International Journal of Engineering & Technology, 7 (2.15) (2018) 72-75

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Challenges of event detection from social media streams

Wafa Zubair Al-Dyani, Adnan Hussein Yahya, Farzana Kabir Ahmad

School of Computing, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia
**Email: wafazb1084@yahoo.com*

Abstract

The area of Event Detection (ED) has attracted researchers' attention over the last few years because of the wide use of social media. Many studies have examined the problem of ED in various social media platforms, like Twitter, Facebook, YouTube, etc. The ED task for social networks involves many issues, including the processing of huge volumes of data with a high level of noise, data collection and privacy issues, etc. Hence, this article discusses and presents the wide range of challenges encountered in the ED process from unstructured text data for the most popular Social Networks (SNs), such as Facebook and Twitter. The main goal is to aid the researchers to understand the main challenges and to discuss the future directions in the ED area.

Keywords*; Challenges; Event Detection; Facebook; Social Network; Twitter.*

1. Introduction

With the rapid advancement of technology, a large amount of information has generated in an exponential manner. This information comes from various sources, including traditional media (i.e., Radio, TV and newspaper), or modern sources, such as Social Networks (SNs) (i.e., Facebook, Twitter and YouTube). In addition, data from sources can be collected by using either their Application Programming Interfaces (APIs) or web-crawlers [1]. This motivates researchers to do data mining on SNs. Additionally, researchers have observed that a substantial percentage of conversations and responses on SNs are in textual format and generally related to recently merged “events” [2].

These events serve as a summary of the vast amount of information on social media [3]. There are various types of events, like natural disasters (e.g., floods, tsunami, volcanic eruption), political events (e.g., presidential elections), spread of diseases (e.g., Ebola, swine flu), death of celebrities (e.g., Michael Jackson), etc. However, the impacts of the same event may differ from one SN to another depending on the generated data volume on SNs [2]. Hence, many researchers have used a text mining approach on social media content in various applications such as sentiment analysis[4], study of social scientists [5], marketing trend detection [6] , etc. ED on the other hand, remained the most prominent and challenging task of all previous applications due to its social impact and the difficulties during implementation [1]. The rest of this article is organized as follows: Section II presents the various definitions for an event; while Section III describes the various categories of ED. Section IV presents and discusses the challenges and issues of ED for SNs. Finally, Section V presents the conclusions and future recommendations

2. Event definition

In general, an event is defined as “an occurrence at a specific time and place”[7]. However, there are other definitions from different perspectives. In the social media context, an event is not necessarily happening in a physical location. Therefore, it defines the event as “an occurrence causing change in the volume of text data that discusses the associated topic at a specific time”. This occurrence is characterized by topic and time, and often associated with entities, such as people and location [8]. On the other hand, Aggarwal, et al. [3] states that a news event is “something that happens at a specific time and place, but it is also an object of interest to the news media”. Similarly, McMinn et al. [9] define an event as something significant happening in a specific time and place beside it lead to discussions by the news media. This event might be a political event, natural disaster, terror attack or a protest, etc.

3. Event detection categories

Event Detection (ED) is classified into two categories depending on the type of its task; New Event Detection (NED) and Retrospective Event Detection (RED). NED focuses on detecting a newly occurred event from online text streams, while RED aims to discover the unknown events from the historical data in an offline approach. On the one hand, NED is characterized as an automated process which makes a binary decision as to whether an incoming document discusses a new event that has not been identified previously. Thus, this type of NED is considered as a very powerful system where new information needs to be extracted and analyzed from rapidly growing data, especially in some domains (e.g., natural disaster, stock markets, news analyses, intelligence gathering), with the goal to support decision-making. However, in practice, NED comprises of two important subtasks: Retrospective NED and Online NED. The former identifies previously undiscovered events from a collection of documents, while the latter focuses on online detection of new events from live stream texts. On the other hand, RED has been studied for a long time and many efforts have been undertaken to improve RED methods further in order to overcome the problem of high dimensionality data. In the following section, the various key challenges for ED in social streams are presented and discussed.

4. Event detection challenges

Social media streams report almost everything from daily life stories to latest local and global events. This rich and continuous flow of social media content may lead to the problem of information overload. In general, this information is characterized by huge data challenges, which include: volume (data size), velocity (the speed of change), noise and variety (different types of data) [1] . Consequently, mining such social stream information has become a more challenging task compared to the traditional text streams [10]. That is due to the dynamic characteristics of the social media as well as the existence of both text content and network structure within the streams [11] . Thus, the process of information filtering, analyzing data, and especially, detecting and monitoring the interesting events from social media text, has become the most difficult task [18]. In the following paragraphs, the existing challenges caused by SNs for ED are divided into two categories: General text mining challenges for ED from SNs; and specific challenges of the ED methodology.

4.1. General text mining challenges for ED from SNs

4.1.1. Volume and velocity issues

SN sites, such as Facebook and Twitter, contain a massive volume of User Generated Content (UGC). This content can reflect anything from daily activities to real-world events as they happened; sometimes, it may even precede the news channels in spreading the news about events. However, the large volume and variety of UGC on SNs give rise to the issue of extracting useful information out of it. Take Twitter for instance; Twitter produces over 340 million tweets everyday [14] through 500 million active users as reported in 2012 [15] . In contrast, a billion pieces of UGC are created by Facebook users on a daily basis [16] . Comparatively, recently, online news channels on social media streams (e.g., YouTube, Facebook, and Twitter), have gained great popularity as the easiest way to know, consume and understand the real-world events. Additionally, many electronic news reports write and publish daily on different SNs whereby they contain information about either different or similar event. As a result, all these causes the problem of information overload [34], whereby readers face common problems of comprehensively discovering significant events [31]. This motivates researchers to analyze social media news streams to get the most useful information out of it as well as to detect the significant events and their development along the timelines [34] . Thus, highly scalable and efficient methods are required in order to deal with a high and continuous data streams, especially for real-time ED [28] .

4.1.2. Volume and velocity issues

Another challenge that has been identified in SNs for ED, is the availability of public datasets. The conditions of social media companies restrict the use of collected data [1]. It is clearly observed that most studies have been done on Twitter data. This is because of the accessibility and usability of the Twitter's API. However, studies that depend on only a single platform face many risks, as it leads to the repetition of experiments and comparison between the approaches[1]. Thus, data collection is one of the obstacles that stand in the way of the ED task for some social media sources, such as Facebook. Facebook poses several challenges regarding data collection, such as a limited access (only public data) through its graphical API due to its privacy policies and the authorization process. Berger et al.[29] stated that the authorization policy makes it difficult for automatic data crawling; besides that, the code for crawling must be updated frequently as the API changes over time [30]. Additionally, Facebook's API does not support the process of receiving posts in real-time form (online data collection). However, UGC on Facebook could give valuable insights as it is the largest SN[16]. Given all above difficulties, the number of studies relating to ED for Facebook is relatively scarce compared to other social media platforms, such as Twitter, which takes the largest share of studies in this area because of the ease of data collection, especially in real-time form [11].

4.1.3. Volume and velocity issues

ED from SNs poses new challenges that vary from traditional media. Compared to the structured, well-written and edited news, UGC on SNs can be written by anyone. Thus, it may include irregular, abbreviated and informal terms [32]. Besides that, users may write using mixed languages, improper sentence structure and slang. In addition, their writing may contain many grammatical and linguistic errors or misspelling [39]. In addition, UGC on SNs has a limited length (e.g., only 140 characters for a tweet). Therefore, ideas are represented briefly with an insufficient amount of information, hence creating additional challenges to the traditional text analyzing methods [34]. As an illustration, even though all news articles describing events usually answer the major questions about what, when, where and who, this information, however, remains hidden within the text and requires the reader to manually extract it by reading the article [36]. This is not feasible for tweets/posts due to their restricted length which prevents the provision of all necessary information about an event. According to [1] , in spite of attention to the efforts in this area, it is observed that no method has addressed and answered all these questions (e.g., what, when, where, who) clearly, and thus, it is still a challenging task.

4.2. ED methodology challenges

In SNs, ED is considered as a complex problem, especially when the evaluation process of the suggested approaches of ED is raised. To emphasize this point, there are two major approaches to ED, i.e., Document Pivot and Feature Pivot approaches [1]. On the one hand, for the document pivot approach, clustering techniques are used to organize documents based on the similarity measurement and related documents are identified using direct comparison. Additionally, this approach has been used mainly for Topic Detection and Tracking (TDT) challenges, but they often concentrate on detecting just frequent events through mining either co-occurrence term-relations [37] or semantic information [38]. However, they ignore the idea of combing the two kinds of relations, which can help in detecting incomplete information. Moreover, this approach cannot discover the hidden co-occurrence relations from the noisy data collection by using either the bridge terms or the context prevents the significant rare events from being discovered [25]. Nevertheless, TDT is not applicable for SNs, such as Facebook and Twitter. This is because not all tweets/posts are related to an event and it cannot handle a large volume of data [1] . Moreover, TDT can only organize news stories as events in a flat hierarchical structure without illustrating how these events evolve within a topic [34]. Still, this does not satisfy the news readers nowadays as they are not just interested in detecting the significant events, but also in how these events have evolved along the timeline [56]. However, fulfilling this requirement has proven to be a very challenging task, especially for high-level rate stream[18]. On the other hand, the Feature Pivot approach depends on detecting burst features from SN text streams and focuses on the variation of detected features [1]. Notably, almost all TDT techniques which have been applied in SNs are Feature Pivot algorithms [1]. From another point of view, the popular methods for ED are categorized into either supervised classification or unsupervised clustering methods [43]. Supervised classification methods have obtained good results, but they are time-consuming and require a large amount of data to be trained, and required a lot of human effort [45]. On the other hand, despite the popularity of clustering techniques for ED and the fact that they do not require labeled data, it is still a challenging task to build an automated unsupervised method which can deal with high dimensionality of news stream data without human effort and cost [46]. In practice, various ED approaches are available in the literature which affects the quality of results [12]. However, these approaches have suffered from the problem of high dimensionality data due to the limitations of proposed techniques in each stage of ED, such as data processing [12], data representation, feature selection, data categorization and evaluation [45].

4.2.1 Pre-processing and representation of challenges

Social media streams are filled with noise [39] (i.e., advertisements, spam messages, hoaxes, URL, etc.). In practice, not all UGC contains useful information; in fact, a substantial amount contains meaningless contents that are not relevant to real-world events [46]. Consequently, this generates a lot of noise, and eventually, affects negatively on ED accuracy and performance. Thus, this is a new challenge for ED from SNs whereby identifying tweets/posts that describe truly real-world events from among the polluted contents, has become a necessary step [39]. In addition, data pre-processing has a positive effect on the quality of detected events as it is observed that most tweets/posts usually contain noisy components (e.g., stop words, URL) which may affect the performance of the ED process [2]. On the other hand, data representation has diverse techniques, such as Bag of Words (BOW) and Term Vector (TV). However, both techniques have proven to be not useful for ED as BOW makes the distinguishing task between the events within the same topic a difficult process, while TV increases the computational cost [62] .

4.2.2 Feature extraction challenges

SN streams consist of a huge number of features [26]. Hence, extracting the correct set of features is a very crucial and challenging task for the ED process [3], as these key features are very vital for representing the events and capturing the most important information. In addition, the critical factor is the dependency between these selected features since more than one event may be represented by an identical set of features leading to ambiguity [48]. Moreover, these features are used to differentiate between the events within the same topic, since the variation between these events may be relatively minor [49]. In practice, these features might be either content-based features (e.g., TF-IDF scores, emoticons, number of tags) or non-textual features, called meta-data (e.g., number of comments or friends (Facebook), or number of followers (Twitter)). This additional data could be used to extract other characteristics of the detected events [50]. For example, if most tweets/posts related to a specific event have identical geographical information, one can conclude that the event may originate at that location. Furthermore, other meta-data can be integrated to support the ranking of events (e.g., number of retweets, number of comments or number of shares). To clarify, the authors [51] compared between these two types. They find that applying both textual and non-textual features can produce significantly precise results. In addition, in social media news streams, reading a single article from one news channel may give a biased and incomplete picture of a specific event [36] . Thus, it would be good to find relevant articles from other news publishers. This process also poses a problem since it mostly depends on searching related articles using one or more relevant key features [36]. Thus, selecting the accurate features and utilizing them in either supervised or unsupervised approaches would enhance the performance of the ED process [48].

4.2.3 Categorization challenges

The problem of ED in social media streams is relatively similar to clustering problem, in that most approaches handle the ED task, at least in the first stage, as a text clustering process [1] , where the recognized clusters are organized into either “non-event clusters” or “event-clusters” [1]. However, data clustering for social media streams faces the main problem of “combinatorial explosion”, which is an issue that frequently occurs when dealing with large volumes of data sets [44]. In addition, for NED task, where it does not require any prior knowledge about the event, methods which solely rely on predefined queries are not suitable. Similarly, partitioning clustering algorithms (i.e., k-mean, k-median and k-medoid, etc.) are not appropriate as they require prior knowledge about the number of clusters (k)[28]. In spite of all the above issues for clustering methods, k-mean is still considered as a very well-known and popular algorithm which has been used for most ED studies [41]. However, k-mean has a fundamental drawback, i.e., falling into local optima, and hence, producing weak results [52]. In order to overcome this limitation, several options are applicable for potential optimization. For example, using advanced techniques for setting the threshold of incremental clustering algorithms for NED task. Another option is applying the heuristic approach by searching for global optima with the aid of an optimization algorithm [52]. In practice, combing existing data mining algorithms with bio-inspired algorithms to create a hybrid method is still in its early stage for ED [52].
To clarify further, various algorithms have been used for the ED task, but they suffer from some kind of limitation. For example, Latent Dirichlet Allocations (LDA) and Nearest Neighbor algorithms (i.e., KD-trees and Indexing Trees) require expensive computational procedures and are not easy to apply for fast and large volumes of SN streams. Conversely, Locality Sensitive Hashing (LSH) hashes similar documents into the same bucket. However, it is a randomized approach and errors may occur. Thus, multiple LSH has been applied recently to reduce the error rate, but this has led to an increase in implementation time[53].

4.2.4 Evaluation challenges

For critical applications (e.g., emergency events), the events should be detected as soon as possible to support decision- making. Thus, the methods and techniques used for these kinds of events should be evaluated in terms of how fast they can be identified rather than just evaluating based on precision and recall measurements[1]. Unfortunately, there are very few ED evaluation datasets [9]. The TDT5 dataset has been utilized by many studies [43], to evaluate precision. However, these datasets are quite different in their nature and the results obtained from them could significantly differ from the results obtained from Facebook and Twitter, where TDT5 originates from newswire documents and includes high well-formed texts which are not suitable for the informally written tweets/posts on SNs. Therefore, most researchers have built their own corpora which are manually annotated [44,45], with a fewer number of events. Hence, the results are often subject to bias and lead to the repeatability of experiments and comparison between methods [1].

5. Conclusion and future recommendations

This article highlights the main challenges of ED in the most popular and largest SN streams: Facebook and Twitter. The various definitions of events suggest that there are many domains where ED has been implemented so far. ED stands out because of its complexity and social impact. Despite making progress in the ED process, analyzing and monitoring the events from various SN platforms remain a challenging task where no standard ED approach has yet to be recognized. More extensive effort is required to obtain an effective and efficient ED model. This can be achieved, for instance, by enhancing feature extraction and query generation techniques as well as improving the detection algorithms. In addition, the methods can be expanded to combine and analyze information from multiple SN resources and languages. Finally, the summarization can be improved as well as new visualization tools which can help the readers to understand and get a more comprehensive picture of significant events.

Acknowledgement

This work was supported in part by National Grant under MOHE.

References

[1] N. Panagiotou, I. Katakis, and D. Gunopulos, “Detecting events in online social networks: Definitions, trends and challenges,”, vol. 9580, Springer, 2016, pp. 42–84.

[2] S. B. Kaleel, M. AlMeshary, and A. Abhari, “Event detection and trending in multiple social networking sites,”, 2013, p. 5.

[3] W. Dou, X. Wang, W. Ribarsky, and M. Zhou, “Event detection in social media data,” in IEEE, 2012, pp. 971–980.

[4] A. Bifet and E. Frank, “Sentiment knowledge discovery in twitter streaming data,”, 2010, pp. 1–15.

[5] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, “Outtweeting the twitterers-predicting information cascades in microblogs.,” WOSN, vol. 10, pp. 3–11, 2010.

[6] M. Mathioudakis and N. Koudas, “Twittermonitor: trend detection over the twitter stream,”, 2010, pp. 1155–1158.

[7] C. C. Yang, X. Shi, and C.-P. Wei, “Discovering event evolution graphs from news corpora,”, vol. 39, no. 4, pp. 850–863, 2009.

[8] W. Dou, X. Wang, D. Skau, W. Ribarsky, and M. X. Zhou, “Leadline: Interactive visual analysis of text data through event identification and exploration,”, 2012, pp. 93–102.

[9] A. J. McMinn, Y. Moshfeghi, and J. M. Jose, “Building a large-scale corpus for evaluating event detection on twitter,” , 2013, pp. 409–418.

[10] H. Zhou, H. Yu, R. Hu, and J. Hu, “A survey on trends of cross-media topic evolution map,” Knowledge-Based Syst., 2017.

[11] H. Sayyadi, M. Hurst, and A. Maykov, “Event detection and tracking in social streams.,” in Icwsm, 2009.

[12] S. D. Tembhurnikar and N. N. Patil, “Topic detection using BNgram method and sentiment analysis on twitter dataset,” in 2015 4th Trends and Future Directions, ICRITO 2015, 2015, pp. 1–6.

[13] H.-P. Chen, K.-W. Hsu, and S.-I. Chiu, “Event Detection in an Ego Network on Facebook,” 2016.

[14] T. Wasserman, “Twitter says it has 140 million users,”, 2012.

[15] T. Team, “Twitter turns six,” Twitter Blog, pp. 21–23, 2012.

[16] I. P. Cvijikj and F. Michahelles, “Monitoring trends on facebook,”, 2011, pp. 895–902.

[17] S. Khatdeo, S. Shrawane, P. Kumbhare, and P. M. S. Nimbarte, “Detection and Visualization of Events from Online News,” vol. 1, no. 21, pp. 241–244, 2017.

[18] D. Huang, S. Hu, Y. Cai, and H. Min, “Discovering event evolution graphs based on news articles relationships,” in 2014, pp. 246–251.

[19] C. Wu, B. Wu, and B. Wang, “Event Evolution Model Based on Random Walk Model with Hot Topic Extraction,”, 2016, pp. 591–603.

[23] G. Leban, B. Fortuna, and M. Grobelnik, “Using News Articles for Real-time Cross-Lingual Event Detection and Filtering.,” in 2016, pp. 33–38.

[24] C. Zhang, H. Wang, W. Wang, C. Ma, J. Li, Y. Wang, and F. Xu, “EventPanorama: A Framework for Event Detection and Visualization from Online News,” in HICSS, 2016, pp. 3739–3748.

[25] T. M. Beigh, S. Upadhyaya, and G. Gopal, “Event Identification in Social News Streams Using Keyword Analysis,” Int. Res. J. Eng. Technol., vol. 3, no. 5, 2016.

[26] Z. Lu, W. Yu, R. Zhang, J. Li, and H. Wei, “Discovering Event Evolution Chain in Microblog,” in HPCC, 2015, pp. 635–640.

[27] F. Atefeh and W. Khreich, “A survey of techniques for event detection in Twitter,” Comput. Intell., vol. 31, no. 1, pp. 133–164, 2015.

[28] P. Berger, P. Hennig, T. Klingbeil, M. Kohnen, S. Pade, and C. Meinel, “Mining the Boundaries of Social Networks: Crawling Facebook and Twitter for BlogIntelligence,” in IKE, 2013, p. 1.

[29] L. C. Passaro, A. Bondielli, and A. Lenci, “FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis,” in CLiC-it 2016., 2016.

[30] D. Richter, P. D. D. K. Riemer, and J. vom Brocke, “Internet social networking,”, vol. 53, no. 2, pp. 89–103, 2011.

[31] J. Deng, F. Qiao, H. Li, X. Zhang, and H. Wang, “An Overview of Event Extraction from Twitter,”, 2015, pp. 251–256.

[32] R. Parikh, “ET : Events from Tweets,”, pp. 613–620, 2013.

[33] W. Simm, M.-A. Ferrario, S. Piao, J. Whittle, and P. Rayson, “Classification of short text comments by sentiment and actionability for voiceyourview,”, 2010, pp. 552–557.

[34] D. Metzler, S. Dumais, and C. Meek, “Similarity measures for short segments of text,”, 2007, pp. 16–27.

[35] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik, “Event Registry – Learning About World Events From News,”, 2014, pp. 107–110.

[36] H. Sayyadi and L. Raschid, “A graph analytical approach for topic detection,” ACM Trans. Internet Technol., vol. 13, no. 2, p. 4, 2013.

[37] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” J. Mach. Learn. Res., vol. 3, no. Jan, pp. 993–1022, 2003.

[38] R. Ahirrao and S. Patel, “An Overview on Event Evolution Technique,” Int. J. Comput. Appl., vol. 77, no. 10, pp. 7–11, 2013.

[39] X. Li, Y. Zheng, and Y. Dong, “Discovering evolution of complex event based on correlations between events,” in Proceedings - 11th WISA 2014, 2014, pp. 47–50.

[40] F. Hogenboom, F. Frasincar, U. Kaymak, F. De Jong, and E. Caron, “A Survey of event extraction methods from text for decision support systems,” Decis. Support Syst., vol. 85, pp. 12–22, 2016.

[41] F. Hogenboom, F. Frasincar, U. Kaymak, and F. De Jong, “An Overview of Event Extraction from Text,” in CyberC, on, 2015, pp. 251–256.

[42] G. Li, “A clustering based approach on sentiment analysis,” in 2010, pp. 331–337.

[43] A. Mohamad, S. Syed Mustapha, and M. Razali, “Automatic Event Detection on Reuters News,” 2010.

[44] Q. H. Ramadan and M. Mohd, “A review of retrospective news event detection,” in STAIR, International Conference on, 2011, pp. 209–214.

[45] H. Becker, M. Naaman, and L. Gravano, “Selecting Quality Twitter Content for Events.,” ICWSM, vol. 11, 2011.

[46] Y. Xi, B. Li, and Y. Liu, “A Semantic Aspect-Based Vector Space Model to Identify the Event Evolution Relationship within Topics,” J. Comput. Sci. Eng., vol. 9, no. 2, pp. 73–82, 2015.

[47] S. Lavanya, R. Kavipriya, Y. Yang, J. Q. Carbonell, R. D. Brown, B. Archibald, and X. Liu, “A Survey on Event Detection in News Streams,” vol. 2, no. 5, pp. 33–35, 2014.

[48] A. Kumar, R. Khorwal, and S. Chaudhary, “A Survey on Sentiment Analysis using Swarm Intelligence,” Indian J. Sci. Technol., vol. 9, no. 39, 2016.

[49] A. Weiler, M. Grossniklaus, and M. H. Scholl, “Event identification and tracking in social media streaming data,” in EDBT/ICDT, 2014, pp. 282–287.

[50] H. Becker, D. Iter, M. Naaman, and L. Gravano, “Identifying content for planned events across social media sites,” in 2012, pp. 533–542.

[51] R. Tang, S. Fong, and S. Deb, “Integrating Nature-inspired Optimization Algorithms to K-means Clustering,” in 2012, pp. 116–123.

[52] S. Petrović, M. Osborne, and V. Lavrenko, “Streaming first story detection with application to twitter,” in, 2010, pp. 181–189.

[53] M. Karkali, F. Rousseau, A. Ntoulas, and M. Vazirgiannis, “Efficient Online Novelty Detection in News Streams.,” in WISE (1), 2013, pp. 57–71.

[54] H. Becker, M. Naaman, and L. Gravano, “Beyond trending topics: Real-world event identification on Twitter,” Icwsm, vol. 11, no. 2011, pp. 1–17, 2011.