Review

ENCYCLOPEDIA WITH SEMANTIC COMPUTING AND ROBOTIC INTELLIGENCE
Vol. 1, No. 1 (2017) 1630002 (7 pages)

DOI: 10.1142/S2425038416300020

World Scientific
www.worldscientific.com

Event identification in social networks

Fattane Zarrinkalam*,†,‡ and Ebrahim Bagheri*
*Laboratory for Systems, Software and Semantics (LS3)
Ryerson University, Toronto, Canada

†Department of Computer Engineering, Ferdowsi University of Mashhad
Mashhad, Iran
‡fattane.zarrinkalam@gmail.com

Received 26 June 2016; Accepted 20 July 2016; Published 17 March 2017

Social networks enable users to freely communicate with each other and share their recent news, ongoing activities or views about different topics. As a result, they can be seen as a potentially viable source of information to understand the current emerging topics/events. The ability to model emerging topics is a substantial step to monitor and summarize the information originating from social sources. Applying traditional methods for event detection which are often proposed for processing large, formal and structured documents, are less effective, due to the short length, noisiness and informality of the social posts. Recent event detection techniques address these challenges by exploiting the opportunities behind abundant information available in social networks. This article provides an overview of the state of the art in event detection from social networks.

Keywords: Event detection; social network analysis; topic detection and tracking.

1. Overview

With the emergence and the growing popularity of social networks such as Twitter and Facebook, many users extensively use these platforms to express their feelings and views about a wide variety of social events/topics as they happen, in real time, even before they are released in traditional news outlets.[1] This large amount of data produced by various social media services has recently attracted many researchers to analyze social posts to understand the current emerging topics/events. The ability to identify emerging topics is a substantial step towards monitoring and summarizing the information on social media and provides the potential for understanding and describing real-world events and improving the quality of higher level applications in the fields of traditional news detection, computational journalism[2] and urban monitoring,[3] among others.

1.1. *Problem definition*

The task of topic detection and tracking (TDT) to provide the means for news monitoring from multiple sources in order to keep users updated about news and developments. One of the first forums was the TDT Forum, held within TREC.[4] There has been a significant interest in TDT in the past for static documents in traditional media.[5–8] However, recently the focus has moved to Social Network data sources. Most of these works use Twitter as their source of information, because the information that the users publish on Twitter are more publicly accessible compared to other social networks.

To address the task of detecting topics from social media streams, a stream is considered to be made of posts which are generated by users in social network (e.g., tweets in the case of Twitter). Each post, in addition to its text formed by a sequence of words/terms, includes a user id and a timestamp. Additionally, a time interval of interest and a desired update rate is provided. The expected output of topic detection algorithms is the detection of emerging topics/events. In TDT, a topic is "a seminal event or activity, along with all directly related to events and activities."[4] where event refers to a specific thing that happens at a certain time and place.[9–11] A topic can be represented either through the clustering of documents in the collection or by a set of most important terms or keywords that are selected.

1.2. *Challenges*

The works proposed within the scope of the TDT have proven to be relatively well-established for topic detection in traditional textual corpora such as news articles.[11] However, applying traditional methods for event detection in the social media context poses unique challenges due to the distinctive features of textual data in social media[12] such as Time Sensitivity, Short Length, Unstructured Phrases, and Abundant Information.

‡Corresponding author.

- **Time sensitivity.** Different from traditional textual data, the text in social media has real-time nature. Besides communicating and sharing ideas with each other, users in social networks may publish their feelings and views about a wide variety of recent events several times daily.[7,13,14] Users may want to communicate instantly with friends about "What they are doing (Twitter)" or "What is on their mind" (Facebook).
- **Short length.** Most of social media platforms restrict the length of posts. For example, Twitter allows users to post tweets that are no longer than 140 characters. Similarly, Picasa comments are limited to 512 characters, and personal status messages on Windows Live Messenger are restricted to 128 characters. Unlike standard text with lots of words and their resulting statistics, short messages consist of few phrases or sentences. They cannot provide sufficient context information for effective similarity measure, the basis of many text processing methods.[12,29,57]
- **Unstructured phrases.** In contrast with well-written, structured, and edited news releases, social posts might include large amounts of meaningless messages, polluted and informal content, irregular, and abbreviated words, large number of spelling and grammatical errors, and improper sentence structures and mixed languages. In addition, in social networks, the distribution of content quality has high variance: from very high-quality items to low-quality, sometimes abusive content, which negatively affect the performance of the detection algorithms.[61,12]
- **Abundant information.** In addition to the content itself, social media in general exhibit a rich variety of information sharing tools. For example, Twitter allows users to utilize the "#" symbol, called hashtag, to mark keywords or topics in a Tweet; an image is usually associated with multiple labels which are characterized by different regions in the image; users are able to build connection with others (link information). Previous text analytics sources most often appear as <user, content> structure, while the text analytics in social media is able to derive data from various aspects, which include user, content, link, tag, timestamps and others.[62–64]

In the following, we explain different methodologies that have been proposed in the state of the art to tackle challenges in social networks.

2. Background Literature

According to the availability of the information about events, event detection algorithms can be classified into specified and unspecified techniques.[13] The specified techniques rely on specific information and features that are known about the event, such as a venue, time, type, and description. On the other hand, when there are no prior information available about the event, unspecified event detection technique rely on the social media streams to detect the occurrence of a real-world event.

2.1. *Specified event detection*

Specified event detection aims at identifying known social events which are partially or fully specified with its content or metadata information such as location, time, and venue. For example, Sakaki *et al.*[14] have focused on monitoring tweets posted recently by users to detect earthquake or rainbow. They have used three types of features: the number of words (statistical), the keywords in a tweet message, and the words surrounding users queries (contextual), to train a classifier and classify tweets into positive or negative cases. To identify the location of the event a probabilistic spatiotemporal model is also built. They have evaluated their proposed approach in an earthquake-reporting system in Japan. The authors have found that the statistical features provided the best results, while a small improvement in performance has been achieved by the combination of the three features.

Popescu and Pennacchiotti[15] have proposed a framework to identify controversial events. This framework is based on the notion of a Twitter snapshot which consists of a target entity, a given period, and a set of tweets about the entity from the target period. Given a set of Twitter snapshots, the authors first assign a controversy score to each snapshot and then rank the snapshots according to the controversy score by considering a large number of features, such as linguistic, structural, sentiment, controversy and external features in their model. The authors have concluded that Hashtags are important semantic features to identify the topic of a tweet. Further, they have found that linguistic, structural, and sentiment features provide considerable effects for controversy detection.

Benson *et al.*[16] have proposed a model to identify a comprehensive list of musical events from Twitter based on artist–venue pairs. Their model is based on a conditional random field (CRF) to extract the artist name and location of the event. The input features to CRF model include word shape; a set of regular expressions for common emoticons, time references, and venue types; a bag of words for artist names extracted from external source (e.g., Wikipedia); and a bag of words for city venue names. Lee and Sumiya[17] have proposed a geosocial local event detection system, to identify local festivals. They have collected Twitter geotagged data for a specific region and used k-means algorithm applied to the geographical coordinates of the collected data to divide them into several regions of interest (ROI). The authors have found that an increased user activity, i.e., moving inside or coming to an ROI, combined with an increased number of tweets provides strong indicator of local festivals.

Becker *et al.*[18] have used a combination of simple rules and query building strategies to identify planned events from Twitter. They have identified tweets related to an event by utilizing simple query building strategies that derive queries from the structured description of the event and its associated aspects (e.g., time and venue). To provide high-precision tweets, they have asked an annotator to label the results

returned by each strategy, then they have employed term-frequency analysis and co-location techniques to improve recall to identify descriptive event terms and phrases, which are then used recursively to define new queries. Similarly, Becker *et al.*[19] have proposed centrality-based approaches to extract high-quality, relevant, and useful related tweets to an event. Their approach is based on the idea that the most topically central messages in a cluster are more likely to reflect key aspects of the event than other less central cluster messages.

2.2. *Unspecified event detection*

The real-time nature of social posts reflect events as they happen about emerging events, breaking news, and general topics that attract the attention of a large number of users. Therefore, these posts are useful for unknown event detection. Three main approaches have been studied in the literature for this purpose: topic-modeling, document-clustering and feature-clustering approaches[1]:

2.2.1. *Topic modeling methods*

Topic modeling methods such as LDA assume that a document is a mixture of topics and implicitly use co-occurrence patterns of terms to extract sets of correlated terms as topics of a text corpus.[20] More recent approaches have extended LDA to provide support for temporality including the recent topics over time (ToT) model,[21] which simultaneously captures term co-occurrences and locality of those patterns over time and is hence able to discover more event-specific topics.

The majority of existing topic models including LDA and TOT, focus on regular documents, such as research papers, consisting of a relatively small number of long and high quality documents. However, social posts are shorter and noisier than traditional documents. Users in social networks are not professional writers and use very diverse vocabulary, and there are many abbreviations and typos. Moreover, the online social media websites have a social network full of context information, such as user features and user-generated labels, which have been normally ignored by the existing topic models. As a result, they may not perform so well on social posts and might suffer from the sparsity problem.[22–25] To address this problem, some works aggregate multiple short texts to create a single document and discover the topics by running LDA over this document.[25–27] For instance, Hong and Davison,[30] have combined all the tweets from each user as one document and apply LDA to extract the document topic mixture, which represents the user interest. However, in social networks a small number of users usually account for a significant portion of the content. This makes the aggregation process less effective.

There are some recent works that deal with the sparsity problem by applying some restrictions to simplify the conventional topic models or develop novel topic models for short texts. For example, Zhao *et al.*[28] have proposed the Twitter-LDA model. It assumes that a single tweet contains only one topic, which differs from the standard LDA model. Diao *et al.*[29] have proposed biterm topic model (BTM), a novel topic model for short texts, by learning the topics by directly modeling the generation of word co-occurrence patterns (i.e., biterms) in the whole corpus. BTM is extended by Yan *et al.*[57] by incorporating the burstiness of biterms as prior knowledge for bursty topic modeling and proposed a new probabilistic model named bursty biterm topic model (BBTM) to discover bursty topics in microblogs.

It should be noted that applying such restrictions and the fact that the number of topics in LDA is assumed to be fixed can be considered strong assumptions for social network content because of the dynamic nature of social networks.

2.2.2. *Document clustering methods*

Document-clustering methods extract topics by clustering related documents and consider each resulting cluster as a topic. They mostly represent textual content of each document as a bag of words or *n*-grams using TF/IDF weighting schema and utilize cosine similarity measures to compute the co-occurrence of their words/*n*-grams.[31,32] Document-clustering methods suffer from cluster fragmentation problems and since the similarity of two documents may be sensitive to noise, they perform much better on long and formal documents than social posts which are short, noisy and informal.[33] To address this problem, some works, in addition to textual information, take into account other rich attributes of social posts such as timestamps, publisher, location and hashtags.[33,35,36] These works typically differ in that they use different information and different measures to compute the semantic distance between documents.

For example, Dong *et al.*[34] have proposed a wavelet-based scheme to compute the pairwise similarity of tweets based on temporal, spatial, and textual features of tweets. Fang *et al.*[35] have clustered tweets by taking into consideration multi-relations between tweets measured using different features such as textual data, hashtags and timestamp. Petrovic *et al.*[31] have proposed a method to detect new events from a stream of Twitter posts. To make event detection feasible on web-scale corpora, the authors have proposed a constant time and space approach based on an adapted variant of locality sensitive hashing methods. The authors have found that ranking according to the number of users is better than ranking according to the number of tweets and considering entropy of the message reduces the amount of spam messages in the output. Becker *et al.*[39] have first proposed a method to identify real-world events using an a classical incremental clustering algorithm. Then, they have classified the clusters content into real-world events or nonevents. These nonevents includes Twitter-centric topics, which are trending activities in Twitter that do not reflect any real-world occurrences. They have trained the classifier on the variety of features

including temporal, social, topical, and Twitter-centric features to decide whether the cluster (and its associated messages) contains real-world event.

In the context of breaking news detection from Twitter, Sankaranarayanan *et al.*[37] have proposed TwitterStand which is a news processing system for Twitter to capture tweets related to late breaking news that takes into account both textual similarity and temporal proximity. They have used a naive Bayes classifier to separate news from irrelevant information and an online clustering algorithm based on weighted term vector to cluster news. Further, they have used hashtags to reduce clustering errors. Similarly, Phuvipadawat and Murata[38] have presented a method for breaking news detection in Twitter. They first sample tweets using predefined search queries, and then group them together to form a news story. Similarity between posts is based on tf-idf with an increased weight for proper noun terms, hashtags, and usernames. They use a weighted combination of number of followers (reliability) and the number of retweeted messages (popularity) with a time adjustment for the freshness of the message to rank each cluster. New messages are included in a cluster if they are similar to the first post and to the top-k terms in that cluster.

2.2.3. *Feature clustering methods*

Feature clustering methods try to extract features of topics from documents. Opics are then detected by clustering features based on their semantic relatedness. As one of the earlier work that focused on Twitter data, Cataldi *et al.*[40] have constructed a co-occurrence graph of emerging terms selected based on both the frequency of their occurrence and the importance of the users. The authors have applied a graph-based method in order to extract emerging topics. Similarly, Long *et al.*[41] have constructed a co-occurrence graph by extracting topical words from daily posts. To extract events during a time period, they have applied a top-down hierarchical clustering algorithm over the co-occurrence graph. After detecting events in different time periods, they track changes of events in consecutive time periods and summarize an event by finding the most relevant posts to that event. The algorithm by Sayyadi *et al.*[42] builds a term co-occurrence graph, whose nodes are clustered using a community detection algorithm based on betweenness centrality. Additionally, topic description is enriched with the documents that are most relevant to the identified terms. Graphs of short phrases, rather than of single terms, connected by edges representing lexical inclusion or similarity have also been used.

There are also some works that utilize signal processing techniques for event detection from social networks. For instance, Weng and Lee[43] have used wavelet analysis to discover events in Twitter streams. First, they have selected bursty words by representing each word as a frequency-based signal and measuring the bursty energy of each word using autocorrelation. Then, they build a graph whose nodes are bursty words and edges are cross-correlation between each pair of bursty words and used graph-partitioning techniques to discover events. Similarly, Cordeiro[44] has used wavelet analysis for event detection from Twitter. This author has constructed a wavelet signal for each hashtag, instead of words, over time by counting the hashtag mentions in each interval. Then, he has applied the continuous wavelet transformation to get a time-frequency representation of each signal and used peal analysis and local maxima detection techniques to detect an event within a given time interval. He *et al.*[6] have used Discrete Fourier Transform to classify the signal for each term based on its power and periodicity. Depending on the identified class, the distribution of appearance of a term in time is modeled using one or more Gaussians, and the KL-divergence between the distributions is then used to determine clusters.

In general, most of these works are based on terms and compute similarity between pairs of terms based on their co-occurrence patterns. Petkos *et al.*[45] have argued that the algorithms that are only based on pairwise co-occurrence patterns cannot distinguish between topics which are specific to a given corpus. Therefore, they have proposed a soft frequent pattern mining approach to detect finer grained topics. Zarrinkalam *et al.*[46] have inferred fine grained users' topics of interest by viewing each topic as a conjunction of several concepts, instead of terms, and benefit from a graph clustering algorithms to extract temporally related concepts in a given time period. Further, they compute inter-concept similarity by customizing the concepts co-occurrences within a single tweet to an increased, yet semantic preserving context.

3. Application Areas

There are a set of interesting applications of event/topic detection systems and methods. Health monitoring and management is an application in which the detection of events plays an important role. For example, Culotta[49] have explored the possibility of tracking influenza by analyzing Twitter data. They have proposed an approach to predict influenza-like illnesses rates in a population to identify influenza-related messages and compare a number of regression models to correlate these messages with U.S. Centers for disease control and prevention (CDC) statistics. Similarly, Aramaki *et al.*[52] have identified flu outbreaks by analyzing tweets about Influenza. Their results are similar to Google-trends based flu outbreak detection especially in the early stages of the outbreak

Paul and Dredze[50] have proposed a new topic model for Twitter, named ailment topic aspect model (ATAM), that associates symptoms, treatments and general words with diseases. It produces more detailed ailment symptoms and tracks disease rates consistent with published government statistics (influenza surveillance) despite the lack of supervised influenza training data. In Ref. 51, the authors have used Twitter to identify posts which are about health issues

and they have investigated what types of links the users consult for publishing health related information.

Natural events detection (Disasters) is another application for the automatic detection of events from social network. For example, Sakaki *et al.*[14] have proposed an algorithm to monitor the real-time interaction of events, such as earthquakes in Twitter. Their approach can detect an earthquake with high probability by monitoring tweets and detects earthquakes promptly and sends e-mails to registered users. The response time of the system is shown to be quite fast, similar to the Japan Meteorological Agency. Cheong and Cheong[53] have analyzed the tweets during Australian floods of 2011 to identify active players and their effectiveness in disseminating critical information. As their secondary goal, they have identified the most important users among Australian floods to be: local authorities (Queensland Police Services), political personalities (Premier, Prime Minister, Opposition Leader and Member of Parliament), social media volunteers, traditional media reporters, and people from not for profit, humanitarian, and community associations. In Ref. 54, the authors have applied visual analytics approach to a set of georeferenced Tweets to detect flood events in Germany providing visual information on the map. Their results confirmed the potential of Twitter as a distributed "social sensor". To overcome some caveats in interpreting immediate results, they have explored incorporating evidence from other data sources.

Some applications with marketing purpose have also utilized event detection methods. For example, Medvent *et al.*[55] have focused on detecting events related to three major brands including Google, Microsoft and Apple. Examples of such events are the release of a new product like the new iPad or Microsoft Security Essential software. In order to achieve the desired outcome, the authors study the sentiment of the tweets. Si *et al.*[56] have proposed a continuous Dirichlet Process Mixture model for Twitter sentiment, to help predict the stock market. They extract the sentiment of each tweet based on its opinion words distribution to build a sentiment time series. Then, they regress the stock index and the Twitter sentiment time series to predict the market.

There are also some works that model user's interests over detected events from social networks. For example, Zarrinkalam *et al.*[47] have proposed a graph-based link prediction schema to model a user's interest profile over a set of topics/events present in Twitter in a specified time interval. They have considered both explicit and implicit interests of the user. Their approach is independent of the underlying topic detection method, therefore, they have adopted two types of topic extraction methods: feature clustering and LDA approaches. Fani *et al.*[48] have proposed a graph-based framework that utilizes multivariate time series analysis to tackle the problem of detecting time-sensitive topic-based communities of user who have similar temporal tendency with regards to topics of interests in Twitter. To discover topics of interest from Twitter, they have utilized an LDA-based topic model that jointly captures word co-occurrences and locality of those patterns over time.

4. Conclusion and Future Directions

Due to the fast-growing and availability of social network data, many researchers has recently become attracted to event detection from social networks. Event detection aims at finding real-world occurrences that unfold over space and time. The problem of event detection from social networks has faced different challenges due to the short length, noisiness and informality of the social posts. In this paper, we presented an overview of the recent techniques to address this problem. These techniques are classified according to the type of target event into specified or unspecified event detection. Further, we provided some potential applications in which event detection techniques are utilized.

While there are many works related to event detection from social networks, one challenge that has to be addressed in this research area is the lack of public datasets. Privacy issues along with Social Network companies' terms of use hinder the availability of shared data. This obstacle is of great significance since it relates to the repeatability of experiments and comparison between approaches. As a result, most of the current approaches have focused on a single data source, specially the Twitter platform because of the usability and accessibility of the Twitter API. However, being dependent on a single data source entails many risks. Therefore, one future direction can be monitoring and analyzing the events and activities from different social network services simultaneously. As an example, Kaleel[58] have followed this idea and utilized Twitter posts and Facebook messages for event detection. They have used LSH to classify messages. The proposed algorithm first independently identifies new events (first stories) from both sources (Twitter, Facebook) and then hashes them into clusters.

As another future direction, there is no method in the field of event detection from social networks which is able to automatically answer the following questions for each detected event: what, when, where, and by whom. Therefore, improving current methods to address these questions can be a new future direction. As a social post is often associated with spatial and temporal information, it is possible to detect when and where an event happens.

Several further directions can be explored to achieve efficient and reliable event detection systems such as: investigating how to model the social streams together with other data sources, like news streams to better detect and represent events,[60] designing better feature extraction and query generation techniques, designing more accurate filtering and detection algorithms as well as techniques to support multiple languages.[59]

References

[1]L. M. Aiello, G. Petkos, C. Martin and D. Corney, Sensing trending topics in twitter, *IEEE Trans. Multimedia* **15**(6) 1268.

[2]S. Cohen, J. T. Hamilton and F. Turner, Computational journalism, *Comm. ACM* **54**, 66 (2011).

[3]D. Quercia, J. Ellis, L. Capra and J. Crowcroft, Tracking "Gross Community Happiness" from Tweets, in *CSCW: ACM Conf. Computer Supported Cooperative Work*, New York, NY, USA, (2012), pp. 965–968.

[4]J. Fiscus and G. Duddington, Topic detection and tracking overview, *Topic Detection and Tracking: Event-Based Information Organization* (2002), pp. 17–31.

[5]J. Kleinberg, Bursty and hierarchical structure in streams, *Proc. Eighth ACM SIGKDD International Conf. Knowledge Discovery and Data Mining, KDD '02*, ACM, New York, NY (2002), pp. 91–101.

[6]Q. He, K. Chang and E.-P. Lim. Analyzing feature trajectories for event detection, *Proc. 30th Annual Int. ACM SIGIR Conf. Research and Development in Information Retrieval, SIGIR '07*, ACM, New York, NY (2007), pp. 207–214.

[7]X. Wang, C. Zhai, X. Hu and R. Sproat, Mining correAn event extraction model based on timelinelated bursty topic patterns from coordinated text streams, *Proc. 13th ACM SIGKDD International Conf. Knowledge Discovery and Data Mining, KDD '07*, ACM, New York, NY (2007), pp. 784–793.

[8]S. Goorha and L. Ungar, Discovery of significant emerging trends, *Proc. 16th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD '10*, ACM, New York, NY (2010), pp. 57–64.

[9]R. Troncy, B. Malocha and A. T. S. Fialho, Linking events with media, *Proc. 6th Int. Conf. Semantic Systems, I-SEMANTICS '10*, ACM, New York, NY (2010), pp. 42:1–42:4.

[10]L. Xie, H. Sundaram and M. Campbell, Event mining in multimedia stream, *Proc. IEEE* **96**(4), 623 (2008).

[11]J. Allan, J. Carbonell, G. Doddington, J. Yamron and Y. Yang, Topic detection and tracking pilot study final report, *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA (1998), pp. 194–218.

[12]X. Hu and H. Liu, Text analytics in social media, *Mining Text Data* (Springer US, 2012), pp. 385–414.

[13]F. Atefeh and W. Khreich, A survey of techniques for event detection in Twitter, *Comput. Intell.* **31**(1), 132 (2015).

[14]T. Sakaki, M. Okazaki and Y. Matsuo, Earthquake shakes Twitter users: Real-time event detection by social sensors, *Proc. 19th WWW* (2010).

[15]A.-M. Popescu and M. Pennacchiotti, Detecting controversial events from twitter, *Proc. CIKM '10 Proc. 19th ACM Int. Conf. Information and Knowledge Management* (2010), pp. 1873–1876.

[16]E. Benson, A. Haghighi and R. Barzilay, Event discovery in social media feeds, *Proc. HLT '11 Proc. 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (2011), pp. 389–398.

[17]R. Lee and K. Sumiya, Measuring geographical regularities of crowd behaviors for Twitter-based geosocial event detection, *Proc. 2nd ACM SIGSPATIAL Int. Workshop on Location Based Social Networks, LBSN '10*, ACM, New York, NY (2010), pp. 1–10.

[18]H. Becker, F. Chen, D. Iter, M. Naaman and L. Gravano, Automatic identification and presentation of Twitter content for planned events, *Int. AAAI Conf. Weblogs and Social Media* (Barcelona, Spain, 2011).

[19]H. Becker, M. Naaman and L. Gravano, Selecting quality Twitter content for events, *Int. AAAI Conf. Weblogs and Social Media* (Barcelona, Spain, 2011).

[20]D. Blei, Probabilistic topic models, *Commun. ACM* **55**(4), 77 (2012).

[21]X. Wang and A. McCallum, Topics over time: A non-Markov continuous-time model of topical trends, *Proc. 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (2006), pp. 424–433.

[22]M. Michelson and S. A. Macskassy, Discovering users' topics of interest on twitter: A first look, *4th Workshop on Analytics for Noisy Unstructured Text Data (AND'10)* (2010), pp. 73–80.

[23]B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu and M. Demirbas, Short text classification in twitter to improve information filtering, *33rd Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (2010), pp. 841–842.

[24]C. Xueqi, Y. Xiaohui, L. Yanyan and G. Jiafeng, BTM: Topic modeling over short texts, *IEEE Trans. Knowl. Data Eng.* **26**(12), 2928 (2014).

[25]R. Mehrotra, S. Sanner, W. Buntine and L. Xie, Improving lda topic models for microblogs via tweet pooling and automatic labeling, *36th Int. ACM SIGIR Conf. Research and Development in Information Retrieval* (2013).

[26]J. Weng, E. P. Lim, J, Jiang and Q. He, TwitterRank: Finding topic-sensitive influential twitterers, *3rd ACM Int. Conf. Web Search and Data Mining (WSDM '10)* (2010), pp. 261–270.

[27]N. F. N. Rajani, K. McArdle and J. Baldridge, Extracting topics based on authors, recipients and content in microblogs, *37th Int. ACM SIGIR Conf. Research & Development in Information Retrieval* (2014), pp. 1171–1174.

[28]W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan and X. Li, Comparing twitter and traditional media using topic models, *33rd European Conf. Advances in Information Retrieval* (2011), pp. 338–349.

[29]Q. Diao, J. Jiang, F. Zhu and E. Lim, Finding bursty topics from microblogs, *50th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (ACL HLT, 2012), pp. 536–544.

[30]L. Hong and B. Davison, Empirical study of topic modeling in Twitter, *1st ACM Workshop on Social Media Analytics* (2010).

[31]S. Petrović, M. Osborne and V. Lavrenko, Streaming first story detection with application to Twitter, *Proc. HLT: Ann. Conf. North American Chapter of the Association for Computational Linguistics* (2010), pp. 181–189.

[32]G. Ifrim, B. Shi and I. Brigadir, Event detection in twitter using aggressive filtering and hierarchical tweet clustering, SNOW-DC@WWW (2014), pp. 33–40.

[33]G. P. C. Fung, J. X. Yu, P. S. Yu and H. Lu, Parameter free bursty events detection in text streams, *Proc. 31st Int. Conf. Very Large Data Bases, VLDB '05* (2005), pp. 181–192.

[34]X. Dong, D. Mavroeidis, F. Calabrese and P. Frossard, Multiscale event detection in social media, CoRR abs/1406.7842 (2014).

[35]Y. Fang, H. Zhang, Y. Ye and X. Li, Detecting hot topics from Twitter: A multiview approach, *J. Inform. Sci.* **40**(5), 578 (2014).

[36]S-H. Yang, A. Kolcz, A. Schlaikjer and P. Gupta, Large-scale high-precision topic modeling on twitter, *Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (2014), pp. 1907–1916.

[37]J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman and J. Sperling, Twitterstand: News in tweets, *Proc. 17th ACM*

SIGSPATIAL Int. Conf. Advances in Geographic Information Systems (2009), pp. 42–51.
[38]S. Phuvipadawat and T. Murata, Breaking news detection and tracking in Twitter, *IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Vol. 3, Toronto, ON (2014), pp. 120–123.
[39]H. Becker, M. Naaman and L. Gravano, Beyond trending topics: Real-world event identification on Twitter, ICWSM (Barcelona, Spain, 2011).
[40]M. Cataldi, L. D. Caro and C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, *10th Int. Workshop on Multimedia Data Mining, MDMKDD '10* (USA, 2010), pp. 4:1–4:10.
[41]R. Long, H. Wang, Y. Chen, O. Jin and Y. Yu, Towards effective event detection, tracking and summarization on microblog data, *Web-Age Information Management*, Vol. 6897 (2011), pp. 652–663.
[42]H. Sayyadi, M. Hurst and A. Maykov, Event detection and tracking in social streams, *Proc. ICWSM 2009* (USA, 2009).
[43]J. Weng and F. Lee, Event detection in Twitter, *5th Int. AAAI Conf. Weblogs and Social Media* (2011), pp. 401–408.
[44]M. Cordeiro, Twitter event detection: Combining wavelet analysis and topic inference summarization, *Doctoral Symp. Informatics Engineering* (2012).
[45]G. Petkos, S. Papadopoulos, L. M. Aiello, R. Skraba and Y. Kompatsiaris, A soft frequent pattern mining approach for textual topic detection, *4th Int. Conf. Web Intelligence, Mining and Semantics (WIMS14)* (2014), pp. 25:1–25:10.
[46]F. Zarrinkalam, H. Fani, E. Bagheri, M. Kahani and W. Du, Semantics-enabled user interest detection from twitter, *IEEE/WIC/ACM Web Intell. Conf.* (2015).
[47]F. Zarrinkalam, H. Fani, E. Bagheri and M. Kahani, Inferring implicit topical interests on twitter, *38th European Conf. IR Research, ECIR 2016*, Padua, Italy, March 20–23 (2016), pp. 479–491.
[48]H. Fani, F. Zarrinkalam, E. Bagheri and W. Du, Time-sensitive topic-based communities on twitter, *29th Canadian Conf. Artificial Intelligence, Canadian AI* 2016, Victoria, BC, Canada, May 31–June 3 (2016), pp. 192–204.
[49]A. Culotta, Towards detecting influenza epidemics by analyzing twitter messages, *KDD Workshop on Social Media Analytics* (2010), pp. 115–122.
[50]J. M. Paul and M. Dredze, A model for mining public health topics from twitter, Technical Report, Johns Hopkins University (2011).
[51]E. V. D. Goot, H. Tanev and J. Linge, Combining twitter and media reports on public health events in medisys, *Proc. 22nd Int. Conf. World Wide Web Companion, International World Wide Web Conf. Steering Committee* (2013), pp. 703–705.
[52]E. Aramaki, S. Maskawa and M. Morita, Twitter catches the flu: Detecting influenza epidemics using Twitter, *Proc. Conf. Empirical Methods in Natural Language Processing* (2011), pp. 1568–1576.
[53]F. Cheong and C. Cheong, Social media data mining: A social network analysis of tweets during the 2010–2011 australian floods, *PACIS*, July (2011), pp. 1–16.
[54]G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe and H. Stange, Tracing the German centennial flood in the stream of tweets: First lessons learned, *Proc. Second ACM SIGSPATIAL Int. Workshop on Crowdsourced and Volunteered Geographic Information* (ACM, New York, NY, USA, 2013), pp. 31–38.
[55]E. Medvet and A. Bartoli, Brand-related events detection, classification and summarization on twitter, *IEEE/WIC/ACM Int. Conf. Web Intelligence and Intelligent Agent Technology* (2012), pp. 297–302.
[56]J. Si, A. Mukherjee, B. Liu, Q. Li, H. Li and X. Deng, Exploiting topic based twitter sentiment for stock prediction, *ACL* (2) (2013), pp. 24–29.
[57]X. Yan, J. Guo, Y. Lan, J. Xu and X. Cheng, A probabilistic model for bursty topic discovery in microblogs, *AAAI Conf. Artificial Intelligence* (2015), pp. 353–359.
[58]S. B. Kaleel, Event detection and trending in multiple social networking sites, *Proc. 16th Communications & Networking Symp. Society for Computer Simulation International* (2013).
[59]G. Lejeune, R. Brixtel, A. Doucet and N. Lucas, Multilingual event extraction for epidemic detection, *Artif. Intell. Med.* **65**(2), 131 (2015).
[60]W. Gao, P. Li and K. Darwish, Joint topic modeling for event summarization across news and social media streams, *Proc. 21st ACM Int. Conf. Information and Knowledge Management, CIKM '12* (New York, NY, USA, ACM, 2012), pp. 1173–1182.
[61]P. Ferragina and U. Scaiella, Fast and accurate annotation of short texts with wikipedia pages, *J. IEEE Softw.* **29**(1), 70 (2012).
[62]S. Yang, A. Kolcz, A. Schlaikjer and P. Gupta, Large-scale high-precision topic modeling on twitter, *Proc. 20th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining* (2014), pp. 1907–1916.
[63]D. Duan, Y. Li, R. Li, R. Zhang, X. Gu and K. Wen, LIMTopic: A framework of incorporating link based importance into topic modeling, *IEEE Trans. Knowl. Data Eng.* (2013), 2493–2506.
[64]M. JafariAsbagh, E. Ferrara, O. Varol, F. Menczer and A. Flammini, Clustering memes in social media streams, *Soc. Netw. Anal. Min.* (2014).