and generating the signal accordingly is also very prominent way for accurate event detection. However, such kind of signal generation required online domain classification. During the time of change in the domain of discussion, the underlying data distribution of the text stream also changes significantly.

3.2 Burst Detection in Text Data Stream

Burst detection is a process of detecting high-frequency period of in time series data analysis. The burst detection methods are utilized in variety of ways:

1. Fixed threshold value.
2. Minimum period between two bursts.
3. Duration of bursts.
4. Adaptive threshold parameter.

4 Conclusion

With the exponential growth of users on OSNs, the OSNs now resemble the corresponding real-world community very cohesively. The events occurred in the real world are often discussed on the virtual world of OSNs; hence, the analysis of contents of online OSNs provides immense possibility to track the real-world happening from the horizon of virtual world formed over social network. In keeping the view of the feasibility of OSNs in real-time tracking of the new events through OSNs, this paper discusses the general framework for real-world event detection through online social networking sites.

References

1. Web 2.0: January 2016, cited 2016. Available from: https://en.wikipedia.org/wiki/Web_2.0
2. Kaplan, A.M., Haenlein, M.: Users of the world, unite! The challenges and opportunities of social media. Bus. Horiz. **53**(1), 59–68 (2010)

3. Rosa, K.D., Ellen, J.: Text classification methodologies applied to micro-text in military chat. In: ICMLA'09. International Conference on Machine Learning and Applications. IEEE (2009)
4. Jackoway, A., Samet, H., Sankaranarayanan, J.: Identification of live news events using Twitter. In: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM (2011)
5. Srivastava, R., et al.: Analyzing Delhi assembly election 2015 using textual content of social network. In: Proceedings of the Sixth International Conference on Computer and Communication Technology. ACM (2015)
6. Sakaki, T., 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. ACM (2010)
7. Weimann, G.: New Terrorism and New Media. Wilson Center Common Labs (2014)
8. O'Connor, B., et al.: From tweets to polls: linking text sentiment to public opinion time series. ICWSM **11**(122–129), 1–2 (2010)
9. Yang, Y., Pierce, T., Carbonell, J.: A study of retrospective and on-line event detection. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (1998)
10. Allan, J.: Topic Detection and Tracking: Event-Based Information Organization, vol. 12. Springer Science & Business Media (2012)
11. Allan, J.: Introduction to topic detection and tracking. In: Topic Detection and Tracking, pp. 1–16. Springer, Berlin (2002)
12. AlSumait, L., Barbará, D., Domeniconi, C.: On-line LDA: adaptive topic models for mining text streams with applications to topic detection and tracking. In: Eighth IEEE International Conference on Data Mining. IEEE (2008)
13. Fiscus, J.G., Doddington, G.R.: Topic detection and tracking evaluation overview. In: Topic Detection and Tracking, pp 17–31. Springer, Berlin
14. Brants, T., Chen, F., Farahat, A.: A system for new event detection. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2003)
15. Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Comput. Intell. **31**(1), 132–164 (2015)
16. Cordeiro, M.: Twitter event detection: combining wavelet analysis and topic inference summarization. In: Doctoral Symposium on Informatics Engineering (2012)
17. Bahir, E., Peled, A.: Real-time major events monitoring and alert system through social networks. J. Conting. Crisis Manag. **23**(4), 210–220 (2015)
18. Cui, L., et al.: Topical event detection on Twitter. In: Australasian Database Conference. Springer, Berlin (2016)
19. Srivastava, R., Bhatia, M.: Ensemble methods for sentiment analysis of on-line micro-texts. In: International Conference on Recent Advances and Innovations in Engineering (ICRAIE). IEEE (2016)
20. Srivastava, R., Bhatia, M.: Challenges with sentiment analysis of on-line micro-texts. Int. J. Intell. Syst. Appl. **9**(7), 31 (2017)
21. Srivastava, R., et al.: Exploiting grammatical dependencies for fine-grained opinion mining. In International Conference on Computer and Communication Technology (ICCCT). IEEE (2010)
22. Srivastava, R., Bhatia, M.: Offline versus online sentiment analysis: issues with sentiment analysis of online micro-texts. Int. J. Inf. Retr. Res. (IJIRR) **7**(4), 1–18 (2017)
23. Srivastava, R., Bhatia, M.: Real-time unspecified major sub-events detection in the twitter data stream that cause the change in the sentiment score of the targeted event. Int. J. Inf. Technol. Web Eng. (IJITWE) **12**(4), 1–21 (2017)