Journal of Computer Science

Review

A Survey on Event Detection Models for Text Data Streams

[1,2]Wafa Zubair AL-Dyani, [2]Farzana Kabir Ahmad and [2]Siti Sakira Kamaruddin

[1]*Department of Computer Science, College of Computing and Information Technology, Hadramout University, Hadramout, Yemen*
[2]*School of Computing, College of Arts and Science, Universiti Utara Malaysia, Sintok Kedah, Malaysia*

Article history
Received: 22-04-2020
Revised: 23-06-2020
Accepted: 13-07-2020

Corresponding Author:
Farzana Kabir Ahmad,
School of Computing, College of Arts and Science, Universiti Utara Malaysia, Sintok Kedah, Malaysia
Email: farzana58@uum.edu.my

Abstract: Event Detection (ED) is a study area that attracts the attention of decision-makers from various disciplines in order to help them in taking the right decision. ED has been examined on various text streams like Twitter, Facebook, Emails, Blogs, Web Forums and newswires. Many ED models have been proposed in literature. In general, ED model consists of six main phases: Data collection, pre-processing, feature selection, event detection, performance evaluation and result representation. Among these phases, event detection phase has a vital rule in the performance of the ED model. Consequently, numerous supervised, unsupervised, semi-supervised detection methods have been introduced for this phase. However, unsupervised methods have been extensively utilized as ED process is considered as unsupervised task. Hence, such methods need to be categorized on such a way so it can help researchers to understand and identified the limitations lay in these methods. In this survey, ED models for text data from various Social Network sites (SNs) are analyzed based on domain type, detection methods, type of detection task. In addition, main categories for unsupervised detection methods are explicitly mentioned with revising their related works. Moreover, the major open challenges faced by researchers for building ED models are explained and discussed in detail. The main objective of this survey paper is to provide a complete view of the recent developments in ED field. Hence, help scholars to identify the limitations of existing ED models for text data and help them to recognize the interesting future works directions.

Keywords: Event Detection Model, Text Data, Challenges, Detection Methods\Techniques

Introduction

Due to the increasing popularity of Social Networks sites (SNs), many people utilize such platforms to share their opinions, sentiments and news about different real-world events (Alkubaisi *et al*., 2018). Consequently, a huge amount of structured and unstructured data is generated which comes in various forms such as text, video, photo and audio (Verma *et al*., 2016). Among these types, text data streams represent about 80% of the total data that generates from different sources like news websites, web forums, emails, blogs, SNs e.g., Facebook and Twitter (Singh, 2016). Such data stream is defined as an environment in which text elements arrive online and the hosted system often has no control on the order the data items arrive to be processed. In addition, text data streams are boundless in size and once they have been processed, it is either archived or discarded. In fact, it is reported that majority of text data streams are generally talking about real-world events (Nurwidyantoro and Winarko, 2013). Thus, it has attracted and encouraged many researchers from different disciplines to collect and analyze this data in order to identify the emerging events as well as to monitor and summarize the information related to these events. Event Detection (ED) is a process of automatically identifying real-world events from different data streams and contains information about what has happened, where and when it has happened and who was involved (Fu *et al*., 2014).

In literature, many review papers have been introduced which have studied, investigated and discussed different proposed ED models for various SNs

(Dou *et al*., 2012a; Goswami and Kumar, 2016; Panagiotou *et al*., 2016; Zarrinkalam and Bagheri, 2017). In contrast, several studies were done specifically for Twitter data (Atefeh and Khreich, 2015; Deng *et al*., 2015; Hasan *et al*., 2018; Weng and Lee, 2011). Dou *et al*. (2012b) reviewed ED studies based on different tasks of ED e.g., new ED, retrospective ED, event tracking, event summarization and event association. Panagiotou *et al*. (2016) provided a detail description about ED concepts as well as reviewed ED works based on various ED methods e.g., clustering, anomaly, first story, topic specific techniques. In the same year, (Goswami and Kumar, 2016) introduced the recent proposed ED models for various text data streams like newswire, email, web forums, blogs and microblogs as well as discussed, their open challenges. In contrast, (Zarrinkalam and Bagheri, 2017) focused only on four general challenges for ED models e.g., short length, volume, time sensitivity and style of writing. In addition, they have categorized existing ED studies into specified and unspecified events, which further classified into topic modelling, document clustering and feature clustering methods.

Despite the existing of such ED review papers for text data, yet there is a need for a study that specializes on collecting and analyzing ED studies in terms of the methods used in the ED phase, more specifically, unsupervised methods. This is because ED process is essentially considered to be unsupervised task as no information is available in advance about events. Besides that, there is a need to mention and explain in detail the most important problems and challenges facing the researchers in ED field during the process of building ED model for SNs' text data. Given such gap, this survey paper comes to present a comprehensive analysis of recent ED studies which have utilized unsupervised methods to detect real-world events. It investigates the related ED studies according to different aspects: Domain type (i.e., open and specific), detection methods used (i.e., supervised, unsupervised and semi-supervised) and detection task (NEW or RED). Furthermore, studies using unsupervised methods have classified by the authors of this survey into different five categories i.e., query based, statistical based, probabilistically based, clustering based and graphical based. Moreover, major open challenges for building ED models have explained and discussed to identify the future direction for researchers in ED field.

The rest of this paper is organized as follows. Sections two delivers brief introduction about ED's main concepts. An analysis of existing ED models is presented in Section three. Section four describes the different categories of ED methods and reviews their related works. Section five discusses the main open challenges for building ED model and provide future recommendations to solve such issues followed by conclusion in Section six.

Concepts of Event Detection

ED originally is addressed by a research program called Topic Detection and Tracking (TDT) (Lavanya *et al*., 2014), which is a joint project of CMU, DARPA and Dragon systems (Goswami and Kumar, 2016). Topic Detection and Tracking (TDT) initially organizes news stories under same topic in a way so that people can easily recognize the significant real-world events (Dai *et al*., 2010). A news topic is not just a collection of news stories, but rather it contains a set of events. The main layers of any news are topic, event and news story Fig. 1. Topic is "a collection of events/stories that talk about the same subject". TDT differentiates between event and topic on the basis that event is characterized by time and location.

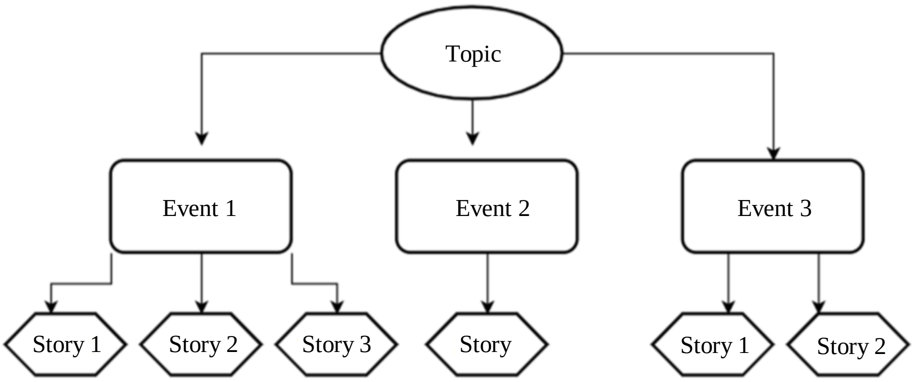


Fig. 1: Hierarchy layers of news in TDT

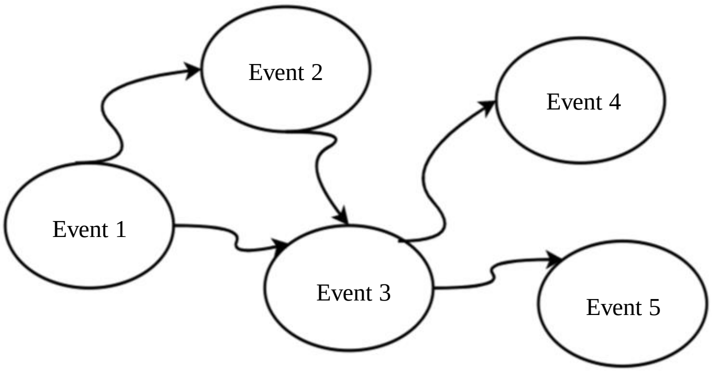
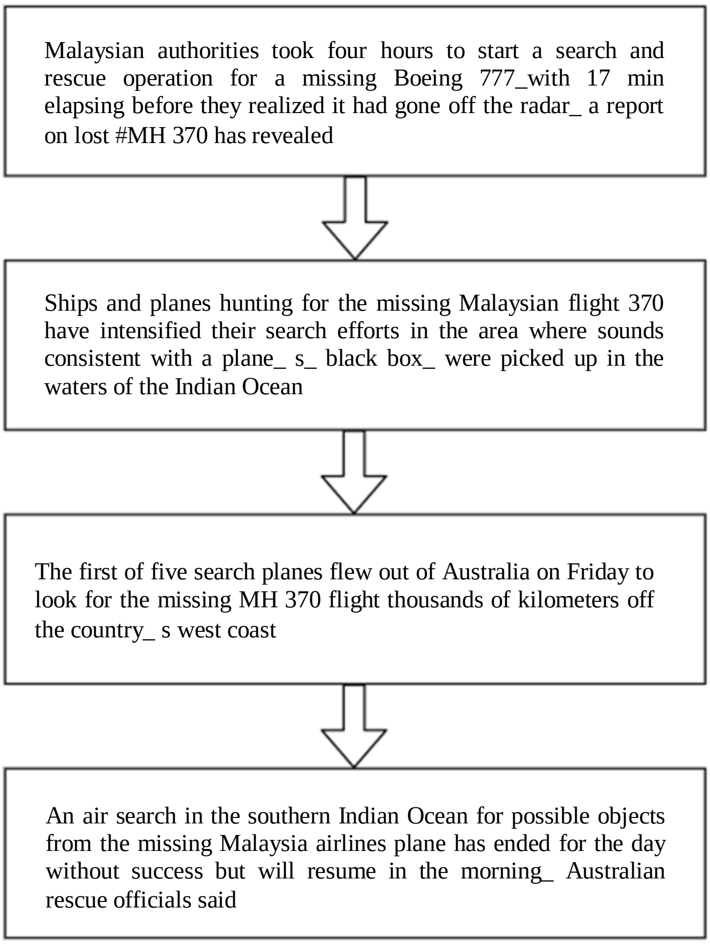


Fig. 2: Event association



Malaysian authorities took four hours to start a search and rescue operation for a missing Boeing 777_with 17 min elapsing before they realized it had gone off the radar_ a report on lost #MH 370 has revealed

Ships and planes hunting for the missing Malaysian flight 370 have intensified their search efforts in the area where sounds consistent with a plane_ s_ black box_ were picked up in the waters of the Indian Ocean

The first of five search planes flew out of Australia on Friday to look for the missing MH 370 flight thousands of kilometers off the country_ s west coast

An air search in the southern Indian Ocean for possible objects from the missing Malaysia airlines plane has ended for the day without success but will resume in the morning_ Australian rescue officials said

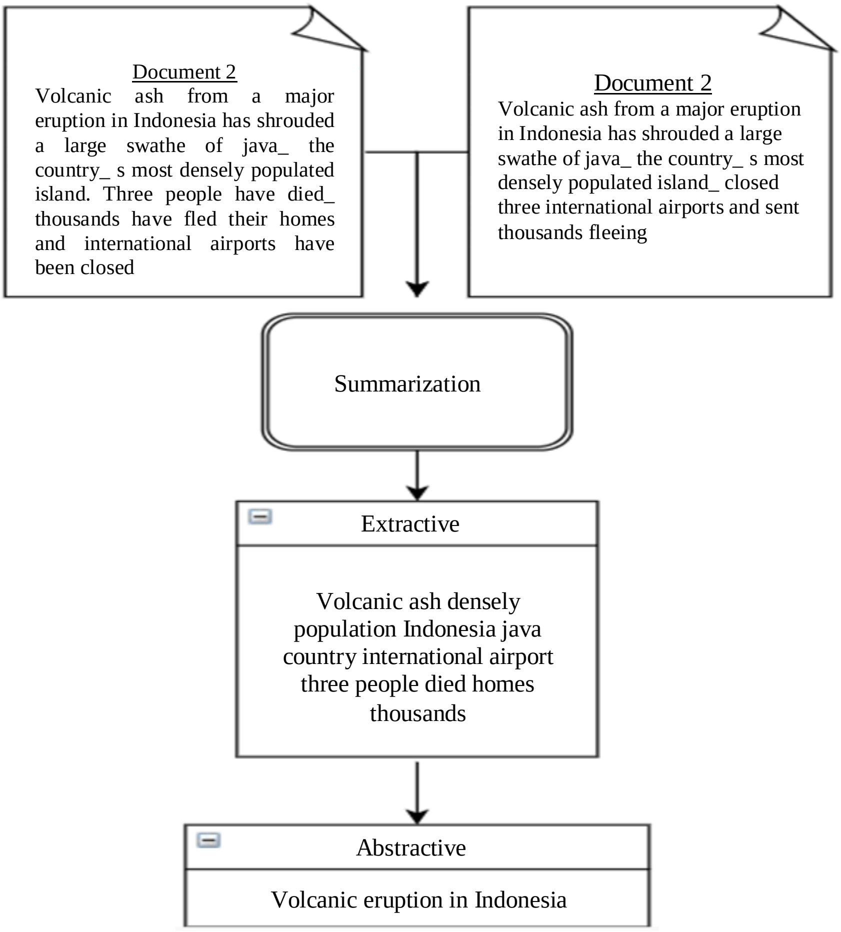
Fig. 3: Event evolution over time

Therefore, an event is defined as "a specific incident that happened in specific time and location" and should cover answers to questions like what has happened, when, where and who was involved. For instance, "Mumbai Terrorist Attack on 26 November 2008", represents as an event, while "terrorist attacks" is more common topic (Goswami and Kumar, 2016). Lastly, story is usually forming the body of a news article/post and a single event may include diverse stories. Several studies have interested on topic detection (Blei, 2012). On contrary, different researchers have detected event and topic interchangeably i.e., they detected events as topics from news stories and vice versa (Huang *et al*., 2013). TDT include four main subtasks: Event detection, event association, event tracking (evolving) and event summarization (Nurwidyantoro and Winarko, 2013). Event detection automatically identifies events from SNs streams. Event association identifies the relationship among events Fig. 2.

Where direction of → represents the different associations (relationships) among various events. Event 1 has a direct relationship with event 2 and 3, meanwhile it has indirect links with event 4 and 5. On the other hand, event 4 and 5 have no any relationships among them. Event tracking recognizes the evolvement process of an event over time Fig. 3.

Event summarization précises the event from the corpus Fig. 4. This task includes two main steps: Extractive summarization and abstractive summarization. In the former step, the most informative sentence is selected, while in the later step a sentence is generated that describes the contents of the documents.

Generally, ED Models have two main types called New Event Detection (NED) models also known as online ED and Retrospective Event Detection (RED) models or called offline ED (Panagiotou *et al*., 2016). NED focuses on detecting newly occurred events from online data streams. NED is a powerful model where novel information is extracted and analyzed from a rapidly growing data with intension to support decision makers in some domains like natural disaster, stock markets, news analyses, etc. On the other hand, RED concerns on discovering past and unseen events from historical repository in offline manner in order to study the situation and answers questions related to the detected events (Atefeh and Khreich, 2015). RED has been studied for a long time but remains as an active research area due to its wide area applications in sport, education, financial and news (Chandran *et al*., 2017).



Document 2
Volcanic ash from a major eruption in Indonesia has shrouded a large swathe of java_ the country_ s most densely populated island. Three people have died_ thousands have fled their homes and international airports have been closed

Document 2
Volcanic ash from a major eruption in Indonesia has shrouded a large swathe of java_ the country_ s most densely populated island_ closed three international airports and sent thousands fleeing

Volcanic ash densely population Indonesia java country international airport three people died homes thousands

Volcanic eruption in Indonesia

Fig. 4: Event Summarization

Event Detection Models

From Table 1, it is clear that majority of ED models were built for detecting events from Twitter data and official news articles. On the other hand, insufficient ED models were built for Facebook, Blogs and Emails. This has happened due to the accessibility and flexibility of Twitter's Application Programming Interface (API), domain type (i.e., open and specific), detection tasks (i.e., NED and RED) and ED methods i.e., unsupervised, supervised and semi-supervised. It is obvious that, a large number of ED models were open domain type especially, the ones developed for news corpus and such models have employed unsupervised detection methods such as probabilistic, clustering and graph-based detection methods. In addition, it goes without saying that most of proposed models are fall under RED models, which focus on identifying unseen events from various historical corpus. which allows users to easily collect data (Goswami and Kumar, 2016). Different from Twitter, some SNs (e.g., Facebook) have privacy issues that limit the collection process to just offline publicly available data that have been given permissions to be collected (Chen *et al*., 2016). Table 2 shows a summary of ED models based on domain type (i.e., open and specific), detection tasks (i.e., NED and RED) and ED methods i.e., unsupervised, supervised and semi-supervised. It is obvious that, a large number of ED models were open domain type especially, the ones developed for news corpus and such models have employed unsupervised detection methods such as probabilistic, clustering and graph-based detection methods. In addition, it goes without saying that most of proposed models are fall under RED models, which focus on identifying unseen events from various historical corpus.

Table 1: ED models for different text data

Data sources	Studies
Twitter	Becker *et al*. (2011a; Phuvipadawat and Murata, 2010; Sakaki *et al*., 2010; Popescu *et al*., 2011; Ritter *et al*., 2012; Li *et al*., 2012a; Weng and Lee, 2011; Sankaranarayanan *et al*., 2009; Culotta, 2010; Osborne *et al*., 2012; Subašić and Berendt, 2011; Becker *et al*., 2012; Abhik and Toshniwal, 2013; Petrović *et al*., 2010; Ishikawa *et al*., 2012; Mathioudakis and Koudas, 2010; Long *et al*., 2011; Rosa *et al*., 2011; Popescu and Pennacchiotti, 2010; Benson *et al*., 2011; Lee and Sumiya, 2010; Mehrotra *et al*., 2013; Rajani *et al*., 2014; Weng *et al*., 2010; Zhao *et al*., 2011; Diao *et al*., 2012; Dong *et al*., 2015; Fang *et al*., 2014; Cataldi *et al*., 2010; Cordeiro, 2012; Kwan *et al*., 2013; Weiler *et al*., 2014; Aggarwal and Subbian, 2012; Tembhurnikar and Patil, 2015; Katragadda *et al*., 2017; Manaskasemsak *et al*., 2016; Zhang *et al*., 2015; Huang *et al*., 2013; Kaleel *et al*., 2013; Rafea and Mostafa, 2013; Unankard *et al*., 2014; Nguyen *et al*., 2019; GabAllah and Rafea, 2019; Melvin *et al*., 2017)
News Articles	Fung *et al*. (2005; He *et al*., 2007; Zhao *et al*., 2011; Yang *et al*., 1998; 1999; Menon, 2010; Lam *et al*., 2001; Leban *et al*., 2016; Mele and Crestani, 2017; Huang *et al*., 2013; Beigh *et al*., 2016; Mohamad *et al*., 2010; Khatdeo *et al*., 2017; Dai *et al*., 2010; Zhang *et al*., 2016; Yang *et al*., 2018; Rasouli *et al*., 2019; Yu and Wu, 2018; Florence *et al*., 2017; Hu *et al*., 2017; Chen *et al*., 2017; Wei *et al*., 2018; Moutidis and Williams, 2020)
Facebook	Chen *et al*. (2016; Cvijikj and Michahelles, 2011; Kaleel *et al*., 2013; Passaro *et al*., 2016; Salloum, 2017; Salloum *et al*., 2017a; 2017b; Dewan and Kumaraguru, 2015; Al-Rawi, 2016; Duwairi and Alfaqeeh, 2015)
	Sayyadi *et al*. (2009; Huang *et al*., 2013; Vavliakis *et al*., 2013)
E-mail	Zhao and Mitra (2007; Wasi *et al*., 2011; Aggarwal and Subbian, 2012)

Table 2: Summary of ED studies

		Domain		ED Methods			Unsupervised Method					Detection Task	
Year	Author	*Open*	*Specific*	*Un-sup*	*Sup*	*Semi-sup*	*A*	*B*	*C*	*D*	*E*	*NED*	*RED*
2020	(Moutidis and Williams, 2020)		X	X							X		X
2019	(Rasouli *et al*., 2019)	X		X							X	X	X
2019	(Nguyen *et al*., 2019)	X		X						X		X	
2019	(GabAllah and Rafea, 2019)		X	X				X					X
2018	(Wei *et al*., 2018)	X				X					X	X	X
2018	(Yang *et al*., 2018)	X		X							X		X
2018	(Yu and Wu, 2018)	X		X						X			X
2018	(Akachar *et al*., 2018)	X		X						X			X
2018	(Cracs and Porto, 2018)	X		X				X					X
2017	(Chen *et al*., 2017)	X		X							X		X
2017	(Melvin *et al*., 2017)	X		X							X		X
2017	(Katragadda *et al*., 2017)	X		X							X	X	
2017	(Hu *et al*., 2017)	X		X						X		X	
2017	(Florence *et al*., 2017)		X	X						X			X
2017	(Salloum, 2017)	X		X				X					X
2017	(Mele and Crestani, 2017)	X		X					X				X
2016	(Beigh *et al*., 2016)	X		X				X				X	
2016	(Chen *et al*., 2016)	X		X						X			X
2016	(Leban *et al*., 2016)	X		X						X		X	

Continued (**Table 2:** Summary of ED studies)

2016	(Sy *et al*., 2016)		X	X							X		X
2016	(Passaro *et al*., 2016)	X		X							X		X
2016	(Alashri *et al*., 2016)	X		X					X				X
2015	(Dong *et al*., 2015)	X		X				X					X
2015	(Tembhurnikar and Patil, 2015)		X	X						X			X
2015	(Lu *et al*., 2015)		X		X								
2014	(Fang *et al*., 2014)		X	X						X			X
2014	(Rajani *et al*., 2014)	X		X					X				X
2014	(Sowmiya and Chandrakala, 2014)		X	X					X				X
2014	(Leban *et al*., 2014)	X		X								X	
2014	(Weiler *et al*., 2014)	X		X				X				X	
2014	(Zhang *et al*., 2014)	X		X							X		X
2013	(Mehrotra *et al*., 2013)	X		X					X				X
2013	(Kaleel *et al*., 2013)	X		X				X					X
2013	(Parikh, 2013)	X		X						X			X
2013	(Vavliakis *et al*., 2013)	X		X					X				X
2013	(Huang *et al*., 2013)	X		X							X		X
2013	(Nanba *et al*., 2013)	X			X								X
2013	(Abhik and Toshniwal, 2013)		X	X						X			X
2013	(Kwan *et al*., 2013)	X		X							X		X
2013	(Zhang *et al*., 2013)	X		X							X		X
2013	(Wang *et al*., 2013)	X		X							X		X
2013	(Rafea and Mostafa, 2013)	X		X						X			X
2012	(Baldwin *et al*., 2012)	X		X						X		X	
2012	(Wang *et al*., 2012)	X				X							X
2012	(Cordeiro, 2012)	X		X					X				X
2012	(Diao *et al*., 2012)	X		X				X					X
2012	(Ishikawa *et al*., 2012)	X		X						X		X	
2012	(Osborne *et al*., 2012)	X		X					X			X	
2012	(Li *et al*., 2012b)	X		X				X					X
2012	(Ritter *et al*., 2012)	X			X								X
2012	(Aggarwal and Subbian, 2012)	X		X						X			X
2011	(Rosa *et al*., 2011)	X			X								X
2011	(Long *et al*., 2011)	X		X						X			X
2011	(Zhao *et al*., 2011)	X		X					X				X
2011	(Benson *et al*., 2011)		X		X								X
2011	(Subašić and Berendt, 2011)		X	X				X					X
2011	(Ahn *et al*., 2011)	X		X					X				X
2011	(Becker *et al*., 2011b)	X				X	X						X
2011	(Becker *et al*., 2011c)	X		X						X		X	
2011	(Weng and Lee, 2011)	X		X				X					X
2011	(Popescu *et al*., 2011)	X			X								X
2011	(Becker *et al*., 2011b)	X		X			X					X	X
2011	(Cvijikj and Michahelles, 2011)		X	X				X					X
2011	(Motooka *et al*., 2011)	X		X						X			X
2010	(Petrović *et al*., 2010)	X		X				X					X
2010	(Mathioudakis and Koudas, 2010)	X		X				X				X	
2010	(Cataldi *et al*., 2010)	X		X							X		X
2010	(Lee and Sumiya, 2010)		X	X						X			X
2010	(Popescu and Pennacchiotti, 2010)		X		X								X
2010	(Sakaki *et al*., 2010)		X		X								X
2010	(Phuvipadawat and Murata, 2010)	X		X				X					X
2010	(Phuvipadawat and Murata, 2010)		X	X				X					X
2010	(Dai *et al*., 2010)	X		X						X			X
2010	(Mohamad *et al*., 2010)	X		X				X					X
2010	(Newman *et al*., 2010)	X		X					X				X

A: Query-Based Methods, B: Statistical-Based Methods, C: Probabilistic-Based Methods, D: Clustering-Based Methods, E: Graph-Based Methods, Un-sup: Unsupervised Methods, Sup: Supervised Methods, Semi-sup: Semi-supervised Methods

Event Detection Methods

Diverse categorization of ED methods is presented by several authors (Atefeh and Khreich, 2015; Lavanya *et al*., 2014; Panagiotou *et al*., 2016; Zarrinkalam and Bagheri, 2017). The scholars of this paper have adapted the classifications proposed by previous researchers and categorize ED methods into supervised and unsupervised methods as it is shown in Fig. 5. Subsequently, unsupervised methods are further classified into five categories: Query-based method, statistical based methods, probabilistic based methods, clustering based methods, graph-based methods.

Many supervised ED methods have been developed *and presented throughout the time in literature to achieve different objectives (*Ab Aziz *et al*., 2016; Mustaffa *et al*., 2014; Yusof *et al*., 2015). Sakaki *et al*. (2010) applied Support Vector Machine (SVM) classifier to detect earthquake and forecast earthquake's center in real-time from Twitter dataset. Cheong and Cheong (2011) analyzed tweets during natural disaster floods that happened in Australia. Yamanaka *et al*. (2010), introduced a model that detects events based on GPS information for a particular area using SVM. Comparatively, SVM in combination with incremental clustering technique was applied to detect social and real-world events from photos posted on Flicker site (Wang *et al*., 2012).

A decision tree classifier called gradient boosted was used to anticipate weather the tweets consist of an event concerned the target entity or not. Conditional Random Field (CRF) classifier was learned to extract the artist name and location of music events from a corpus of tweets (Benson *et al*., 2011). Likewise, CRF applied to extract full information about events happening in all tourist spots from news articles and web pages (Nanba *et al*., 2013). Moreover, CRF was trained by (Ritter *et al*., 2012) to extract temporal expression about events. Despite the good results that supervised methods have obtained, yet they are time consuming, have high complexity learning, restricted in scope and need a large amount of labelled data to train the classifier (Atefeh and Khreich, 2015). On the other hand, numerous unsupervised methods are introduced by various scientists and which are grouped by the authors of this study into various categories that are described in the following subsections.

Query-Based Methods

Examples of such methods are simple rules and built-in query strategies which were proposed to identify planned events from multiple websites e.g., Twitter, YouTube, Flicker (Becker *et al*., 2012; 2011b; 2011c). In these studies, the authors extracted temporal and spatial information of an event and subsequently, such information was utilized to enquiry other SNs to obtain relevant documents. Their study exhibited that information from one SN can be used to identify related documents from different SNs. However, these methods are limited to specific events and can’t be generalized to all events. On top of that, query based methods always require predefine key words for each event which is not appropriate if there is a large number of events as it requires a substantial amount of time (Ishikawa *et al*., 2012).

Statistical Based Methods

Under this category various methods were introduced by different researchers. For instance, GabAllah and Rafea, (2019) calculated the average frequency of unigrams to find the significant unigrams and consequently, combine those unigrams (keywords) to represent the trending topics. Subašić and Berendt (2011) tried to identify burst features (i.e., unigram) over different time windows in order to detect the hot events. Signals for each unigram feature has computed then converted into frequency domain using Discrete Fourier Transformation (DFT) technique. However, DFT did not identify the time periods when there is a burst which is very important for ED process. Therefore, another technique called Wavelet Transformation (WT) has proposed by Weng and Lee (2011) to assign signals for individual unigram feature. WT technique differs from DFT in terms of localized in both time and frequency domain. Hence, provide better results for ED. Phuvipadawat and Murata (2010) improved ED model by LDA was extensively applied for topic extraction from document streams (Mehrotra *et al*., 2013). Later, it was exploited to discover events from various sources such as Wikipedia pages and their reviews (Osborne *et al*., 2012). Twitter (Rajani *et al*., 2014) and news articles (Mele and Crestani, 2017). Alashri *et al*. (2016) studied Facebook posts using LDA. which published by the candidates of U.S 2016 Presidential Election to identify the significant events. Cordeiro (2012) combined Continues Wavelet Transformation (CWT) analysis and LDA into one ED model. Mixed models of ED and sentiments were introduced in (Passaro *et al*., 2016). Vavliakis *et al*. (2013) proposed a framework which consists of different integrated unsupervised techniques. For instance, LDA, NER, AGH, bipartite graph clustering algorithm based on betweenness centrality scores to identify hidden events and extract their important information such as time, location and people that have been involved.

Increasing weights for the proper nouns features that were identified by Named Entity Relation (NER). Li *et al*. (2012a) first applied tweet segmentation to get phrases consists of one or more serial words rather than using unigrams. Later, they computed TFIDF of these phrases and user frequency and classified them using K-Nearest Neighbor (KNN) to identify the events from tweets published by Singapore users.

Mohamad *et al*. (2010) extracted keywords from a set of news articles to find the similar articles which share identical keywords to group them into one cluster. Cvijikj and Michahelles (2011) classified public Facebook posts into three trending topics using clustering by distribution as well clustering by co-occurrence. Comparatively, Facebook news posts published by (16) English news channels were analyzed to identify the most frequent linked terms among different news channels by (Salloum *et al*., 2017b). Weiler *et al*. (2014) used shifts of terms computed by Inverse Document Frequency (IDF) over simple sliding window model to detect events and trace their evolution. In same fashion, (Beigh *et al*., 2016) divided the news streams into different time sliding windows and select burst features whose frequencies above specific thresholds. However, selection of burst features based on statistics can generate a huge number of features, especially when unigrams are used. Besides that, defining events using single terms is not sufficient and difficult to understand by human (Mele and Crestani, 2017). Moreover, specifying an appropriate threshold to select the burst features as well as determining the size of time window are stated to be challenging tasks (Li *et al*., 2012b).

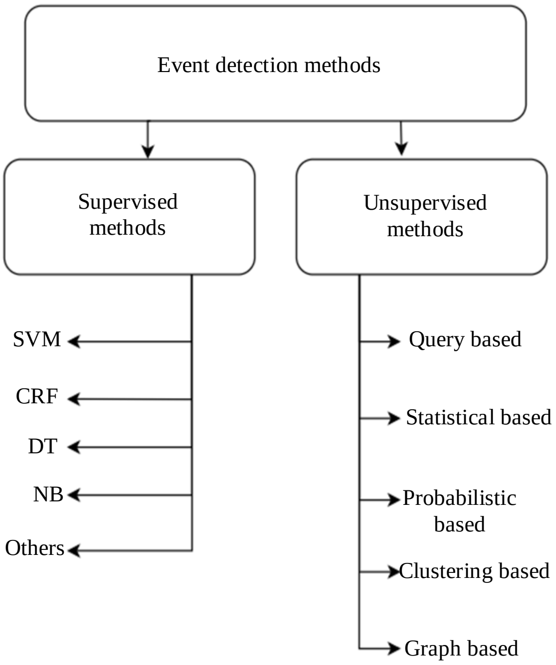


Fig. 5: Classification of ED methods

Locality Sensitive Hashing (LSH) was modified and used by (Petrović *et al*., 2010) to perform First Story Detection (FSD) task on Twitter data stream. Meanwhile, (Kaleel *et al*., 2013) applied LSH first to identify events from each SNs individually then implemented detect cross-over events for the both SNs. However, parameters of LSH are required to be set in advance by users and basic LSH produce high variance results and performs poorly for FSD task (Nurwidyantoro and Winarko, 2013). Furthermore, LSH is a randomized technique and error can occur. Therefore, LSH was applied multiple times recently to reduce the error rate, but this leads to increase the computational time (Petrović *et al*., 2010).

Probabilistic Based Methods

ED has been attempted utilizing topic modelling methods such as Latent Dirichlet Allocation (LDA) and Probabilistic Latent Semantic Indexing (PLSI), etc. To begin with, in LDA each document is presented as a mixture of different topics where each document is assumed to have a set of topics that are allocated to it via LDA Fig. 6.

Though LDA was good in discovering events from academic abstracts and news articles, but it has not worked perfectly with short text documents (Mehrotra *et al*., 2013). Therefore, (Mehrotra *et al*., 2013) improved LDA model through tweet pooling schemes and automatic labelling. Pooling schemes such as basic scheme, author scheme (i.e., tweets published by the same author), burst terms scheme (i.e., aggregate tweets that share common burst terms over a particular time window), temporal scheme (i.e., tweet generated in the same time window) and hashtag scheme tweets published under same hashtag. The experiments results proved that hashtag scheme produced the best clusters results. In the same context, other authors solved the problem of identifying events from short and long text documents by developing a method that incorporates LDA, NER and temporal analysis of burst features (Mele and Crestani, 2017). The results demonstrate better performance and obtained high precision clusters. However, LDA has a problem of specifying the number of topics and number of terms per topic in advance which can be less effective and difficult to determine when implementing it over SNs contents (Vavliakis *et al*., 2013; Zarrinkalam and Bagheri, 2017).

Clustering-Based Methods

In contrast to supervised methods that need labelled data in order to predict the events in the future, this type of methods do not require such data, but rather they rely heavily on the process of selection the most informative features which contribute in detecting events more accurately Fig. 7. In literature, many clustering algorithms have been employed over various text data streams. However, the most famous one is K-means clustering algorithm (Chen *et al*., 2016; Tembhurnikar and Patil, 2015; Yu and Wu, 2018). Yu and Wu (2018) propose a novel dual-level clustering model based on news representation with time2vec to detect events from chines news articles. Tembhurnikar and Patil (2015) applied K-means to identify the clusters and implemented Agglomerative Hierarchal Clustering (AHC) technique to merge the clusters. Similarly, (Florence *et al*., 2017) applied constrained hierarchal k-means to identify 10 specific categories of events through utilizing meta data associated with the news articles such as temporal information, geographical data, name of people and organizations. Dai *et al*. (2010) employed two layers clustering method over chinses news articles to solve overlapping features between events. AHC was used to detect events from Tweets by (Parikh and Karlapalem, 2013). Rafea and Mostafa (2013) applied bisecting k-mean algorithm to identify Arabic hot events from Twitter. Hu *et al*. (2017) proposed novel document representation method based on word embed- dings to reduce dimensionality feature space and applied new adaptive single pass clustering method for online news event detection.

Named entities and central centroids in incremental clustering algorithm were used to group the most similar tweets from Events 2012 dataset (Nguyen *et al*., 2019). Leban *et al*. (2016) constructed event registry system to perform online clustering in order to group news articles into different events and used NER with TFIDF to extract core information like location, date, what has happened and who was involved. A model based on K-means was utilized to detect local festivals events from geotagged tweets (Lee and Sumiya, 2010). In addition to previous clustering methods, many researchers have incorporated various platform's features into clustering methods to improve the performance of ED model. For example, (Ishikawa *et al*., 2012) identified spot hot topics for local area from tweets using used geotags and temporal features. Aggarwal and Subbian (2012) designed a method based on temporal and structural contents (e.g., user's interaction) to identify the events from SNs streams. A monitoring model was introduced based on a comprehensive location dictionary and historical tweets that were generated by different users to fill up missing location information (Zhang *et al*., 2015).

Furthermore, hashtags were exploited by different authors to group tweets into different real-world events (Long *et al*., 2011). Recently, a framework called multi-views topic detection was proposed by (Rajani *et al*., 2014) to identify hot topics from Twitter streams by incorporating social relations, reply and retweet relations, temporal relations, hashtag and geotag. However, a single tweet can have multiple hashtags and usually the number of tweets associated with a hashtag is relatively very small compared to the huge volume of tweets published per day (Ishikawa *et al*., 2012). Besides that, platform's features for single SN are huge and complex and employing such features has been always a major challenge for many scientists (Kwan *et al*., 2013). Moreover, such features could be missing or not trustworthy (Goswami and Kumar, 2016). On top of all that, every clustering technique has its own drawbacks. For instance, K-means is extremely sensible to the process of setting up its parameters like the number (k) clusters as well as determining the initial locations of its centroids and dealing with the falling into local optima solution (Mohammed *et al*., 2016). In contrast, it is very difficult to determine when to stop the combining or splitting process for hieratical clustering techniques which are static i.e., objects within one cluster cannot move to another cluster. Hence, lead to a poor performance, especially when the separation of overlapping clusters is existed (Jensi and Jiji, 2014). Despite the existence of many clustering techniques that have been used to detect events, yet there is no single ideal clustering technique is found (Shukla and Naganna, 2014).

Graph-Based Clustering Methods

ED has been explored also through analyzing graphs, which is known as graph clustering method, also known as community detection methods. In general, a graph is composed of a set of nodes\vertices which represent entities and a set of edges\links that represent relationships between nodes (Saritha, 2019). Valuable information can be extracted from these graphs through grouping a set of nodes based on the set of edges. Each generated group form what is called a cluster\graph structure or also known as community, cluster or module Fig. 8. The links between different nodes are called intra-edges, meanwhile links that connect different communities are called inter-edges (Fortunato, 2010). In literature, many graph clustering algorithms have been proposed. Sayyadi *et al.* (2009) applied cut-off technique applied for large graphs while basic score was implemented for small once. Cataldi *et al.* (2010) identified the emergence topics from directed weighted graph that was constructed from the tweets generated by active users. However, the authors have not applied pre-processing steps or FS technique to reduce the high dimensionality of data. In addition, the used cut-off graph technique has not assured to identify ideal emerging topics as it depends on a threshold defined by a user.

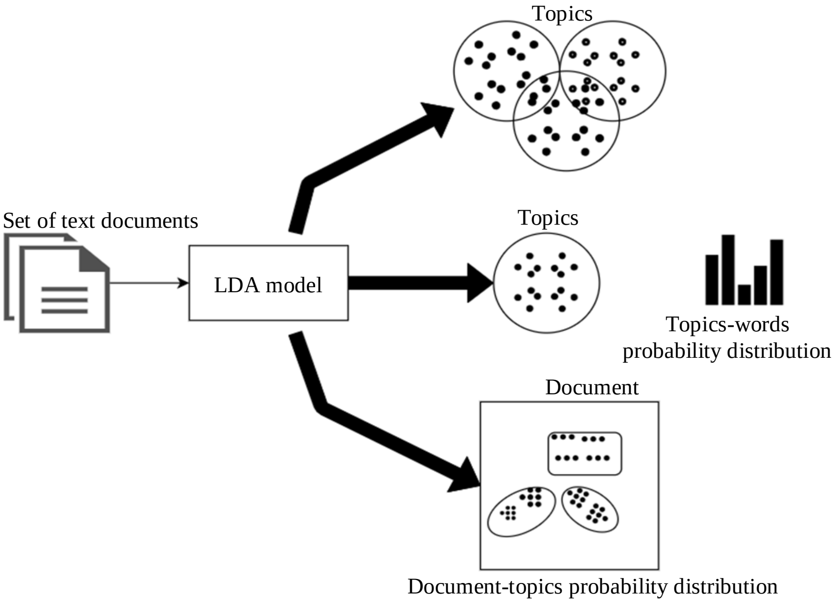


Fig. 6: Topic modeling LDA

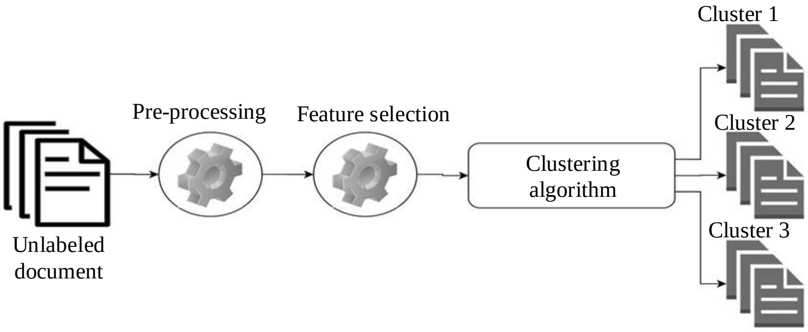


Fig. 7: Clustering based methods

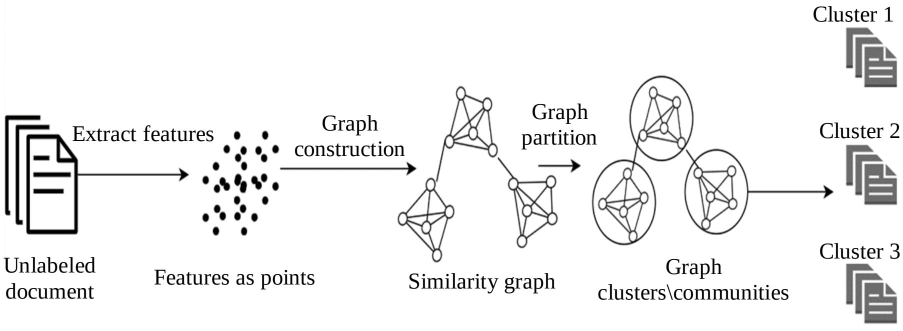


Fig. 8: Graph based clustering methods

Zhang *et al.* (2016) identified the significant events from news articles and visualize them graphically to improve human recognition. Unfortunately, their method was computationally expensive and has not addressed the problem of obtaining event's temporal, spatial information. On top of that, it was implemented over a relatively very few documents which contains a long text content. Quite similar, (Wei *et al*., 2018) built a comprehensive framework which employs a domain dictionary and location ontology to detect overlapping and specific events using news articles collected from a large, noisy news corpus generated from various news sources. However, this framework ignored the high dimensionality of feature space generated from merging a large volume of news articles published by different sources. Moutidis and Williams (2020) identified events through finding peaks within entity knowledge graph and summarizing of events was done by applied community detection method on KeyGraph that linking noun-phrases and entities. This study was implemented on small size of manually annotated dataset.

Similarly, (Sy *et al*., 2016) recognized the events from graphs and then grouped them using the proposed AHC technique that was proposed in (Fortunato, 2010). Later, the authors applied pruning technique to filter out the clusters with higher co-occurrence frequency. This method has the issue of defining a proper threshold for purring, where various values can lead to different results. Kwan *et al*. (2013) built a directed weighted graph from the keywords that were extracted from different time windows. Later, they implemented cut off technique to identify three types of events e.g., one shot, long run and non-events. However, many thresholds must be defined in advance by the users for filtering and selecting candidate keywords. In addition, the precision percentage that has obtained was not high and the model has not given a clear summary for the discovered events. Manaskasemsak *et al*. (2016) divided Twitter stream into two-time windows of 15 days. Subsequently, Markov Clustering Algorithm (MCA) was employed to detect events from the undirected weighted graph which has built from features extracted using TFIDF. However, MCA suffers from generating many tiny and scattered clusters which could represent meaningless events. Moreover, researchers had to define a threshold in advance to select the top ranked events from each generated cluster. Katragadda *et al*. (2017) proposed a model that effectively merged the information from Twitter and Tumblr in three styles; during graph generation, after graph pruning and combine post-clustering. This study confirmed that utilizing information from different data sources improves the task of ED. However, the proposed model has high computational time and it has not implemented pre-processing steps, even though the contents of SNs characterized by the existence of noisy data.

Recently, Rasouli *et al*. (2019) has applied betweenness centrality community detection method on Weighted Bursty KeyGraph to identify events from web news documents in both RED and NED style. The authors have introduced improved feature selection technique and adopted some graph sampling technique in order to reduce the size of features and graph, respectively. However, their proposed model has many parameters that required to be set up in advance. Melvin *et al*. (2017) introduced ED model called phrase network, which detect events using Louvain graph detection method and provide summary about the detected events using high peak phrases. However, their model has many thresholds than required to be define in advance and they ignore the high dimensionality of feature space. Same detection method was used by (Chen *et al*., 2017) to detect topics from semantic graph constructed using wordnet corpus. However, the proposed model is not suitable for the high dynamic volume of generated text data from SNs. In summary, despite the popularity of unsupervised methods in that they do not require labelled data, yet it is still a challenging task to design unsupervised method which is able to deal with high dimensionality of text data streams (Parikh and Karlapalem, 2013). Table 3 summarizes the main limitations of existing ED methods that have been used to detect events from text data which was generated by different SNs.

A lot of efforts have been done to enhance or develop new ED models in order to overcome various problems which have a direct impact on the performance of the methods and techniques used in various phases of ED model (Tembhurnikar and Patil, 2015). In general, ED model consists of six main phases as it is shown in Fig. 9.

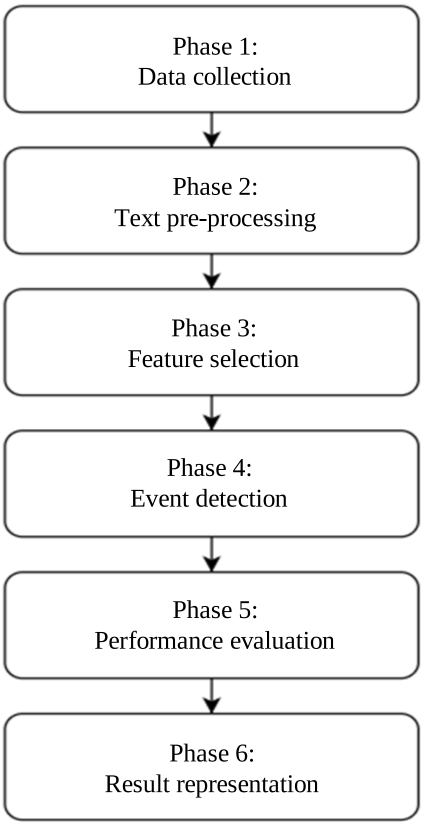


Fig. 9: Main phases of ED model

Table 3: Limitations of ED methods

	ED Method	Problems
	Supervised methods	• Require a large amount of labelled data. • Take long time to learn. • Usually limited to specific domain.
Unsupervised methods	Query-based techniques	• Require a predefined key word for each event
	Statistical-based techniques	• Generate a huge number of features. • The detected events usually represent using a set of single terms (unigrams) which may not be not sufficient and hard for human recognition • Difficult to choose the appropriate thresholds to select the burst features
	Probabilistic-based techniques	• Do not work perfectly with short text documents. • Require specifying the number of topics as well as the number of terms per topic in advance
	Clustering-based techniques	• Partitioning clustering (e.g., K-means): Difficult to determine the number of (k) clusters, difficult to specify the initial locations for the centroids, fall into local optima solutions • It is very difficult to determine when to stop the process of combining or splitting clusters for HCAs. • HCAs are static i.e., objects within one cluster cannot move to another cluster
	Graph-based techniques	• Majority of them have applied on static and small-scale graphs with just few numbers of nodes and edges. • Most of the techniques have applied over graphs that were constructed from noisy, sparsity and high dimensionality feature space.

These phases are data collection phase, pre-processing phase, Feature Selection (FS) phase, ED phase, evaluation phase, result representation phase which consists of visualization, summarization or evolution (Ramadan and Mohd, 2011; Tembhurnikar and Patil, 2015). Regardless of all the existing ED models in literature, yet several challenges are still existed that need to be addressed further. In the following sub-sections, a detail description and discussion about key challenges that faced researchers from ED field in building ED models for text data from different SNs is presented.

Main Challenges Involved in Event Detection Models

Designing high accurate detection model is a challenging task (Al-Dyani *et al*., 2018, Goswami and Kumar, 2016; Nurwidyantoro and Winarko, 2013; Zarrinkalam and Bagheri, 2017). Especially, for a critical area like news, health, politics and finance. Whereby, if there is a miner mistake in the detection process, this could make power holders to take wrong decisions (Abdullah *et al*., 2012). The accuracy of ED model and the quality of results depend primarily on the performance of the methods which are employed under each phase of ED model (Goswami and Kumar, 2016). In the following subsections major challenges that affect the process of building ED models for text data from SNs is described briefly.

Writing Style and Noise Contents

Usually, data on SNs are written informally (Nurwidyantoro and Winarko, 2013), which is opposite of well formatted and high quality writing documents like news articles and academic articles (Zarrinkalam and Bagheri, 2017). Informal documents contain large number of misspelling, grammar errors, slang language, irregular abbreviations, mixed languages and improper sentences (Deng *et al*., 2015; Zarrinkalam and Bagheri, 2017). Thus, current ED models should be enhanced to handle such kind of contents (Goswami and Kumar, 2016; Zarrinkalam and Bagheri, 2017). In addition, data on SNs usually contain different types of noisy contents like spam messages, advertisements, hoaxes, internet memes, URLs and disambiguation semantic (Panagiotou *et al*., 2016). In the same context, news posts which are published on SNs by different news channels also include noisy data like URLs, meaningless terms, duplicated posts and empty posts. Moreover, unlike official news articles, not all news posts published on SNs represent an event, whereby, there are various non-related event posts. For instance, posts include opinions (i.e., writes by variety of people), questions (i.e., writes by page's manger to ask news readers), posts contain links or very few words i.e., maximum three words that have no meaning and do not indicate anything. To eliminate such posts, researchers first remove them using specific techniques or classify the posts into event or non-event posts using machine learning classifiers (Zarrinkalam and Bagheri, 2017).

Build Open Domain Event Detection model

ED model that is suitable for one domain might not be applicable for other domains. For instance, events of political election differ from the events of any natural disaster incident like earthquake. That is because every domain has its unique parameters, variables and metrics. Therefore, building open domain ED models becomes a challenging task as it includes events from different fields. Example of such models is ED model for news data, whereby this data covers events from different domains such as politics, natural disaster, airplane crash, conflict, sports, education and so forth. However, build open domain ED model for news data that is published by several sources is very challenging task (Beigh *et al*., 2016; Chen *et al*., 2016; Garg and Kumar, 2016; Goswami and Kumar, 2016; Ramadan and Mohd, 2011; Zarrinkalam and Bagheri, 2017; Zhou *et al*., 2017). This is due to the high dimensional data which includes irrelevant, duplicated and noisy features that eventually, decrease the overall accuracy of ED model (Allahyari *et al*., 2017; Beigh *et al*., 2016; Bharti and Singh, 2016; Chen *et al*., 2016; Panagiotou *et al*., 2016). Despite the existence of such challenges, yet it has become a hot research topic in recent years and has motivated several researchers to develop open domain models for news data (Goswami and Kumar, 2016).

Short Text Issues

Contents on SNs are characterized to be short text documents (Nurwidyantoro and Winarko, 2013). Consequently, it introduces new challenges for the traditional text mining and Natural Language Processing (NLP) methods (Panagiotou *et al*., 2016; Zarrinkalam and Bagheri, 2017). For instance, they don't provide sufficient information about an event like location, people and activity (Deng *et al*., 2015; Zarrinkalam and Bagheri, 2017). Such information is necessary to answer questions related to an event like what has happened, when, where and who were involved (Panagiotou *et al*., 2016). Additionally, they do not provide enough statistical information for calculating the similarity between two documents (Deng *et al*., 2015). Moreover, majority of existing ED models have implemented on long text documents and when same models will applied over short text length documents, they might perform poorly (Zarrinkalam and Bagheri, 2017). To solve such problems, several studies have aggregated multiple short text posts to generate one single document and identify events through applying LDA over this document. The aggregation was based on different basics like tweets published by the same user or tweets that share same location tag. However, aggregation method could be less effective on other SNs platforms as they have different structures and features. For such reason, other methods are required to be more researched to overcome the issue of the short text length for ED model.

Feature Selection

Accuracy of ED model is extremely influenced by the high dimensionality of feature space which includes various type of features such as redundant, irrelevant and noisy features (Beigh *et al*., 2016). Such features increase the computational complexity of the underline mining algorithms/techniques used for various phases of the ED model. As a result, the overall accuracy of the detection model is decreased(Al-Dyani *et al*., 2018, Allahyari *et al*., 2017). Most current ED models have used only single Feature Selection technique (FS) for FS phase such as Term Frequency Inverse Document Frequency (TFIDF) (Figueira, 2018; Salloum, 2017; Salloum *et al*., 2017a; 2017b), Term Frequency (TF) (Passaro *et al*., 2016; Salloum, 2017; Salloum *et al*., 2017a; 2017b). However, applying single technique has proved by several researchers from text mining field to be insufficient to remove all unnecessary features and select the optimal ones (Allahyari *et al*., 2017; Bharti and Singh, 2014a; 2015; 2016; Harish and Revanasiddappa, 2017; Kumar and Minz, 2014). Therefore, recently in the context of text mining tasks, several dimension reduction methods have been utilized such as filter, wrapper, embedded and hybrid methods (Bharti and Singh, 2014b; Jeyaraj, 2018: Kashef and Nezamabadi-Pour, 2014). Hybrid filter-wrapper methods have proved to achieve best results due to their good balance between the computational efficiency of a filtering techniques and the high accuracy performance of the wrapper techniques (Alsaeedi *et al*., 2017; Bharti and Singh, 2014a; 2014b; 2015; 2016; Dastider *et al*., 2015; Taha *et al*., 2015). For the wrapper part, different traditional searching strategies have been utilized such as exponential, sequential and random search (El Aboudi and Benhlima, 2016). However, such strategies have the problem of nesting effect in which the selected features cannot be discarded, or deleted features cannot be reselected (Srividhya and Mallika, 2018). As a consequence, different Meta-Heuristic Algorithms (MHA) have been employed recently as search strategy and have achieved promising results because of their powerful global search ability in exploring the feature space more effectively and efficiently (Arora and Anand, 2019; Bharti and Singh, 2016; Uğuz, 2011; Xue *et al*., 2016). Given such advantages of MHAs, there is a need to develop FS method based on any MHA to solve the problem of high dimensional feature space of ED model (Panagiotou *et al*., 2016).

Design Unsupervised Event Detection Method

Traditional ED methods like query-based and key extraction-based methods have been extensively used to detect real-world events (Goswami and Kumar, 2016). Unfortunately, such methods assume that number of events is already known in advance, which is not applicable in real-time situation where events usually happen without any prior knowledge or prediction (Panagiotou *et al*., 2016). Hence, a model based on unsupervised ED method is required to detect the hidden events through investigating the textual features of the corpus (Goswami and Kumar, 2016). In the same context, many scientists treat ED problem as text clustering problem and therefore, different clustering methods have been proposed to identify the events (Huang *et al*., 2016). On the other hand, researchers have gone towards optimizing clustering method's performance through integrating them with various optimization algorithms (Huang *et al*., 2016). In practice, performance of a clustering method depends mainly on its capacity to handle the high dimensionality feature space as well as on the selection of most appropriate evaluation measurements (Uğuz, 2011). All together make the process of constructing unsupervised method for ED phase becomes more challenging task (Panagiotou *et al*., 2016).

Exploiting Features of Platforms

Existing ED models have utilized only text contents that may lead to incorrect detection of events (Zhou and Chen, 2014). As a result, ED models are required to take advantage of some available features on SNs. Whereby, it is reported that features such as tags, timestamps, links, meta-data, user's involvement equally important in improving the detection accuracy of the given ED model (Huang *et al*., 2016). For instance, set of posts that share same location tags indicate that they are talking about the same event. Hashtag "#" is used to detect the hot topics/events in Twitter through identifying the related tweets. Recently, hashtag has been also introduced into Facebook platform. In addition, number of followers in Twitter and number of friends in Facebook as well as number of engagements can be used to identify the most interesting posts among users (Uğuz, 2011). Moreover, user involvements (e.g., retweets, comments and relations among users) can contribute in detecting events more effectively (Atefeh and Khreich, 2015). For instance, more information about an event (e.g., earthquake) can be obtained through exploiting user's posting comments and retweets, which hold valuable information such as proper names, leading phrases that describe the amount of damage or mention the number of injuries and missing people, etc. Due to all above features, it is recommended to build ED models that are able to incorporate such features to improve the performance of ED model (Goswami and Kumar, 2016).

Evaluation Problem

Evaluation of ED model may differ according to the used dataset either labelled or unlabeled dataset (Zarrinkalam and Bagheri, 2017). In literature, many researchers used benchmark annotated dataset named TDT5 to evaluate their proposed models in terms of Precision (Karkali *et al*., 2013; Petrović *et al*., 2010). However, this data differs from the primary datasets that were extracted from SNs in terms of their nature and length, thus, the obtained results are significantly varied. In addition, using the same dataset TDT5 to evaluate different models would lead to repeat the experiments and make the corresponding studies to become just comparative studies that compare different ED models (Panagiotou *et al*., 2016; Zarrinkalam and Bagheri, 2017). To overcome such problems, several scientists have gone toward employing query-based technique over different sources (e.g., news feeds, search engines and Wikipedia pages) in order to collect and annotate the corpus at the same time. However, such techniques require predefined keywords which rise another issue, since events, in real-life usually happen suddenly without any warning. Under those circumstances, some scholars have resorted to manually annotated their collected datasets with very few numbers of events (McMinn *et al*., 2013). The annotation process is usually done through hiring more than one annotator and then measure the rate of their agreement using different measurement tool such as Cohen's Kappa (Wood, 2007). Subsequently, just high rate agreements annotations are granted, meanwhile agreements with low rate are removed. For that reason, it is become necessary to design or improve an existing annotation method to label the primary datasets automatically with less human interfere.

Summarize Information of the Detected Events

It is very important and necessary to summarize and present information regarding the discovered events properly so that readers can have a complete picture about an event (Goswami and Kumar, 2016; Nurwidyantoro and Winarko, 2013). Event summarization is defined as "creating a summary of events based on burst features identified from a text corpus", (Dou *et al*., 2012a). Several studies have tried to address this task through utilizing both textual contents (e.g., using burst features) and meta-information (e.g., hashtags, date and geolocations), sentiments and named entities, etc. However, identifying hidden semantics information for an event from the textual contents is considered to be a difficult task (Dou *et al*., 2012b). Besides that, organizing the obtained information and present it in a comprehensive summary is still a challenging to deal with (Goswami and Kumar, 2016; Nurwidyantoro and Winarko, 2013). Moreover, the summary of an event might be affected by the event's evolution over time. Therefore, changing track of the event should be considered while creating a summary of the event (Goswami and Kumar, 2016).

Correctly Identifying Evolution of Events

Event evolution is defined as "how event unfold, or how to track the development of an event over time and extract important information to support situational awareness during crisis and inform public polices", (Dou *et al*., 2012b). Event evolution is recognized through spotting the semantic shifting of features over various time windows (Goswami and Kumar, 2016). Existing studies have used cosine similarity and time window analysis that investigate the semantic shift of event's keywords. However, such methods are not optimal as they sometimes cannot corporate all available information. Consequently, enhanced methods are required; which are capable to incorporate offered data to identify event evolution correctly and provide the users with a complete picture of an event over timeline.

Conclusion

The rapid spread of information on different SNs in the form of text has encouraged many researchers to investigate and extract important information about various real-world events. whereby, such information help decision makers in different disciplines to make the right decisions in various situations. In this study, a comprehensive review of recent ED studies for text data from different SNs is presented. In particular, different categories of unsupervised ED methods are demonstrated and their related works are reviewed and investigated. In addition, key open challenges involved in building ED models are explained and discussed in order to assist scholars to figure out the gaps in literature. All together can provide guidelines for scientists from ED field to better understand the awareness about ED methods and their limitations to enhance them or develop a new one in order to achieve high accuracy detection of ED model.

Acknowledgement

Authors would like to thank Universiti Utara Malaysia for their financial support, editors and anonymous reviewers for valuable comments.

Author's Contributions

Wafa Zubair Al-Dyani: Analyzed the articles and wrote the manuscript.

Farzana Kabir Ahmad: Provided proper guidelines to prepare the paper and shared ideas.

Siti Sakira Kamaruddin: Monitored and defined work area, read and approve the final manuscript.

Ethics

Authors declare that there are no ethical issues in the paper.

References

Ab Aziz, A., Ahmad, F., Yusof, N., Ahmad, F. K., and Yusof, S. A. M. (2016). Designing a robot-assisted therapy for individuals with anxiety traits and states. 2015 International Symposium on Agents, Multi-Agent Systems and Robotics (ISAMSR), 98–103.

Abdullah, F., Ku-Mahamud, K. R., Ahmad, F., A Ghani, N. F., and M Kasim, M. (2012). Relative efficiency assessment of projects using data envelopment analysis: A case study. International Journal of Digital Content Technology and Its Applications, 6(9), 310–318.

Abhik, D., and Toshniwal, D. (2013). Sub-event detection during natural hazards using features of social media data. Proceedings of the 22nd International Conference on World Wide Web, 783–788.

Aggarwal, C. C., and Subbian, K. (2012). Event detection in social streams. Proceedings of the 2012 SIAM International Conference on Data Mining, 624–635.

Al-Rawi, A, (2016). News values on social media: News organizations' Facebook use. Journalism, 1–19. https://doi.org/10.1177/1464884916636142

Ahn, B. G., Van Durme, B., and Callison-Burch, C. (2011). WikiTopics: What is popular on Wikipedia and why. Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, 33–40.

Akachar, E., Ouhbi, B., and Frikh, B. (2018). Community detection in social networks using structural and content information. Proceedings of the 20th International Conference on Information Integration and Web-Based Applications & Services, 282–288.

Al-Dyani, W. Z., Ahmad, F. K., and Kamaruddin, S. S. (2019). Event detection model for Facebook news posts. International Journal of Innovative Technology and Exploring Engineering, 9(1), 98–102. https://doi.org/10.35940/ijitee.A3930.119119

Al-Dyani, W. Z., Yahya, A. H., and Ahmad, F. K. (2018). Challenges of event detection from social media streams. International Journal of Engineering & Technology, 7(2.15), 72–75.

Alashri, S., Kandala, S. S., Bajaj, V., Ravi, R., Smith, K. L., and Desouza, K. C. (2016). An analysis of sentiments on facebook during the 2016 US presidential election. Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference On, 795–802.

Alkubaisi, G. A. A. J., Kamaruddin, S. S., and Husni, H. (2018). Conceptual framework for stock market classification model using sentiment analysis on twitter based on Hybrid Naïve Bayes Classifiers. International Journal of Engineering & Technology, 7(2.14), 57–61.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., and Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. ArXiv Preprint ArXiv:1707.02919.

Alsaeedi, A., Fattah, M. A., and Aloufi, K. (2017). A hybrid feature selection model for text clustering. System Engineering and Technology (ICSET), 2017 7th IEEE International Conference On, 7–11.

Arora, S., and Anand, P. (2019). Binary butterfly optimization approaches for feature selection. Expert Systems with Applications, 116, 147–160.

Atefeh, F., and Khreich, W. (2015). A survey of techniques for event detection in Twitter. Computational Intelligence, 31(1), 133–164. https://doi.org/10.1111/coin.12017

Baldwin, T., Cook, P., Han, B., Harwood, A., Karunasekera, S., and Moshtaghi, M. (2012). A support platform for event detection using social intelligence. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, 69–72.

Becker, H., Naaman, M., and Gravano, L. (2011a). Beyond trending topics: Real-world event identification on Twitter. Icwsm, 11(2011), 1–17. https://doi.org/10.1.1.221.2822

Becker, H., Naaman, M., and Gravano, L. (2011b). Selecting Quality Twitter Content for Events. ICWSM, 11.

Becker, H., Chen, F., Iter, D., Naaman, M., and Gravano, L. (2011c). Automatic Identification and Presentation of Twitter Content for Planned Events. ICWSM.

Becker, H., Iter, D., Naaman, M., and Gravano, L. (2012). Identifying content for planned events across social media sites. Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, 533–542.

Beigh, T. M., Upadhyaya, S., and Gopal, G. (2016). Event Identification in Social News Streams Using Keyword Analysis. International Research Journal of Engineering and Technology (IRJET), 3(5).

Benson, E., Haghighi, A., and Barzilay, R. (2011). Event discovery in social media feeds. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, 389–398.

Bharti, K. K., and kumar Singh, P. (2014). A survey on filter techniques for feature selection in text mining. Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012, 1545–1559.

Bharti, K. K., and Singh, P. K. (2014). A three-stage unsupervised dimension reduction method for text clustering. Journal of Computational Science, 5(2), 156–169.

Bharti, K. K., and Singh, P. K. (2015). Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering. Expert Systems with Applications, 42(6), 3105–3114.

Bharti, K. K., and Singh, P. K. (2016). Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering. Applied Soft Computing, 43, 20–34.

Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77–84.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. Proceedings of the Tenth International Workshop on Multimedia Data Mining, 4.

Chandran, T. R., Reddy, A. V, and Janet, B. (2017). Text Clustering Quality Improvement using a hybrid Social spider optimization. International Journal of Applied Engineering Research, 12(6), 995–1008.

Chen, H.-P., Hsu, K.-W., and Chiu, S.-I. (2016). Event Detection in an Ego Network on Facebook. Pacific Asia Conference on Information Systems, PACIS 2016 - Proceedings.

Chen, Q., Guo, X., and Bai, H. (2017). Semantic-based topic detection using Markov decision processes. Neurocomputing, 242, 40–50.

Cheong, F., and Cheong, C. (2011). Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods. PACIS, 11, 46.

Cordeiro, M. (2012). Twitter event detection: combining wavelet analysis and topic inference summarization. Doctoral Symposium on Informatics Engineering, 11–16.

Cracs, C. S., and Porto, P. (2018). A Three-Step Data-Mining Analysis of Top-Ranked Higher Education Institutions ’ Communication on Facebook.

Culotta, A. (2010). Towards detecting influenza epidemics by analyzing Twitter messages. Proceedings of the First Workshop on Social Media Analytics, 115–122.

Cvijikj, I. P., and Michahelles, F. (2011). Monitoring trends on Facebook. Proceedings - IEEE 9th International Conference on Dependable, Autonomic and Secure Computing, DASC 2011, 895–902. https://doi.org/10.1109/DASC.2011.150

Dai, X., He, Y., and Sun, Y. (2010). A Two-layer text clustering approach for retrospective news event detection. Artificial Intelligence and Computational Intelligence (AICI), 2010 International Conference On, 1, 364–368.

Dastider, S. G., Kashyap, H., Mandal, S., Ghosh, A., and Goswami, S. (2015). Feature Subset Selection for Clustering using Binary Particle Swarm Optimization. Information Technology (ICIT), 2015 International Conference On, 159–164.

Deng, J., Qiao, F., Li, H., Zhang, X., and Wang, H. (2015). An Overview of Event Extraction from Twitter. Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), 2015 International Conference On, 251–256.

Dewan, P., and Kumaraguru, P. (2015). Towards automatic real time identification of malicious posts on Facebook. Privacy, Security and Trust (PST), 2015 13th Annual Conference On, 85–92.

Diao, Q., Jiang, J., Zhu, F., and Lim, E.-P. (2012). Finding bursty topics from microblogs. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1, 536–544.

Dong, X., Mavroeidis, D., Calabrese, F., and Frossard, P. (2015). Multiscale event detection in social media. Data Mining and Knowledge Discovery, 29(5), 1374–1405.

Dou, W., Wang, X., Ribarsky, W., and Zhou, M. (2012a). Event detection in social media data. IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content, 971–980.

Dou, W., Wang, X., Skau, D., Ribarsky, W., and Zhou, M. X. (2012b). Leadline: Interactive visual analysis of text data through event identification and exploration. Visual Analytics Science and Technology (VAST), 2012 IEEE Conference On, 93–102.

Duwairi, R. M., and Alfaqeeh, M. (2015). RUM Extractor: A Facebook Extractor for Data Analysis. Future Internet of Things and Cloud (FiCloud), 2015 3rd International Conference On, 709–713.

El Aboudi, N., and Benhlima, L. (2016). Review on wrapper feature selection approaches. 2016 International Conference on Engineering & MIS (ICEMIS), 1–5.

Fang, Y., Zhang, H., Ye, Y., and Li, X. (2014). Detecting hot topics from Twitter: A multiview approach. Journal of Information Science, 40(5), 578–593.

Florence, R., Nogueira, B., and Marcacini, R. (2017). Constrained hierarchical clustering for news events. Proceedings of the 21st International Database Engineering & Applications Symposium, 49–56.

Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3–5), 75–174.

Fu, Z., Sun, X., Shu, J., and Zhou, L. (2014). Plain Text Zero Knowledge Watermarking Detection Based on Asymmetric Encryption. 48(Cia), 126–134.

Fung, G. P. C., Yu, J. X., Yu, P. S., and Lu, H. (2005). Parameter free bursty events detection in text streams. Proceedings of the 31st International Conference on Very Large Data Bases, 181–192.

GabAllah, N. A., and Rafea, A. (2019). Unsupervised Topic Extraction from Twitter: A Feature-pivot Approach.

Garg, M., and Kumar, M. (2016). Review on event detection techniques in social multimedia. Online Information Review, 40(3), 347–361.

Goswami, A., and Kumar, A. (2016). A survey of event detection techniques in online social networks. Social Network Analysis and Mining, 6(1), 107.

Harish, B. S., and Revanasiddappa, M. B. (2017). A Comprehensive Survey on various Feature Selection Methods to Categorize Text Documents. International Journal of Computer Applications, 164(8).

Hasan, M., Orgun, M. A., and Schwitter, R. (2018). A survey on real-time event detection from the twitter data stream. Journal of Information Science, 44(4), 443–463.

He, Q., Chang, K., and Lim, E.-P. (2007). Analyzing feature trajectories for event detection. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 207–214.

Hu, L., Zhang, B., Hou, L., and Li, J. (2017). Adaptive online event detection in news streams. Knowledge-Based Systems, 138, 105–112.

Huang, W., Li, Z., Zhang, L., and Li, Y. (2016). Review of intelligent microblog short text processing. Web Intelligence, 14(3), 211–228.

Huang, X., Zhang, X., Ye, Y., Deng, S., and Li, X. (2013). A topic detection approach through hierarchical clustering on concept graph. Applied Mathematics & Information Sciences, 7(6), 2285.

Ishikawa, S., Arakawa, Y., Tagashira, S., and Fukuda, A. (2012). Hot topic detection in local areas using Twitter and Wikipedia. ARCS Workshops (ARCS), 2012, 1–5.

Jensi, R., and Jiji, D. G. W. (2014). A survey on optimization approaches to text document clustering. ArXiv Preprint ArXiv:1401.2229.

Jeyaraj, A. (2018). Comparison of Feature Selection Strategies for Classification using Rapid Miner. July 2016. https://doi.org/10.15680/IJIRCCE.2016.

Kaleel, S. B., AlMeshary, M., and Abhari, A. (2013). Event detection and trending in multiple social networking sites. Proceedings of the 16th Communications & Networking Symposium, 5.

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

Kashef, S., and Nezamabadi-pour, H. (2014). An Advanced ACO Algorithm for Feature subset Selection. Neurocomputing. https://doi.org/10.1016/j.neucom.2014.06.067

Katragadda, S., Benton, R., and Raghavan, V. (2017). Framework for real-time event detection using multiple social media sources.

Khatdeo, S., Shrawane, S., Kumbhare, P., and Nimbarte, P. M. S. (2017). Detection and Visualization of Events from Online News. 1(21), 241–244.

Kumar, V. (2014). Feature Selection: A literature Review. The Smart Computing Review, 4(3). https://doi.org/10.6029/smartcr.2014.03.007

Kwan, E., Hsu, P.-L., Liang, J.-H., and Chen, Y.-S. (2013). Event identification for social streams using keyword-based evolving graph sequences. Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference On, 450–457.

Lam, W., Meng, H. M. L., Wong, K. L., and Yen, J. C. H. (2001). Using contextual analysis for news event detection. International Journal of Intelligent Systems, 16(4), 525–546.

Lavanya, S., Kavipriya, R., Yang, Y., Carbonell, J. Q., Brown, R. D., Archibald, B., and Liu, X. (2014). A Survey on Event Detection in News Streams. 2(5), 33–35.

Leban, G., Fortuna, B., Brank, J., and Grobelnik, M. (2014). Event Registry – Learning About World Events From News. Proceedings of the 23rd International Conference on World Wide Web, 107–110. https://doi.org/10.1145/2567948.2577024

Leban, G., Fortuna, B., and Grobelnik, M. (2016). Using News Articles for Real-time Cross-Lingual Event Detection and Filtering. NewsIR@ ECIR, 33–38.

Lee, R., and Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, 1–10.

Li, C., Sun, A., and Datta, A. (2012a). Twevent: segment-based event detection from tweets. Proceedings of the 21st ACM International Conference on Information and Knowledge Management, 155–164.

Li, R., Lei, K. H., Khadiwala, R., and Chang, K. C.-C. (2012b). Tedas: A twitter-based event detection and analysis system. Data Engineering (Icde), 2012 Ieee 28th International Conference On, 1273–1276.

Long, R., Wang, H., Chen, Y., Jin, O., and Yu, Y. (2011). Towards Effective Event Detection, Tracking and Summarization on Microblog Data. WAIM, 652–663.

Lu, Z., Yu, W., Zhang, R., Li, J., and Wei, H. (2015). Discovering Event Evolution Chain in Microblog. High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conferen on Embedded Software and Systems (ICESS), 2015 IEEE 17th International Conference On, 635–640.

Manaskasemsak, B., Chinthanet, B., and Rungsawang, A. (2016). Graph Clustering-Based Emerging Event Detection from Twitter Data Stream. Proceedings of the Fifth International Conference on Network, Communication and Computing, 37–41.

Mathioudakis, M., and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data, 1155–1158.

McMinn, A. J., Moshfeghi, Y., and Jose, J. M. (2013). Building a large-scale corpus for evaluating event detection on twitter. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 409–418.

Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, 889–892.

Mele, I., and Crestani, F. (2017). Event Detection for Heterogeneous News Streams. International Conference on Applications of Natural Language to Information Systems, 110–123.

Melvin, S., Yu, W., Ju, P., Young, S., and Wang, W. (2017). Event detection and summarization using phrase network. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 89–101.

Menon, R. (2010). Spatial - Temporal Random Indexing for Event Detection in Newswire Data.

Mohamad, A., Syed Mustapha, S., and Razali, M. (2010). Automatic Event Detection on Reuters News.

Mohammed, A. J., Yusof, Y., and Husni, H. (2016). Integrated Bisect K-Means and Firefly Algorithm for Hierarchical Text Clustering. J. Eng. Applied Sci, 100(3), 522–527.

Moutidis, I., and Williams, H. T. P. (2020). Complex networks for event detection in heterogeneous high volume news streams. ArXiv Preprint ArXiv:2005.13751.

Mustaffa, Z., Yusof, Y., and Kamaruddin, S. S. (2014). Application of LSSVM by ABC in energy commodity price forecasting. 2014 IEEE 8th International Power Engineering and Optimization Conference (PEOCO2014), 94–98.

Nanba, H., Saito, R., Ishino, A., and Takezawa, T. (2013). Automatic extraction of event information from newspaper articles and web pages. International Conference on Asian Digital Libraries, 171–175.

Newman, D., Baldwin, T., Cavedon, L., Huang, E., Karimi, S., Martinez, D., Scholer, F., and Zobel, J. (2010). Visualizing search results and document collections using topic maps. Web Semantics: Science, Services and Agents on the World Wide Web, 8(2–3), 169–175.

Nguyen, S., Ngo, B., Vo, C., and Cao, T. (2019). Hot Topic Detection on Twitter Data Streams with Incremental Clustering Using Named Entities and Central Centroids. 2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF), 1–6.

Nurwidyantoro, A., and Winarko, E. (2013). Event detection in social media: A survey. ICT for Smart Society (ICISS), 2013 International Conference On, 1–5.

Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., and Ounis, I. (2012). Bieber no more: First story detection using Twitter and Wikipedia. SIGIR 2012 Workshop on Time-Aware Information Access.

Panagiotou, N., Katakis, I., and Gunopulos, D. (2016). Detecting events in online social networks: Definitions, trends and challenges. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (Vol. 9580, pp. 42–84). Springer. https://doi.org/10.1007/978-3-319-41706-6_2

Passaro, L. C., Bondielli, A., and Lenci, A. (2016). FB-NEWS15: A Topic-Annotated Facebook Corpus for Emotion Detection and Sentiment Analysis. Proceedings of the Third Italian Conference on Computational Linguistics CLiC-It 2016.

Petrović, S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 181–189.

Phuvipadawat, S., and Murata, T. (2010). Breaking news detection and tracking in Twitter. Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference On, 3, 120–123.

Popescu, A.-M., and Pennacchiotti, M. (2010). Detecting controversial events from twitter. Proceedings of the 19th ACM International Conference on Information and Knowledge Management, 1873–1876.

Popescu, A.-M., Pennacchiotti, M., and Paranjpe, D. (2011). Extracting events and event descriptions from twitter. Proceedings of the 20th International Conference Companion on World Wide Web, 105–106.

Motooka, R. T. Yumoto, M. Nii, Y. Takahashi, and K. S. (2011). A Similar Event Search System Using Hashtag of Twitter,. The Database Society of Japan, The 3rd Fo, A1-5.

Rafea, A., and Mostafa, N. A. (2013). Topic extraction in social media. Collaboration Technologies and Systems (CTS), 2013 International Conference On, 94–98.

Rajani, N. F. N., McArdle, K., and Baldridge, J. (2014). Extracting topics based on authors, recipients and content in microblogs. Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, 1171–1174.

Ramadan, Q. H., and Mohd, M. (2011). A review of retrospective news event detection. Semantic Technology and Information Retrieval (STAIR), 2011 International Conference On, 209–214.

Rasouli, E., Zarifzadeh, S., and Rafsanjani, A. J. (2019). WebKey: a graph-based method for event detection in web news. Journal of Intelligent Information Systems, 1–20.

Ritter, A., Etzioni, O., and Clark, S. (2012). Open domain event extraction from twitter. Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1104–1112.

Rosa, K. Dela, Shah, R., Lin, B., Gershman, A., and Frederking, R. (2011). Topical clustering of tweets. Proceedings of the ACM SIGIR: SWSM.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web, 851–860.

Salloum, S. A. (2017). Mining Social Media Text: Extracting Knowledge from Facebook. Int. J. Com. Dig. Sys, 6(2).

Salloum, S. A., Al-Emran, M., and Shaalan, K. (2017). Mining Text in News Channels: A Case Study from Facebook. International Journal of Information Technology and Language Studies, 1(1), 1–9.

Salloum, S. A., Mhamdi, C., Al-Emran, M., and Shaalan, K. (2017). Analysis and Classification of Arabic Newspapers' Facebook Pages using Text Mining Techniques. International Journal of Information Technology, 1(2), 8–17.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. Proceedings of the 17th Acm Sigspatial International Conference on Advances in Geographic Information Systems, 42–51.

Saritha, S. K. (2019). Community Detection Methods in Social Network Analysis. In Emerging Technologies in Data Mining and Information Security (pp. 849–858). Springer.

Sayyadi, H., Hurst, M., and Maykov, A. (2009). Event detection and tracking in social streams. Icwsm.

Shukla, S., and Naganna, S. (2014). A review on K-means data clustering approach. International Journal of Information & Computation Technology, 4(17), 1847–1860.

Singh, M. H. (2016). Clustering of text documents by implementation of K-means algorithms. Streamed Info-Ocean, 1(1), 53–63.

Sowmiya, J. S., and Chandrakala, S. (2014). Joint sentiment/topic extraction from text. Advanced Communication Control and Computing Technologies (ICACCCT), 2014 International Conference On, 611–615.

Srividhya, S., and Mallika, R. (2018). Multimodal Feature Selection using Invasive Weed Optimization and Improved BAT for high dimensional Imbalanced datasets. International Journal of Applied Engineering Research, 13(2), 960–966.

Subašić, I., and Berendt, B. (2011). Peddling or creating? investigating the role of twitter in news reporting. European Conference on Information Retrieval, 207–213.

Sy, E., Jacobs, S. A., Dagnino, A., and Ding, Y. (2016). Graph-based clustering for detecting frequent patterns in event log data. Automation Science and Engineering (CASE), 2016 IEEE International Conference On, 972–977.

Taha, A. M., Chen, S.-D., and Mustapha, A. (2015). Bat algorithm based hybrid filter-wrapper approach. Advances in Operations Research, 2015.

Tembhurnikar, S. D., and Patil, N. N. (2015). Topic detection using BNgram method and sentiment analysis on twitter dataset. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2015, 1–6. https://doi.org/10.1109/ICRITO.2015.7359267

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based Systems, 24(7), 1024–1032. https://doi.org/10.1016/j.knosys.2011.04.014

Unankard, S., Li, X., and Sharaf, M. A. (2014). Emerging event detection in social networks with location sensitivity. https://doi.org/10.1007/s11280-014-0291-3

Vavliakis, K. N., Symeonidis, A. L., and Mitkas, P. A. (2013). Event identification in web social media through named entity recognition and topic modeling. Data & Knowledge Engineering, 88, 1–24.

Verma, J. P., Agrawal, S., Patel, B., and Patel, A. (2016). Big data analytics: Challenges and applications for text, audio, video, and social media data.

Wang, H., Xu, F., Hu, X., and Ohsawa, Y. (2013). Ideagraph: a graph-based algorithm of mining latent information for human cognition. Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference On, 952–957.

Wang, Y., Sundaram, H., and Xie, L. (2012). Social event detection with interaction graph modeling. Proceedings of the 20th ACM International Conference on Multimedia, 865–868.

Wasi, S., Shaikh, Z. A., and Shamsi, J. (2011). Contextual event information extractor for emails. Sindh University Research Journal-SURJ (Science Series), 43(1 (a)).

Wei, Y., Singh, L., Buttler, D., and Gallagher, B. (2018). Using semantic graphs to detect overlapping target events and story lines from newspaper articles. International Journal of Data Science and Analytics, 5(1), 41–60.

Weiler, A., Grossniklaus, M., and Scholl, M. H. (2014). Event identification and tracking in social media streaming data. EDBT/ICDT, 282–287.

Weng, J., and Lee, B.-S. (2011). Event detection in twitter. ICWSM, 11, 401–408.

Weng, J., Weng, J., Lim, E., and Jiang, J. (2010). Twitterrank : Finding Topic-Sensitive Influential Twitterers TwitterRank : Finding Topic-sensitive Influential Twitterers.

Wood, J. M. (2007). Understanding and Computing Cohen’s Kappa: A Tutorial. WebPsychEmpiricist. Web Journal at Http://Wpe. Info/.

Xue, B., Zhang, M., Browne, W. N., and Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. IEEE Transactions on Evolutionary Computation, 20(4), 606–626.

Yamanaka, T., Tanaka, Y., Hijikata, Y., and Nishida, S. (2010). A supporting system for situation assessment using text data with spatio-temporal information. Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, 22(6), 691–706.

Yang, S., Sun, Q., Zhou, H., Gong, Z., Zhou, Y., and Huang, J. (2018). A Topic Detection Method Based on KeyGraph and Community Partition. Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, 30–34.

Yang, Y., Carbonell, J. G., Brown, R. D., Pierce, T., Archibald, B. T., and Liu, X. (1999). Learning approaches for detecting and tracking news events. IEEE Intelligent Systems and Their Applications, 14(4), 32–43.

Yang, Y., Pierce, T., and Carbonell, J. (1998). A study of retrospective and on-line event detection. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 28–36.

Yu, S., and Wu, B. (2018). Exploiting structured news information to improve event detection via dual-level clustering. 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), 873–880.

Yusof, Y., Ahmad, F. K., Kamaruddin, S. S., Omar, M. H., and Mohamed, A. J. (2015). Short term traffic forecasting based on hybrid of firefly algorithm and least squares support vector machine. International Conference on Soft Computing in Data Science, 164–173.

Zarrinkalam, F., and Bagheri, E. (2017). Event identification in social networks. Encyclopedia with Semantic Computing and Robotic Intelligence, 1(01), 1630002.

Zhang, C., Wang, H., Wang, W., Ma, C., Li, J., Wang, Y., and Xu, F. (2016). EventPanorama: A Framework for Event Detection and Visualization from Online News. 2016 49th Hawaii International Conference on System Sciences (HICSS), 3739–3748. https://doi.org/10.1109/HICSS.2016.466

Zhang, C., Wang, H., Wang, W., and Xu, F. (2014). An improved ideagraph algorithm for discovering important rare events. Systems, Man and Cybernetics (SMC), 2014 IEEE International Conference On, 3290–3295.

Zhang, C., Wang, H., Xu, F., and Hu, X. (2013). Ideagraph plus: A topic-based algorithm for perceiving unnoticed events. Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference On, 735–741.

Zhang, Y., Szabo, C., and Sheng, Q. Z. (2015). Sense and focus: towards effective location inference and event detection on Twitter. International Conference on Web Information Systems Engineering, 463–477.

Zhao, Q., and Mitra, P. (2007). Event Detection and Visualization for Social Text Streams. ICWSM.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. European Conference on Information Retrieval, 338–349.

Zhou, H., Yu, H., Hu, R., and Hu, J. (2017). A survey on trends of cross-media topic evolution map. Knowledge-Based Systems.

Zhou, X., and Chen, L. (2014). Event detection over twitter social media streams. The VLDB Journal, *23*(3), 381–400.