Online Event Detection in Social Media with Bursty Event Recognition

Wanlun Ma(✉), Zhuo Liu(✉), and Xiangyu Hu(✉)

School of Information and Communication Engineering,
University of Electronic Science and Technology of China,
Chengdu 611731, People's Republic of China
mawanlun0@gmail.com, 201721010807@std.uestc.edu.cn, alxe24@163.com

Abstract. The emergence of social media opens tremendous research opportunities. Many individuals, mostly teens and young adults around the world, share their daily lives and opinions about a wide variety of topics (e.g., crime, sports, and politics) on social media sites. Thus, social media becomes a valuable repository for data of different types, which could provide insights of social events happening around the world. However, it is still a challenge to identify the bursty and disruptive events from the massive and noisy user-generated content on social media sites. In this paper, we present a novel event detection framework for identifying surrounding real-world events that can support decision making and emergency management. Our proposed framework consists of four main components, including data pre-processing, event-related tweets classifying, online clustering, and bursty event recognition. We conducted a series of experiments on the real-world social media dataset collected from Twitter. The experimental results demonstrated the effectiveness of our proposed method.

Keywords: Event detection · Social media · Text mining · Information extraction

1 Introduction

Online Social Networks (OSNs), such as Twitter and Facebook, have become an integral part of individuals' daily life. Due to the fast dissemination nature of information on OSNs, they are now served as the major news consumption tool for users. In a survey recently conducted by Pew Research Centre (PRC), two-thirds of Americans get at least some of their news on social media with one-in-five doing so frequently and about three-in-four Twitter users get news on the site [17]. Social network users share their views and broadcast news and information about ongoing events. Online social media platforms serve as real-time "sensors" for social trends and incidents [21], which is very useful for supporting decision making and public management. Some research has shown the importance of social media in disaster warning system design and emergency management. For

W. Meng and S. Furnell (Eds.): SocialSec 2019, CCIS 1095, pp. 181–190, 2019.
https://doi.org/10.1007/978-981-15-0758-8_14

instance, Palen [15] stated that social media was a significant and accurate tool for the crisis event management during the Virginia Tech shootings and the Southern California wildfire.

Although many researchers have proposed models and techniques for the purpose of detecting events from social media contents, the existing approaches suffer from several key challenges. Firstly, it is hard for real-time event detection due to the large number of social media records (e.g., tweets) and the continuous appearance of new events. In addition, user-generated content in social media consists of incomplete and even wrongly structured sentences due to abbreviations, irregular expressions, abnormal words, and slang terms. Secondly, the event-related tweets classification can filter most noisy and irrelevant content, but the effectiveness of classifying the event-related posts is limited due to the performance of the classifier. Moreover, it is difficult to conduct further analysis of the ongoing events from the massive event clusters generated by online clustering.

To overcome these issues, we propose a novel event detection framework which can recognize the bursty and disruptive events from social media. Our contributions can be summarised as follows:

- We proposed a novel event detection framework which can identify surrounding bursty events, such as terrorist attacks. Our proposed framework consists of four main components, including data pre-processing, event-related tweets classifying, online clustering and bursty event recognition.
- In the bursty event recognition module, we employed the temporal bursty feature and the news value of events to distinguish the significant and bursty events from the event clusters.
- In the evaluation, we conducted extensive experiments on the real-world social media dataset, and the experimental results demonstrated the effectiveness of our proposed method.

2 Related Work

There are many research works that focus on event detection and tracking [1, 6,7], topic discovery and evolution [8,11,20], and information summarization [16]. These works extract various types of events from social media, such as Arab Spring uprisings [1], terrorist attacks [6], sports games [16,20], and disease outbreak [9].

Becker et al. [3] proposed an online clustering and filtering framework to identify events with different types. This approach extracted temporal, social, topical, and twitter centric features of tweets clusters and then classified these clusters into real-world event clusters and non-event clusters. Moreover, Alsaedi et al. [1] implemented a similar clustering method using three sets of features (temporal, spatial and textual features) to identify whether a message cluster belonging to the group of real-world events or not. In the literature, Latent Dirichlet Allocation (LDA) [4] and non-negative matrix factorization (NMF) [13] are two widely used topic models for event detection. Xing et al. [20] extended the

LDA method to model the relationship between hashtags and topics of tweets, and then discovered events based on event-related hashtags. Both Kalyanam [11] and Chen [8] employed the NMF method to model the evolution of topics in social media. Similarly, Shin et al. [18] proposed an NMF-based approach using both spatial and temporal information from tweets to detect anomalous events. Besides, Chen et al. [7] proposed a clustering-based approach using a similarity metric and low dimensional representations of events which were learned from a neural network with an attention mechanism. While Liu et al. [14] had a similar idea that they exploited a recurrent neural network with a cross-lingual attention gate to identify events from multiple languages. In addition to the text source, Schinas et al. [16] proposed a multimodal clustering method using not only the textual information but also the image content from social media.

In general, most existing research firstly models bursty frequency patterns along with time or space, and then extract events using classification or clustering methods. However, most existing approaches showed an unsatisfactory performance of recognizing bursty and important events from the massive extracted events.

3 Online Event Detection

In this section, we describe the framework of our event detection and tracking method, which consists of four components - data pre-processing, event-related tweets classifying, online clustering and bursty event recognition. Figure 1 provides an overview of our proposed event detection framework. First of all, the data pre-processing component applied several text pre-processing techniques to clean the tweets. Then, the event-related tweets classification component identifies event-related tweets from noisy and irrelevant posts. After filtering non-event tweets, the online clustering component groups similar event-related tweets into the same event cluster. Then, the bursty event recognition component distinguishes the significant and bursty events from the massive event clusters.

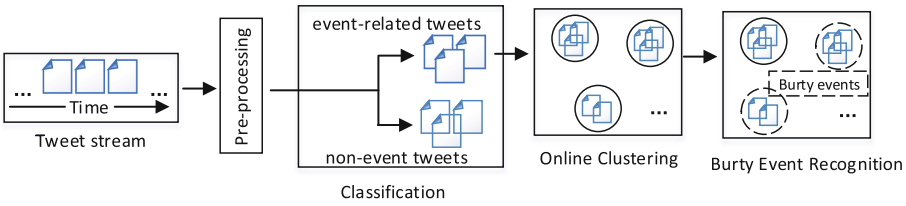

Fig. 1. Event detection framework for twitter stream

3.1 Data Pre-processing

Tweets usually consist of incomplete and even wrongly structured sentences due to abbreviations, irregular expressions, abnormal words, and slang terms. Therefore, we applied some pre-processing techniques to clean the tweets and consequently improve their quality for the subsequent event detection analysis.

In addition to the traditional text pre-processing techniques (e.g., stop words removal), we performed word segmentation and parts-of-speech (POS) tagging for each tweet and then employed name entity recognition (NER) to extract various kinds of name entities, including mentioned users, organizations, locations. An example of the data pre-processing is shown in Fig. 2.

raw tweet		#Trump says second summit with #KimJongUn will happen after the midterms https://t.co/OtXi13wYXA						
after preprocessing	words	Trump	says	second	summit	KimJongUn	happen	midterms
	NER	PER	-	-	-	PER	-	-
	POS	NOUN	VERB	ADJ	NOUN	NOUN	VERB	NOUN

Fig. 2. An example of the data pre-processing. The results of NER denotes names for a certain specific person (PER), location(LOC), or organization (ORG). And each word has one of the ten following POS tags: verbs (VERB), nouns (NOUN), pronouns (PRON), adjectives (ADJ), adverbs (ADV), adpositions (ADP), conjunctions (CONJ), cardinal numbers (NUM), particles or other function words (PRT), and others (X).

3.2 Event-Related Tweets Classifying

As OSNs contain various types of content posted by users, such as personal updates, random thoughts and opinions, and information sharing, it is necessary to separate "event" and "non-event" content on OSNs. Generally, the classification aims to identify event-related tweets from noisy and irrelevant posts and eliminate non-event tweets. Since the following online clustering component only considers the event-related tweets, filtering non-event tweets consequently reduces the number of tweets to be processed in subsequent steps. In the classification, we applied a well-known supervised machine learning algorithms - random forest [5] - for filtering non-event tweets. Features are critical in a machine learning based classification task. Therefore, based on the Sriram's work [19], we chose nine features to represent the tweets, including the number of words, the ratio of capitalized words, the ratio of hashtag words. Table 1 lists all the features used in the task.

We used a manually sampled dataset which was composed of 1737 event-related tweets and 1900 non-event tweets to train the Random Forest classifier. Then, this pre-trained classifier can be used to identify the "event" and "non-event" posts in the streaming tweets.

Table 1. 9 features used in the classification

Features	Examples
The number of words	The count of words in a tweet
The ratio of capitalized words	BBC, CNN, NEWS
The number of name entities	London, Troy, FBI
The ratio of hashtag words	#Brexit, #Tradewar
The ration of mentioned users	@ReutersPolitics, @TheEconomist
The ratio of non-English words*	Trade, summit
The ratio of opinion words	Think, tell, believe, deem
The ratio of personal sentiment words	God, thanks, hell, stupid
The number of question expression word	Where, why, what

*The example words are translated from the originals.

3.3 Online Clustering

The event-related tweets classification separates the event-related tweets from the noisy and messy tweet stream. In order to identify the topic of an event, we applied an unsupervised online clustering approach [2] to group similar event-related tweets into the same event cluster.

Each tweet is represented as a vector T whose values are weighted based on the POS of each word and the results of NER. Based on the empirical study, name entities (e.g., locations, organizations and celebrities) are weighted as 1.2, noun and verb are weighted as 1, and other words are weighted as 0.5. The similarity function used to measure the similarity between the tweet and the existing clusters $(C_1, C_2, ..., C_n)$ is defined as:

$$similarity = \frac{T \cdot C_i}{|T| \cdot |C_i|} \tag{1}$$

Whether a tweet T belongs to an existing cluster C_i or not is determined by the threshold parameters τ. Different from the work [2], we dynamically tune the threshold τ based on the number of words in a tweet. In the empirical study, τ is set as 0.4, 0.45, 0.55 respectively, when the length of the tweet is greater than 13, 6 or 0.

3.4 Bursty Event Recognition

The event-related tweets classification can filter the most noisy and irrelevant content, but the effectiveness of classifying the event-related posts is limited depending on the performance of the classifier. Moreover, it is difficult to conduct a further fine-grained analysis of the ongoing events from the massive event clusters generated by online clustering. In order to address these issues, we applied bursty event recognition to the generated event clusters to distinguish the significant and bursty events for further analysis.

In the bursty event recognition module, we employed the temporal bursty feature and the news value of events to recognize the bursty events from the event clusters. In each event cluster C_i, there are n tweets $(T_1, T_2, ..., T_n)$ which are sorted by the post time. In other words, let PT_k represents the post time of T_k, then PT_j is earlier than PT_k when j is smaller than k. Thus, the temporal bursty feature (TBF) for an event cluster C_i is defined as:

$$TBF = min\,|PT_k - PT_{k+w}| \tag{2}$$

where w is the time window. Intuitively, social media users are likely to update many posts in a short time to share information with others when a bursty event (e.g., terrorist attacks) is happening around them. Thus, the temporal bursty features (TBFs) of bursty events are supposed to be smaller than that of general events. Therefore, the temporal bursty feature (TBF) of an event under the threshold τ is identified as a bursty event.

Furthermore, we applied the news value of events to recognize the bursty events from the event clusters. In order to assess the news value of events, we use three indicators, including the max number of retweets, the number of keywords (e.g., breaking news, breaking, and news post) and the number of name entities about locations (LOC). Table 2 shows the three indicators used to assess the news value of events. The news value of an event cluster is set to 1 when two of the indicators are over the thresholds respectively; otherwise, it is set to 0. In general, an event cluster is identified as a bursty event when TBF is under the threshold τ and the news value is equal to 1.

Table 2. News value indicators

Indicators	Examples
The max number of retweets	The max number of retweets for a tweet in the same cluster
The number of keywords words	Breaking, #news, #breaking, #breakingnews
The number of name entities about locations (LOC)	Manchester, London bridge

4 Experimental Evaluation

4.1 Dataset and Experiment Setting

The performance of the proposed model was evaluated on real-world data collected using Twitter public API. The raw data was collected during 2017 U.K. General Election from May 19th to June 6th, considering two severe terrorist attacks (i.e., the Manchester Arena bombing and 2017 London Bridge attack)

Table 3. The statistics of the real-world datasets

Country	Time period	#Tweets (million)	#Events
UK	05/19/2017–06/06/2017	11.36	120

happened during this time period. We also collected 120 events from May 19th to June 6th through several news websites and evaluated the proposed model based on these background events. The detailed statistics of the datasets are listed in Table 3.

All the experiments were implemented by Python and conducted on a computer running Windows 10 with a memory of 16 G and a processor of 3.0 GHz Intel Core i5.

4.2 Baseline Methods

To validate the effectiveness of the proposed framework, we compare our approach with the following methods:

1. *BHS* [12]: BHS models the tweet stream by an infinite-state automaton, in which the burst appearance of a topic or event is viewed as a state transition process. We set the threshold as 3 for detecting the high-intensity state of the bursty events.
2. *AED* [1]: AED is a detection framework for identifying disruptive events. AED applies a Naive Bayes model for event-related tweets classification and a similar online clustering method to detect disruptive events.

4.3 Evaluation Metrics

In order to evaluate the performance of our proposed method and the baselines, we employed the same evaluation method as in [7,10]. Detected events are judged by two humans whether the events are meaningful and important. Since it is a time-consuming task, only 40 events are judged in the work [7]. Thus we only analyzed 100 events in our experiments. The performance of our approach and baseline methods are evaluated in terms of precision, recall, and F-measure.

4.4 Model Comparisons and Results

We evaluated the proposed approach by comparing with the baseline methods in Sect. 4.2. The experimental results on the real-world dataset are shown in Table 4. We can observe that the performance of our approach is better than the baseline methods in terms of precision, recall, and F-measure. Specifically, the F-measure of our method is improved by 16% and 11% respectively, compared to BHS and AED. BHS considers the bursty and hierarchical structure of the stream tweet data and models the arrival times (i.e., post times) of tweets in an event cluster to identify bursts that have high intensity. Therefore, a non-bursty

Table 4. The results of different methods

Method	Precision	Recall	F-measure
BHS	0.5670	0.7333	0.6395
AED	0.6176	0.7590	0.6810
Our method	**0.7735**	**0.8283**	**0.7999**

Table 5. Sample events detected by our method

Events	Sample tweets
Tory candidate Craig Mackinlay charged over election expenses	Tory election expenses explained: Craig Mackinlay charged https://t.co/sDAO810uBb via @cmackinlay @Channel4News #GE2017
	Craig Mackinlay (Con, South Thanet) has been charged with election offences for the 2015 election campaign in the South Thanet constituency
Election campaigning suspended after London Bridge attack	Safety over freedom: Social media users call for General Election to be suspended following the London Bridge attack https://t.co/ATL0VSJykG
	UK Govt considering indefinitely cancelling election following terror attacks in London Sat. which killed at least 7

event cluster will be wrongly identified as a bursty one if the post timings of several tweets are very close. As a result, BHS has the lowest performance. The main difference between our method and AED is that we incorporated the bursty event recognition for the generated event clusters. This result demonstrates the importance of employing the bursty event recognition component in the proposed framework. To further show the effectiveness of our method, some examples from our detected events are shown in Table 5.

As discussed in Sect. 3.4, the time threshold of the temporal bursty feature (TBF) highly affects the performance of our proposed framework. In other words, an event cluster will be identified as a bursty event under the condition that its TBF is less than the threshold τ. Therefore, we evaluated the performance of our framework under different TBF thresholds. As shown in Fig. 3, the precision decreases with the increase of the threshold τ, in contrast, recall, and F-measure increase. Besides, the result demonstrates that the F-measure almost remains stable when the threshold τ is over 3 min. Thus, we set the threshold τ as 3 in practice.

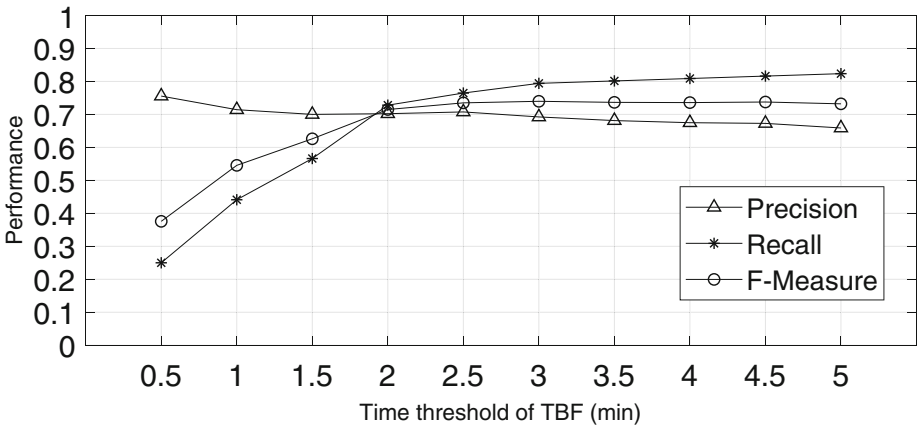

Fig. 3. The performance of our framework under different time thresholds of TBF.

5 Conclusion

In this work, we study the problem of event detection from the massive and noisy content in social media. The major challenge of event detection stems from bursty event recognition on which existing methods showed unsatisfactory performance. In order to tackle this problem, we proposed a novel event detection framework which can identify surrounding bursty events, such as terrorist attacks. Particularly, in the bursty event recognition module, we employed the temporal bursty feature and the news value of events to distinguish the significant and bursty events from the event clusters. Experimental results show that our proposed event detection method achieves significant improvement over existing approaches.

For future work and improvements, there are two main directions including (i) more appropriate semantic representation methods for tweets, such as Word2vec model, (ii) utilizing external information of tweets in the clustering process, like tweets' locations and URLs, (iii) further investigating the temporal and spatial characteristics of bursty events and designing more effective and accurate features to identify bursty events.

References

1. Alsaedi, N., Burnap, P.: Arabic event detection in social media. In: Gelbukh, A. (ed.) CICLing 2015. LNCS, vol. 9041, pp. 384–401. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18111-0_29
2. Alsaedi, N., Burnap, P., Rana, O.: Can we predict a riot? Disruptive event detection using twitter. ACM Trans. Internet Technol. (TOIT) **17**(2), 18 (2017)
3. Becker, H., Naaman, M., Gravano, L.: Beyond trending topics: real-world event identification on twitter. In: ICWSM, vol. 11, pp. 438–441 (2011)

4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
5. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
6. Burnap, P., Williams, M.L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., Voss, A.: Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack. Soc. Netw. Anal. Min. **4**(1), 206 (2014)
7. Chen, G., Kong, Q., Mao, W.: Online event detection and tracking in social media based on neural similarity metric learning. In: IEEE International Conference on Intelligence and Security Informatics, pp. 182–184 (2017)
8. Chen, Y., Zhang, H., Wu, J., Wang, X., Liu, R., Lin, M.: Modeling emerging, evolving and fading topics using dynamic soft orthogonal NMF with sparse representation. In: 2015 IEEE International Conference on Data Mining (ICDM), pp. 61–70. IEEE (2015)
9. Ghenai, A., Mejova, Y.: Catching Zika fever: application of crowdsourcing and machine learning for tracking health misinformation on twitter. arXiv preprint arXiv:1707.03778 (2017)
10. Guille, A., Favre, C.: Mention-anomaly-based event detection and tracking in twitter. In: 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), pp. 375–382. IEEE (2014)
11. Kalyanam, J.: Leveraging social context for modeling topic evolution. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 517–526 (2015)
12. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Min. Knowl. Disc. **7**(4), 373–397 (2003)
13. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature **401**(6755), 788–791 (1999)
14. Liu, J., Chen, Y., Liu, K., Zhao, J.: Event detection via gated multilingual attention mechanism. Statistics **1000**, 1250 (2018)
15. Palen, L.: Online social media in crisis events. Educ. Q. **31**(3), 76–78 (2008)
16. Schinas, M., Papadopoulos, S., Petkos, G., Kompatsiaris, Y., Mitkas, P.A.: Multimodal graph-based event detection and summarization in social media streams. In: ACM International Conference on Multimedia, pp. 189–192 (2015)
17. Shearer, E., Gottfriend, J.: News use across social media platforms 2017. Pew Research Center (2017)
18. Shin, D.S., et al.: STExNMF: spatio-temporally exclusive topic discovery for anomalous event detection. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 435–444. IEEE (2017)
19. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short text classification in twitter to improve information filtering. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 841–842. ACM (2010)
20. Xing, C., Wang, Y., Liu, J., Huang, Y., Ma, W.Y.: Hashtag-based sub-event discovery using mutually generative LDA in twitter. In: AAAI, pp. 2666–2672 (2016)
21. Zhao, L., Sun, Q., Ye, J., Chen, F., Lu, C.T., Ramakrishnan, N.: Multi-task learning for spatio-temporal event forecasting. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1503–1512. ACM (2015)