Computers, Environment and Urban Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Review

Sensing and detecting traffic events using geosocial media data: A review

Shishuo Xu[a,b], Songnian Li[a,*], Richard Wen[a]

[a] Department of Civil Engineering, Ryerson University, 350 Victoria St., Toronto, ON M5B 2K3, Canada
[b] School of Environment Science and Spatial Informatics, China University of Mining and Technology, No. 1 Daxue Road, Xuzhou, Jiangsu 221116, China

ARTICLE INFO

Keywords:
Traffic event
Event detection
Geosocial media
Twitter data stream

ABSTRACT

Social media platforms, or social networks, have allowed millions of users to post online content about topics related to our daily lives. Traffic is one of the many topics for which users generate content. People tend to post traffic related messages through the ever-expanding geosocial media platforms. Monitoring and analyzing this rich and continuous user-generated content can yield unprecedentedly valuable traffic related information, which can be mined to extract traffic events to enable users and organizations to acquire actionable knowledge. A great number of literature has reported on the methods developed for detecting traffic information from social media data, especially geosocial media data when geo-tagged. However, a systematic review to synthesize the state-of-the-art developments is missing. This paper presents a systematic review of a wide variety of techniques applied in detecting traffic events from geosocial media data, arranged based on their adoption in each stage of an event detection framework developed from the literature review. The paper also highlights some challenges and potential solutions. The aim of the paper is to provide a structured view on current state-of-art of the geosocial media based traffic event detection techniques, which can help researchers carry out further research in this area.

1. Introduction

Traffic events, including traffic jams, roadworks, road closures, traffic accidents, and bad weather conditions, pose great challenges to both drivers and traffic management agencies (Gutiérrez, Figuerias, Oliveira, Costa, & Jardim-Goncalves, 2015). Identifying the time, location and type of traffic event in a real-time manner is important for drivers and traffic managers to generate proactive operation strategies to improve traffic conditions (Fu, Lu, Nune, & Tao, 2015; Fu, Nune, & Tao, 2015; Gu, Qian, & Chen, 2016).

Traditional methods applied to detect traffic events mainly focus on measuring traffic speed, traffic density and traffic flow using a wide variety of physical sensors (e.g., imaging sensor, inductive loop, magnetic sensor, acoustic detector, and passive infrared), which are usually installed at fixed locations along roads. They are implicitly embedded with the assumption that significant changes in flow characteristics immediately follow the traffic events (Gu et al., 2016). Guralnik and Srivastava (1999) and Ihler, Hutchins, and Smyth (2006) detected traffic events using the data collected from loop detectors by adopting time series algorithms. This study indicated that the proposed approach performed significantly better than a non-probabilistic threshold-based technique. The accuracy of loop detector data was further investigated by Coifman and Dhoorjaty (2004) using eight detector validation tests, which contrasted the performance of different sensor models and identified hardware problems to correct errors in the loop detector data. A novel event-driven architecture was proposed to deal with the continuous events created by sensors (Dunkel, Fernández, Ortiz, & Ossowski, 2011; Terroso-Sáenz, Valdés-Vela, Sotomayor-Martínez, Toledo-Moreo, & Gómez-Skarmeta, 2012). In terms of video image processors, spatio-temporal Markov random field algorithm (Kamijo, Matsushita, Ikeuchi, & Sakauchi, 2000) and kalman filtering-based approach (Veeraraghavan, Schrater, & Papanikolopoulos, 2005) were developed to automatically monitor traffic scenes and detect accidents at intersections. Li and Porikli (2004) and Porikli and Li (2004) detected highway traffic events in different illumination conditions (i.e., sunny, cloudy, and dark) based on an unsupervised, low-latency traffic congestion estimation algorithm.

However, it costs a lot to install physical sensors at a large scale to sense the traffic dynamics of the city. For example, as reported by Leduc (2008), the average cost to install and maintain an inductive loop detector at an intersection ranged from $9500 to $16,700 annually. Usually, such sensors are sparsely located along major highways other than local arterials. Traffic events may take place anywhere at any time (e.g., car crash on arterial road). Thus, the physical sensor-based method may not be an efficient way to timely detect traffic events due to its high cost, sparsity and limited spatial coverage. Crowdsourcing,

* Corresponding author.
E-mail address: snli@ryerson.ca (S. Li).

https://doi.org/10.1016/j.compenvurbsys.2018.06.006
Received 14 January 2018; Received in revised form 3 May 2018; Accepted 19 June 2018

which refers to a low-cost process of solving a problem through obtaining contributions from a large group of people via online communities (Doan, Ramakrishnan, & Halevy, 2011; Howe, 2006), is likely to be an available solution.

With the wide use of smart phones and mobile devices, crowdsourcing becomes a promising alternative approach to feasibly collect traffic related data with spatial and temporal information (Zheng et al., 2016). These alternative information sources include user-shared traffic information, vehicle and individual GPS trajectories, Bluetooth data, and cellular network, which provide traffic information implicitly through the users' movements. WAZE (Waze Mobile, 2018) is a real-time traffic monitoring and navigation system, which alerts users of abnormal traffic conditions and gives them the best route based on user shared reports of pre-defined categories (e.g., accidents, road hazards, and traffic jams). GPS trajectories data is another typical crowdsourcing information to detect traffic events. Kamran and Haas (2007) utilized the real-time GPS data collected from vehicles to identify the abnormal traffic pattern on motorways based on a multilevel approach. A hierarchical analysis was further conducted to determine the precise location of traffic events. Similarly, INRIX data, which were obtained from individual trajectories, was also used to recognize occurrence of traffic events through a Bayesian structure equation model (Park & Haghani, 2016). Furthermore, a mobile application, named WreckWatch, polled smartphone system sensors (e.g., GPS receiver and accelerometers) and context data (e.g., speed) to automatically detect traffic events (White, Thompson, Turner, Dougherty, & Schmidt, 2011). The emergency notifications were sent to the first responders through WreckWatch to improve situational awareness. Martchouk, Mannering, and Bullock (2011) measured the travel time variability due to adverse weather and traffic breakdown by collecting Bluetooth probe data on freeway segments in Indianapolis. Demissie, de Almeida Correia, and Bento (2013) analyzed the correlation between the cellular networks handover counts and traffic volumes to reveal traffic status by training an Artificial Neural Network (ANN). Such cellular based monitoring contains high measurement errors for the accuracy that are usually proportional to cell size. It was reported that the median error of cellular network positioning was up to 600 m (Zandbergen, 2009). The previously mentioned crowdsourcing data often belong to private operators and the quality of data sharing is often a challenging issue due to the privacy matters. Besides, they do not necessarily enable real-time data processing, which is required for traffic event detection.

Geosocial media platforms allow users to compose and post short statements about their perceptions and/or experiences with geolocations, which plays an increasingly important role in our daily lives (Kelley, 2013). The free-cost features enable users to easily share a variety of information to the public, including photos, video, and blogs, with more spatial and temporal overage compared to physical sensors (Kaplan & Haenlein, 2010). For example, Facebook, Twitter and Weibo (a Chinese version of Twitter), usually have hundreds of millions users, which can generate a large amount of posts attached with timestamps, geolocations, and text contents. In other words, wherever there is a user there is a potential for geosocial media data. Further, geosocial media data can be used to not only identify when and where traffic anomalies take place (i.e., traffic pattern), but also explain the reasons behind the traffic anomalies in a real-time manner due to the abundant semantics of geosocial media content. This provides a significant advantage of geosocial media data over GPS data in detecting traffic patterns (Rashidi, Abbasi, Maghrebi, Hasan, & Waller, 2017), especially in detecting traffic events rather than condition along a road segment. Therefore, it is likely to be an effective way to extract useful information from these abundant messages to detect traffic events.

Specially, Twitter is one of the most popular geosocial media sites all over the most part of the world, from which many studies have been done on detecting traffic information. It provides a free approach to acquire public tweets through open API, such as REST APIs and Streaming APIs (Twitter Inc., 2018). The Twitter REST APIs provide the ability to search by certain keywords or accounts from sample of recent tweets published in the past 7 days. Streaming APIs enable developers to collect real-time tweets with a set of bounding boxes or comma-separated list of phrases. The acquired tweets are often tagged with a pair of longitude and latitude coordinates (if the location-based service is turned on), a timestamp, and a short message limited to 140 characters (280 after November 7, 2017). Recently, Twitter has been adopted as a powerful data source to detect disasters (Dingli, Mercieca, Spina, & Galea, 2015; Kryvasheyeu et al., 2016; Sakaki, Okazaki, & Matsuo, 2010), predict election results (Metaxas & Mustafaraj, 2012) and crimes (X. Wang, Gerber, & Brown, 2012; Zhao, Chen, Lu, & Ramakrishnan, 2015), spread breaking news (Amer-Yahia et al., 2012; Phuvipadawat & Murata, 2010; Sankaranarayanan, Samet, Teitler, Lieberman, & Sperling, 2009) and identify small-scale geosocial events (R. Lee & Sumiya, 2010; Watanabe, Ochi, Okabe, & Onai, 2011), which have happened in the real world.

Traffic is also a popular topic people would like to discuss in their daily lives (Novaco & Gonzalez, 2009). Thus, they tend to post traffic related information via social media networks when there is an accident, car crash, roadwork, or road closure. Geosocial media, especially Twitter, proves to be a valuable source in generating a wide range of traffic related information to detect traffic events to support traffic planning and management (Gal-Tzur et al., 2014; Gal-Tzur, Grant-Muller, Minkov, & Nocera, 2014; Grant-Muller et al., 2014, 2015). An exploratory study conducted in Northern Virginia, which analyzed the correlation between traffic patterns and traffic-related Twitter concentration from a spatiotemporal perspective, revealed that 77.4% of traffic-related Twitter concentrations could be justified by local traffic surge (Zhang et al., 2016). Another similar study showed that tweets tended to be posted within 5-h of the event that they referred to, and were most often sent between 10 and 25 miles of the event's location (Mai & Hranac, 2013). Zhang, Tang, Wang, and Wang (2015) indicated that the spatial distribution of traffic related tweets were clustered mostly within 800 m around traffic incidents in Seattle downtown area. These previous studies have proved that geosocial media data is a valuable data source to detect traffic events.

Up to now, various Twitter based applications have been developed to detect traffic events in a cost-effective way. For example, both Steds (Fu, Lu, et al., 2015) and Butterfly (Fu, Zhong, Lu, & Boedihardjo, 2015) were proposed as novel query expansion methods based on apriori algorithm for extracting traffic related tweets, which were then ranked to better summarize the detected events. TEDS (Liu, Fu, Lu, Chen, & Wang, 2014) adopted spatio-temporal analysis and a wavelet analysis model for traffic events detection. STAR (Semwal, Patil, Galhotra, Arora, & Unny, 2015) analyzed the relationship between co-occurring problems and their causes to train a classifier to predict severe problems for the next day. TrafficWatch (Nguyen, Liu, Rivera, & Chen, 2016), Traffic Observatory (Ribeiro Jr. et al., 2012) and TEDAS (R. Li, Lei, Khadiwala, & Chang, 2012) followed a process of preprocessing tweets to create tokens, identifying traffic related tweets using classification methods, and geocoding them to determine the exact location of events. Moreover, semantic web technologies were applied to interpret the underlying reasons behind traffic events in Dub-Star (Daly, Lécué, & Bicer, 2013) and STAR-CITY (Lécué et al., 2014).

However, applying geosocial media data to detect traffic events still faces many challenges. For example, Twitter messages are restricted in length and written by anyone. Therefore, tweets include large amount of informal, irregular, and abbreviated words, as well as a large number of spelling and grammatical errors, improper sentence structures, and mixed languages. In addition, Twitter streams contain large amount of meaningless messages (Hurlock & Wilson, 2011), polluted content (K. Lee, Eoff, & Caverlee, 2011), and rumors (Castillo, Mendoza, & Poblete, 2011), which negatively affect the performance of the detection algorithms.

Existing studies (Atefeh & Khreich, 2015; Bontcheva & Rout, 2014; Garg & Kumar, 2016; Goswami & Kumar, 2016; Hasan, Orgun, &

Schwitter, 2017; Nurwidyantoro & Winarko, 2013; Weiler, Grossniklaus, & Scholl, 2015) mainly focus on reviewing detections of all types of events based on geosocial media streams, but do not specifically gain insights into traffic domain. Lv, Chen, Zhang, Duan, and Li (2017) summarized the main topics in traffic related research using social media data, and analyzed the current collaboration patterns from perspectives of researchers, institutions, and countries, which did not exhaustively discuss the representative methods adopted in detection processes. The focus of our study was to investigate more processing details of geosocial media based research concerning traffic event detection to complement and extend previous works.

In this paper, we consider the characteristics of traffic events and present a general workflow of extracting traffic events from geosocial media data (further discussed in Section 2), where Twitter is adopted as a major data source platform, while Weibo and Facebook are only mentioned in a few studies due to limited literature found. Aiming at providing readers a clear perspective on the recent literatures, this paper does not provide an exhaustive review of all surveyed works but focus more on discussing the representative techniques applied in each stage of the overall workflow, such as natural language processing, machine learning, information extraction and retrieval, and data mining.

Moreover, this paper sheds light on monitoring traffic conditions in smart transportation platforms, which plays an important role in smart cities (Harrison & Donnelly, 2011). The semantic richness of geosocial media data provides more opportunities to explain the reasons behind the traffic conditions (Djahel, Doolan, Muntean, & Murphy, 2015; Rashidi et al., 2017). Therefore, it is a feasible way to deeply investigate drivers' and pedestrians' feedback of reporting a traffic event in geosocial media platforms, which may assist traffic authorities for better planning of road networks expansion, replacement of road signs, and setting speed limits (Djahel et al., 2015). Meanwhile, reported traffic events provide drivers with situational awareness to more efficiently plan their routes (Doolan & Muntean, 2017; Kousiouris et al., 2018). This will significantly improve traffic flow control and improve road safety in order to ease air pollution severity caused by traffic congestion and enhance citizens' quality of life (Djahel et al., 2015; Doolan & Muntean, 2017).

The remainder of this paper is organised as follows. Section 2 presents a typical traffic-event detection framework and its components. Sections 3–6 describe techniques applied in each stage of traffic event detection with details including querying for traffic related tweets, pre-processing tweets, identifying tweets relevant to a real-world traffic event, and extracting location information. Techniques used to summarize the detected traffic events and notify them to users are presented in Section 7. Some other miscellaneous approaches that do not directly belong to the process mentioned above are illustrated in Section 8. Performance evaluation of different methods employed in the literature, summarized in Section 3–8, are discussed in Section 9. Finally, a number of general observations on different traffic event detection approaches are discussed in Section 10. Section 11 concludes the paper with a brief summary.

2. A typical framework for detecting traffic events from geosocial media data

With respect to characteristics of traffic events defined in the beginning, a typical framework is presented in Fig. 1, which illustrates a number of components to deal with different stages of operation in detecting traffic events from geosocial media data. The workflow consists of querying for traffic related geosocial media data, pre-processing geosocial media data, identifying geosocial media data relevant to real-world traffic events, extracting location information, and summarizing and notifying the detected traffic events to users.

There are also miscellaneous techniques found in some studies in the literature which do not directly belong to the workflow presented in Fig. 1, but are used in the traffic event detection. The details of miscellaneous research are discussed in Section 8. Although the workflow applies to other type of geosocial media data, our discussion hereafter will mainly focus on Twitter data, which has a large number of published studies.

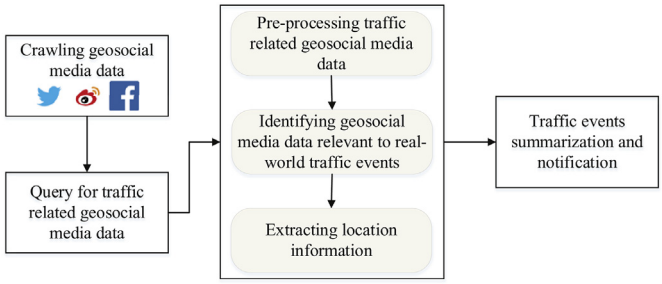


Fig. 1. A typical framework for traffic events detection using geosocial media data.

Twitter open several APIs to the public for obtaining free data, such as Streaming APIs and REST APIs. They allow users to crawl the raw tweets in JSON format with predefined keywords or geospatial bounding box. Keywords used to track tweets are specified by a comma-separated list and the tweets obtained may not have coordinates or place tagged. If a geospatial bounding box, normally specified by a comma-separated list of longitude and latitude pairs, with the southeast corner of the bounding box coming first, is used for tweets collection, only geo-located tweets falling within the requested bounding boxes will be included. Twitter allows users to geotag their tweets with exact coordinates (i.e., a pair of longitude and latitude) or a Twitter place, such as a specific business, landmark, or point of interest (e.g., Time Square, New York, United States). The tagged place corresponds to a bounding box enclosing this place entity rather than a unique pair of longitude and latitude (https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects). To be specific, if the raw tweets are geotagged with coordinates, the pair of longitude and latitude will be tested against the bounding box. If the raw tweets are geotagged with places, the region defined in place is checked for intersection against the bounding box. Any overlap indicates matching the location query. Tweets without coordinates or places will be ignored.

In Section 3, a query with preselected keywords and accounts is employed to search for traffic related tweets from raw tweets. The filtered tweets are regarded as potential sources to report traffic events. The pre-processing component deals with the filtered tweets to tokens through lowercase normalization, stop words removal, Part-of-Speech (POS) tagging and Named Entity Recognition (NER). A detailed discussion of the common pre-processing techniques is presented in Section 4.

Following the pre-processing step, a classification-based method is applied to identify whether a pre-processed tweet relates to a traffic event that takes place in the real world. For instance, traffic related keyword “accident” both exist in the following two tweets: “Quality is never an accident. It is always the result of intelligent effort.” and “The accident occurred around 9:45pm at Nkawkaw, when our bus ran into a stationary vehicle.” The former means a coincidence, while the latter refers to a real traffic accident. The purpose of classification is to discard the former and save the latter. Section 5 presents some insights into this process by reviewing feature selection and classification methods. In order to provide better situational awareness for drivers and traffic planners, it is critical to determine the precise location of the detected traffic event. Geocoding processes presented in Section 6 is then used to extract location information based on tagged GPS coordinates (i.e., longitude and latitude), user profiles information (e.g., Toronto), and tweet contents (e.g., street name).

Some studies also include a component of traffic events summarization and visualization. It creates a summary of tweets associated

with traffic events and then push the detected event to users via a mobile or web-based application. Further discussion of this component is presented in Section 7. Finally, the performance of the traffic events detection techniques mentioned in Section 3–7 are evaluated in Section 9.

3. Querying for traffic related tweets

Despite the huge amount of raw data obtained from geosocial media platforms, we are only interested in the information specific to the traffic domain. The accurate extraction of traffic related tweets is very important, as this is the entry point to the following classification process.

There are two main approaches to extract traffic related information from raw tweets: keywords-based approach and accounts-based approach. When certain traffic related keywords (e.g., accident, traffic, and crash) are detected in the tweet, it indicates that a potential traffic event is observed by the user who posts it. Besides, there are also many organization accounts operated by traffic authorities whose main function is to disseminate real-world traffic events information. In this way, traffic events information is often posted after the traffic management officials are already notified. Personal tweets are more likely to disseminate a traffic event in a more timely manner than organizational accounts (Yazici, Mudigonda, & Kamga, 2017). Table 1 shows some examples of querying for traffic related tweets using keywords-based or accounts-based methods separately, while Table 2 summarizes the studies which combine keywords-based with accounts-based methods to make queries.

In terms of keywords-based approach, some studies (D'Andrea et al., 2015; Mai & Hranac, 2013; Nguyen et al., 2016; Yazici et al., 2017) predefined a bunch of traffic related keywords with common sense to directly retrieve candidate tweets. More rigorously, a corpus of traffic related documents were first investigated to extract traffic related terms by ranking their frequency based on the classical term frequency-inverse document frequency (tf-idf) method. The top terms were gathered and reviewed by the experts to create a list of keywords for the query (Kuflik et al., 2017). Similarly, Fu, Lu, et al. (2015) adopted the tf-idf method to rank the importance of each word from tweets posted by seed influential traffic accounts. Top words were then selected as keywords to acquire more traffic related tweets. This approach is likely to pose a restriction for covering all of the traffic related tweets, and some relevant tweets can be neglected since the pre-defined keyword list is very limited, and are not updated with the streaming data. Some extended query methods are provided to extend the pre-defined keyword list. For example, a thesaurus, such as WordNet, were also used to create a list of similar words and hyponyms so as to extend the seed keywords set (e.g., “incident”, “injury”, “police”, “vehicle”, “accident”, and “road”) (Schulz et al., 2013). For instance, the keyword “accident” was extended with “collision”, “crash”, “wreck”, “injury”, “fatal accident”, and “casualty”.

Considering the fact that streaming tweets are collected continuously, it is better to make some changes to the static keywords list used above to meet the updates. A dynamic query approach was used to fulfill this task. Following the initial data acquisition process using seed keywords, an adaptive data acquisition method was conducted to iteratively update the keywords dictionary so that more potential traffic related tweets could be obtained. This process kept running until the number of traffic related tweets became stable (Gu et al., 2016; Li et al., 2012). Zhang et al. (2015) adopted the Latent Dirichlet Allocation (LDA) model to identify two traffic related topics from Washington State Incident Tracking System (WITS), and then integrated a hierarchical clustering algorithm to iteratively filter out the irrelevant tweets.

The accounts-based approach uses influential traffic accounts owned by traffic authorities to directly collect traffic related tweets. For example, Endarnoto et al. (2011) fetched tweets from @TMCPoldaMetro account of TMC (traffic management center). Ribeiro Jr. et al. (2012) collected tweets from ten accounts whose main purpose was to inform traffic conditions in Belo Horizon and other cities in Brazil including @TransitoBH, @Transito98FM and @waytaxi. Similarly, a dataset of tweets were collected from several official traffic monitoring Twitter accounts in Ireland and Indonesia by Daly et al. (2013) and Kurniawan et al. (2016), respectively. Taking advantages of both accounts-based approach and keywords-based approach, Wang et al. (2015) combined these two approaches to efficiently collect traffic related tweets. Aziz et al. (2015) used specific accounts and hashtags other than keywords to complete this process.

4. Pre-processing traffic related tweets

Due to the limited length of Twitter messages, there exists problems of mining text information, such as language ambiguity, uncertainty, and abbreviation. It is necessary to clean the tweet texts for the feature extraction to identify real-world traffic events. Pre-processing tweets is mainly based on basic natural language processing techniques, including tokenization, normalizing all words to lowercase, removing non-English and duplicate posts (i.e., retweets), removing stop words, links and mentions to other Twitter accounts, correcting spelling errors, slang replacement, stemming, lemmatizing, and POS tagging. A sample tweet pre-processing is shown in Fig. 2.

Table 1
Summary of using the keywords-based or accounts-based methods to query for traffic related tweets.

Articles	Query methods	Details
Mai and Hranac (2013)	Predefining traffic related keywords	“accident”, “crash”, “traffic”, “road, “freeway”, “highway”
D'Andrea, Ducange, Lazzerini, and Marcelloni (2015)		“traffic”, “crash”, “queue”
Nguyen et al. (2016)		“accident”, “crash”, “delay”, “traffic”
Kuflik et al. (2017)		Traffic domain experts review the keywords list ranked by their frequency in a corpus of traffic related documents.
Yazici et al. (2017)		“accident”, “crash”, “traffic”, “road”, “freeway”, “highway”, “lane”, “wreck”, “car”, “cars”, “delay”, “NB”, “northbound”, “SB”, etc.
R. Li et al. (2012)	Extending traffic related keywords	The iteratively refined rules are used to retrieve more tweets based on seed keywords.
Schulz, Ristoski, and Paulheim (2013)		WordNet is used to extend the seed keywords like “incident”, “injury”, “police”, “vehicle”, “accident”, “road”.
S. Zhang et al. (2015)		Topic modeling combined with a hierarchical clustering algorithm are adopted to filter out the irrelevant tweets.
Endarnoto, Pradipta, Nugroho, and Purnama (2011)	Preselecting influential traffic accounts	@TMCPoldaMetro
Ribeiro Jr. et al. (2012)		@TransitoBH, @Transito98FM, @waytaxi
Daly et al. (2013)		@LiveDrive, @AARoadwatch, @GardaTraffic
Kurniawan, Wibirama, and Setiawan (2016)		@ATCS_DIY, @atcs_kotasmrg, @atcs_kotatgr, @atcs_pekalongan, @ntmclantaspolri, etc.

Table 2
Summary of combining the keywords-based with accounts-based methods to query for traffic related tweets.

Articles	Query methods	Details
S. Wang, He, Stenneth, Yu, and Li (2015)	Preselecting influential traffic accounts and predefining traffic related keywords	@ChicagoDrives, @ChiTraTracker, @roadnowChicago, @traffic Chicago, etc. "stuck", "congestion", "jam", "crowded", "pedestrain", "driver", "accident", "crash", etc.
Fu, Lu, et al. (2015)	Preselecting influential traffic accounts and keywords query expansion	Tf-idf method is used to rank the importance of each word based on tweets from seed users (@WTOPtraffic, @VaDOT, @drgridlock, @DCPoliceDep). Top words are then selected as keywords to acquire more traffic related tweets.
Gu et al. (2016)	Preselecting influential traffic accounts and extending traffic related keywords	Influential traffic accounts are collected manually by Google searching, and an adaptive keywords-based method is also used to acquire more traffic related tweets.
Aziz, Prihatmanto, Henriyan, and Wljaya (2015)	Preselecting specific hashtag and account	#lalinbdg, @ lalinbdg

Tokenization aims to break the short tweet text into separate tokens (Aziz et al., 2015; D'Andrea et al., 2015), and all the letters are normalized to lowercase (Aziz et al., 2015; Kurniawan et al., 2016; Yazici et al., 2017) at the same time. Non-English and duplicate posts (Kurniawan et al., 2016), accent marks (Ribeiro Jr. et al., 2012), and links and mentions to other Twitter accounts (Kurniawan et al., 2016; Ribeiro Jr. et al., 2012) are further removed. Filtering stop words, such as punctuations (Yazici et al., 2017) and non-alphanumeric (alphabets and numbers) characters (Kurniawan et al., 2016), are also an important step in the cleaning process (Aziz et al., 2015; D'Andrea et al., 2015; Kumar, Jiang, & Fang, 2014; Nguyen et al., 2016; Schulz et al., 2013; Schulz & Ristoski, 2013; Yazici et al., 2017). Moreover, as users may post a tweet with spelling errors and slangs, a replacement is required to make corrections (Schulz et al., 2013).

Stemming is the process of reducing each word to its stem or root form by removing its suffix or prefix. D'Andrea et al. (2015) and Kumar et al. (2014) used the Porter algorithm to reduce inflected words and transform variants of words into a single stem. Endarnoto et al. (2011), Schulz and Ristoski (2013), Schulz et al. (2013), and Nguyen et al. (2016) made use of the Stanford lemmatization function (Manning et al., 2014) to normalize words, and then applied the Stanford POS tagger (Manning et al., 2014) to filter meaningless categories of words. Furthermore, Gutiérrez et al. (2015) used a POS tagger for analyzing the time expressions and verbal forms to extract temporal information from the tweet texts.

5. Identifying tweets relevant to real-world traffic events

As described in Section 2, some tweets containing traffic related keywords do not actually refer to a real-world traffic event. The main task of this stage is to identify and remove these noisy tweets, in other words, classifying the processed tweets into real traffic events and non-traffic events. To this end, a summarized process including feature extraction and classification is presented in Fig. 3, which illustrates the process of identifying real traffic events. Tables 3 and 4 summarizes a list of techniques applied in this process, which are further discussed in Sections 5.1 and 5.2.

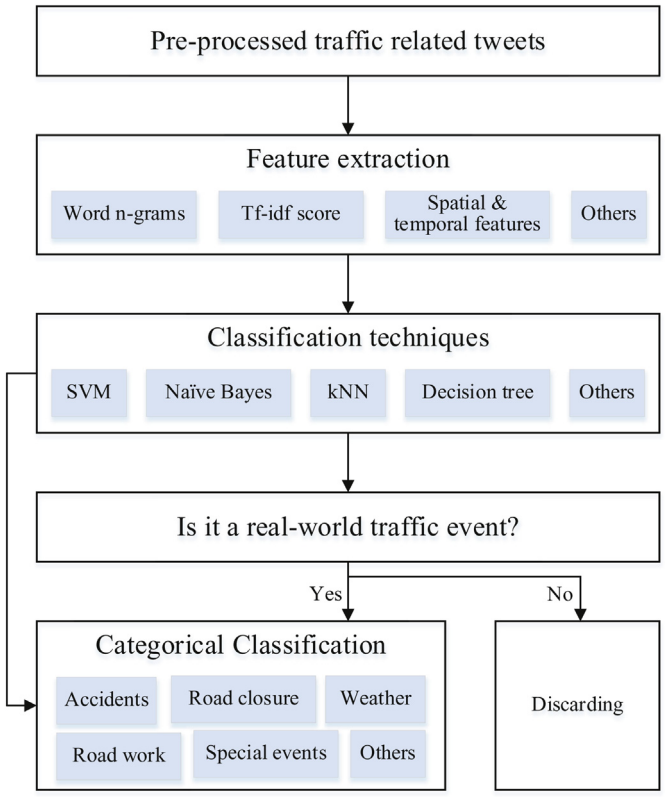


Fig. 3. A flowchart of classification.

5.1. Feature extraction

It is generally acknowledged that tweet texts concerning a real-world traffic event may have one or more traffic related keywords. Thus single words and combinations of some words, which were positively correlated with being a traffic related tweet, were extracted as features

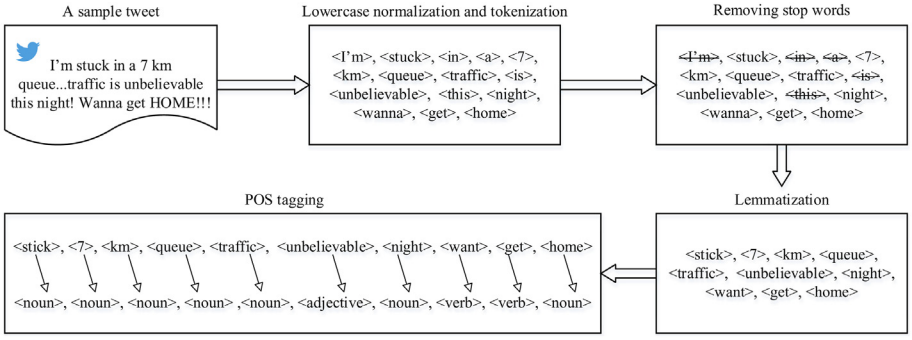


Fig. 2. Pre-processing of a sample tweet (modified from D'Andrea et al., 2015).

Table 3
Summary of classification on real-world traffic events or categorical traffic events.

Articles	Features	Classification	
		Classification on traffic events/ non-traffic events	Classification on traffic events categories
Kurniawan et al. (2016)	Words and their appearance account in a tweet	Naïve Bayes, SVM, Decision tree	N/A
Yazici et al. (2017)	Word n-grams, tf-idf score	Naïve Bayes	N/A
Z. Zhang et al. (2016)	Correlated words with high coefficient in traffic related tweets	Maximum likelihood estimation model (MLE)	N/A
Schulz et al. (2013)	Word n-grams, char n-grams, tf-idf score, syntactic features, spatial & temporal features, FeGeLOD features	Naïve Bayes, Ripper rule learner (JRip), SVM	N/A
Sakaki, Matsuo, Yanagihara, Chandrasiri, and Nawa (2012)	Dependency features, context features, position features, time expression features, word features	N/A	SVM; Heavy traffic, traffic restrictions, police checkpoints, rain, mist
Ribeiro Jr. et al. (2012)	N/A	N/A	Manually creating rules to identify traffic conditions (e.g., slow) and events (e.g., accident)

(Gu et al., 2016). An important and basic statistical model of natural language processing, called the N-gram model, was also used in processing the limited words of a tweet to generate features (Cui et al., 2014; Kuflik et al., 2017).

In addition, the frequency of words in a tweet represents the words' importance to some degree, which can be used as another feature that helps classification. Kurniawan et al. (2016) extracted both words and their appearance count in a tweet as useful features. Tf-idf is one of the most popular term-weighting schemes, which considers both term frequency and inverse document frequency. Gutiérrez et al. (2015) and Yazici et al. (2017) combined tf-idf score with word n-grams as feature representations. Zhang et al. (2016) calculated the coefficients of correlated words and ranked them. Those correlated words whose coefficients were higher than 0.5 were selected as features. Furthermore, stems (D'Andrea et al., 2015), bag of words, lemma, POS and chunk features, pattern recognizer, and bag of tags (Nguyen et al., 2016) can also be adopted as available features.

With regard to detecting traffic events from tweets, the spatial and temporal information embedded in a tweet is likely to provide comprehensive insight. Combining spatial, temporal, and semantic features together lays a solid foundation for the following step of classification (Sakaki et al., 2012; Schulz et al., 2013).

5.2. Classification

The classification process is realized through manually labeling and machine learning techniques. Ribeiro Jr. et al. (2012) manually classified traffic related tweets into two main categories: traffic condition and event. Traffic condition refers to the status at a given location at a given moment (e.g., “slow”), while a traffic event corresponds to a situation that may change the traffic status (e.g., “accident”). The manual method is time consuming and labor intensive, so machine leaning techniques are widely used as an alternative solution to this problem.

In terms of the machine learning based classification approach, some techniques, including support vector machine (SVM) (Gutiérrez et al., 2015; Kuflik et al., 2017; Kurniawan et al., 2016; Nguyen et al., 2016; Schulz et al., 2013), Naïve Bayes (Gu et al., 2016; Kurniawan et al., 2016; Schulz et al., 2013; Yazici et al., 2017), Decision tree (Kurniawan et al., 2016; Nguyen et al., 2016), Ripper rule learner (Schulz et al., 2013), k-nearest neighbor (kNN) (Nguyen et al., 2016), Bayesian Network (Nguyen et al., 2016), and Maximum likelihood estimation model (MLE) (Z. Zhang et al., 2016), are commonly adopted to identify if a tweet is relevant to a real-world traffic event or not. Their performances are evaluated using metrics like accuracy, precision, recall, and F_1-score (Kurniawan et al., 2016). D'Andrea et al. (2015) performed a classification over two classes (i.e., traffic event and non-traffic event) and three classes (i.e., traffic due to external event, traffic congestion or crash, and non-traffic), based on SVM, Naïve Bayes, C4.5 decision tree, kNN and PART techniques. Sakaki et al. (2012) directly collected tweets for five classes, which were heavy traffic, traffic restrictions, police checkpoints, rain, and mist, based on predefined keywords. The SVM method was then used to identify whether the filtered tweets were actually representing the specific event of each

Table 4
Summary of classification on real-world traffic events and categorical traffic events.

Articles	Features	Classification	
		Classification on traffic events/non-traffic events	Classification on traffic events categories
Gu et al. (2016)	Single words and combinations of some words that are positively correlated with being a traffic related tweet	Semi Naïve Bayes	Latent Dirichlet Allocation (LDA); Accidents, road work, hazards & weather, special events, and obstacle vehicles
Cui, Fu, Dong, and Zhang (2014)	Word n-grams	Bayesian classifier	Natural language processing based structural labelling; Traffic flow, traffic accident and traffic control
Kuflik et al. (2017)		SVM	SVM; Expression of an opinion, transport need, reporting an event
Gutiérrez et al. (2015)	Word n-grams, tf-idf score	SVM	Constructing a list of corresponding synonyms of representative keywords for each class; Traffic jam, road work, freight traffic, road closure, ice, wind & snow, traffic accident, and others
D'Andrea et al. (2015)	Relevant stems that are retrieved from traffic related tweets	SVM, Naïve Bayes, C4.5 decision tree, kNN, PART	SVM, Naïve Bayes, C4.5 decision tree, kNN, PART; Traffic due to external event, traffic congestion or crash, and non-traffic
Nguyen et al. (2016)	Bag of words, lemma, POS and chunk features, pattern recognizer, bag of tags	kNN, Bayesian Network, SVM, C4.5 decision tree	Conditional Random Fields (CRFs) labelling; Queue, accident, breakdown, police activities, road work

class.

After the tweet is identified as representing a real-world traffic event, a categorical classification is conducted to identify fine-grained event categories, such as traffic accident, roadwork, and road closure. Cui et al. (2014) applied the natural language processing based structural labelling method to classify the traffic events into traffic flow, traffic accident and traffic control. Gutiérrez et al. (2015) constructed a list of corresponding synonyms of representative keywords for each class. Traffic jam, roadwork, freight traffic, road closure, ice, wind & snow, traffic accident, and others were further identified via keywords filtering. Gu et al. (2016) adopted Latent Dirichlet Allocation (LDA) topic modeling method to classify the traffic events into five categories: accidents, roadwork, hazards & weather, special events, and obstacle vehicles. Nguyen et al. (2016) made a further classification of traffic events into queue, accident, breakdown, police activities, and roadwork based on Conditional Random Fields (CRFs) labelling method. Kuflik et al. (2017) investigated the traffic events and identified them as expression of an opinion, transport need, and reporting an event using a SVM technique.

6. Extracting location information

After the identification of real-world traffic events, some studies further extracted their location information and geocoded them with different tools. The geographic location information carried by tweets is rich but may be very noisy and may not be explicitly available. There are generally three types of location information: tagged GPS coordinates, user profiles, and tweet texts (e.g., geographic names and street names). The use of the location information in traffic event detection are summarized in Table 5.

Some tweets carry a pair of coordinates, namely longitude and latitude, when they are tweeted from smart phones that have location-based service enabled. These coordinates correspond to the more precise locations where users post tweets. They are usually regarded as where the traffic events take place if the tweets are relevant to real-world traffic events (Kumar et al., 2014; Zhang et al., 2015, 2016). However, a major problem about GPS coordinates tagged with tweets is their sparsity. It has been reported that geotagged tweets constitute only 2%–3% of all tweets in Twitter (Giridhar, Abdelzaher, George, & Kaplan, 2015). In fact, this percentage varies with the way the tweets are collected. A higher percentage would be achieved when geospatial bounding box is used to collect Twitter data since the bounding box looks for data satisfying the given criteria for coordinates (Ozdikis, Oğuztüzün, & Karagoz, 2017). For example, around 30.11% and 27.49% of tweets crawled by a bounding box in Streaming API are geotagged with coordinates in Asia and North America, respectively (Morstatter, Pfeffer, Liu, & Carley, 2013). Despite the tweets with exact coordinates, Twitter allows user to label their tweets with places, which are sourced from third party websites, such as Foursquare and Yelp (Ajao, Hong, & Liu, 2015). In existing studies of traffic events detection, extracting location information using places that are tagged with traffic related tweets are not specifically analyzed. There exist confusion and uncertainty to determine the exact location of a place due to the fact a place refers to a bounding box rather than a unique pair of coordinates. As such, user profiles and tweets text are further analyzed to provide alternative solutions for location inference of traffic events.

Table 5
Summary of geocoding techniques.

Articles	GPS coordinates	User profiles	Tweet texts
Wanichayapong, Pruthipunyaskul, Pattara-Atikom, and Chaovalit (2011)	×	×	√
R. Li et al. (2012)	√	√	×
Sakaki et al. (2012)	√	×	√
Ribeiro Jr. et al. (2012)	×	×	√
Daly et al. (2013)	×	×	√
Schulz et al. (2013)	×	×	√
Chen, Chen, and Qian (2014)	√	×	√
Cui et al. (2014)	√	×	√
Kumar et al. (2014)	√	×	×
Gutiérrez et al. (2015)	×	×	√
S. Zhang et al. (2015)	√	×	×
S. Wang et al. (2015)	√	×	√
Tejaswin, Kumar, and Gupta (2015)	×	×	√
Gu et al. (2016)	×	×	√
Nguyen et al. (2016)	√	×	√
Z. Zhang et al. (2016)	√	×	×

Some tweets are posted by accounts whose profiles are shared with the public, such as city, country, and sometimes finer-grained business names and street address of the business. Li et al. (2012) predicted the location of a non-GPS tagged tweet with user's historical GPS tags and his social network, as some of his friends had a particular location on their profiles. It assumed that (1) a user's location was more likely to appear in his tweets than other locations; (2) a user's friends tended to be closer with the user; and (3) a user's location mentioned at least once in his tweets or was the same with at least one of his friends. The geographic location was thus assigned to a tweet using related historical tweets and networks.

Mining the text of tweets is also able to infer the event locations at various granularity levels (e.g., city, district, street name or block number) depending on the event type and can provide more than one location (Ozdikis et al., 2017). With respect to extracting location information of traffic events, specific road names may also be referred in tweet texts, since users tend to add certain locational information when posting a tweet to report a traffic event. For example, “Due to traffic congestion at 15th & Pennsylvania Ave NW, buses are experiencing up to 20 minute delays in both directions.”, both roads 15th & Pennsylvania Ave NW will be located to find the specific coordinates. This tweet text based approach normally builds a place name dictionary and creates entity annotations to generate candidate place names first, and then converts them to a pair of longitude and latitude based on string matching method. With respect to extracting location mentions from tweets, a natural language processing tool named Stanford Named Entity Recognizer (NER) (Manning et al., 2014) can be used to label sequences of words in a text with person, organization, and location. Schulz et al. (2013) applied the Stanford NER model to recognize location mentions in a tweet and labelled them with two entities: location (e.g., cities, streets and landmarks) and place (e.g., “home”, “office”, and “school”) to accurately and abstractly describe where the event took place. Similarly, Gutiérrez et al. (2015) tested the performance of four NER engines, including Alchemy, OpenCalais, Stanford NER, and NERD, and chose NERD as a proposed tool with the best performance of detection. Instead of using a general NER to identify traffic entities in the text, a special tag set, which is more relevant to the task of extracting traffic information, was designed by Nguyen et al. (2016). They also set up an online collaborative environment based on Brat annotation tool to support rapid labelling of the tweets. Tejaswin et al. (2015) used a regular parser to generate candidate entities, which were further disambiguated to accurate locations based on the background information from Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008). The verified results were clustered into grids of varying sizes covering a larger area than single geolocations. Multi-source Linear Referencing Methods (LRM) were adopted for structural labelling based on expert rules in processing Weibo data (Cui et al., 2014).

Further, some studies geocoded the identified location mention to a point based on language matching algorithms, such as exact string matching and fuzzy string matching (Gu et al., 2016; Ribeiro Jr. et al., 2012). A set of location mention n-grams were matched with the geographical database such as GeoNames (Schulz et al., 2013), Wikipedia and Hatena keyword (Sakaki et al., 2012), gazetteer (Gu et al., 2016; Ribeiro Jr. et al., 2012; Wang et al., 2015), and OpenStreetMap data

(Daly et al., 2013; Schulz et al., 2013), so as to complete the fine-grained geo-localization. Moreover, Wanichayapong et al. (2011) geo-coded a traffic event to a link when the location information has all the attributes: Road, Start point and End point.

There is a possibility that a user who posts a tweet about a traffic event is at a different location other than the location of the event, which leads to "location deviation" in geo-locating a traffic event. This phenomenon can be resulted from information diffusion. Once the event has been mentioned in Twitter or reported in traditional media channels, people at distant locations can start posting tweets about that event (Ozdikis et al., 2017). Such problem of location deviation does not exist when tweet texts is analyzed to infer the location of the posted events since the inference is only based on the specific road names mentioned in tweets no matter where the tweets are posted. However, it is a challenge for directly using tagged GPS coordinates as the location of the posted events. Heravi, Morrison, Khare, and Marchand-Maillet (2014) suggested searching for specific predetermined words in tweets in order to identify users who were eyewitnesses to an event. Abdelhaq, Gertz, and Armiti (2017) referred to tweets that mention event-related keywords but posted from distant locations as spatial outliers, and proposed a method that used the co-occurrences of keywords with geo-references in tweet text to minimize the effect of these outliers. Crooks, Croitoru, Stefanidis, and Radzikowski (2013) suggested that early tweets posted shortly after an earthquake provided fast and useful spatial approximation based on their GPS geotags. A possible solution to handle such dynamics can be using a weighing model that assigns varying weights to spatial features (i.e., geographical coordinates, tweet texts and user profiles) in tweets at different stages of events (Ozdikis et al., 2017).

7. Traffic events summarization and visualization

Once a traffic event is detected from the social media platforms, it needs to be presented to the user in a meaningful and informative way, as the tweets themselves do not necessarily provide a good summary of an event. The detected traffic events are visualized in some proposed systems to give the user an intuitive view of the events. Some example screenshots of these visualization systems are illustrated in Fig. 4.

Steds (Fu, Lu, et al., 2015) designed an extractive and centroid based summarization method for redundant traffic related tweets, for deducting multiple similar contents to generate a concise and comprehensible summary of textual contents. Each individual tweet acted as a node in a complete graph, and the idf-modified-cosine similarity of tweets corresponded to the edge between nodes. Those edges with a cosine similarity < 0.6 was removed in order to identify the most salient tweet among the similar tweets set. After the edge removal process, the summarization algorithms based on LexRank was applied to this graph. The top ranked sentences were considered as the summary for the documents. In a similar way, Li et al. (2012) ranked the tweets according to their importance using a learning-to-rank approach, which was based on a linear regression model aggregating content features, user features, and usage features.

TrafficWatch (Nguyen et al., 2016), as shown in Fig. 4(a), aggregated the live tweet stream into different types of traffic events with common keywords used. This was done by applying an unsupervised clustering algorithm combining cosine similarity and Hamming distance evaluation. Users were presented a general view of the popular incident types and growing pattern in the tweets as they emerge over time. After pre-processing and classification, only the relevant and geo-located tweets were loaded onto the Cesium Bing map within a few seconds from time of posting for live traffic monitoring. Kosala and Steven (2012) received the request from users to load traffic information at a specific time and location, and determined the confidence level of the information and tallied the tweets that were directly related to the desired information as a summary. The real-time traffic information was visualized on Google map.

Traffic Observatory (Ribeiro Jr. et al., 2012) displayed the manually labelled traffic related tweets using a kernel density map. Sakaki et al. (2012) (Fig. 4(b)) demonstrated information extracted from both legacy media and social media to notify drivers of important events in a timely manner. Endarnoto et al. (2011) developed an Android mobile application to display map view with the information of traffic condition. Another Android based application publishing traffic status was proposed by Cui et al. (2014) using Weibo data. It was able to interactively communicate with users to acquire more accurate information of time, location, and status based on a Question and Answer (QA) mechanism.

8. Miscellaneous approaches

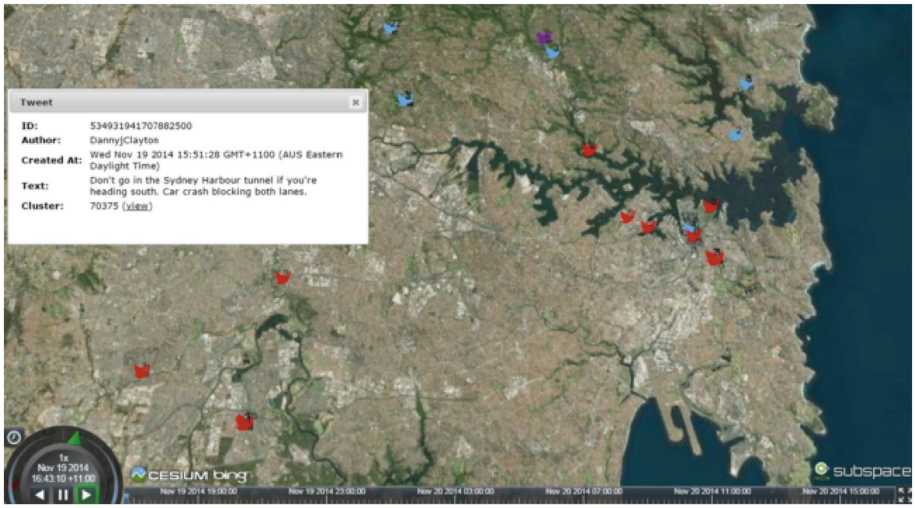
This section discusses the traffic events detection methods that adopt hybrid techniques, which do not directly fall under the typical framework discussed in Section 2. There exists a conflict on the relationship between traffic congestions and traffic events. As described in the beginning, Gutiérrez et al. (2015) regarded traffic congestion as a kind of traffic event. While (Kwon, Mauch, and Varaiya (2006) hold the view that traffic congestion is caused by multiple traffic events including incidents, special events, lane closures, adverse weather, potential ramp metering gain, and excess demand. The articles analyzing the causes to estimate traffic congestion are summarized in this section as miscellaneous approaches.

Liu et al. (2014) presented an application for Traffic Event Detection and Summary (TEDS), which enabled users to analyze traffic events in specific locations of interest in Washington D.C. Tweets were collected from the preselected influential traffic accounts. The location information embedded in tweets was combined with traffic data to build a concept graph, and the incorrect data such as invalid dates and locations were filtered out in the meantime. An individual tweet words unigram was constructed as raw signals, which was applied in Wavelet analysis to display the top-ranked words based on the type of traffic event. The similarity between two signals was measured by pair-wise cross correlation, and trivial words were filtered out. The remaining words were clustered to form events with a modularity-based graph partitioning technique. A natural language processing based summarization approach was adopted to summarize the clustered words to a long abstract instead of some topic words. Finally, the Python Image Library (PIL) was utilized to visualize a heat map showing the number of traffic events in each time interval, which allowed users to understand the traffic events by querying a particular traffic intersection or traffic landmark.

Gong, Deng, and Sinnott (2015) only focused on collecting tweets that were posted along street networks other than tweets with specific traffic related keywords or in a location bounding box. A range of centroids based on the diameter of the road and the central line of the road were calculated to harvest tweets along street networks through Twitter REST APIs. The crawled tweets were spatiotemporally clustered using DBSCAN method based on the assumption that if there were four or more tweets sent in a 1 km stretch within 15 min on main roads, it was considered as a possible traffic event. Triangles and points were used to represent the clustered results (i.e., potential traffic events) and the standalone tweets, respectively. It allowed user to pick a particular date, a time window (morning rush, evening rush or off peak), a day of week or a street to view the mapping of tweets. In this way, users had an easy access to identify rush hour, and/or incidents that may take place around the city of Melbourne.

Moreover, individuals' sentiment can also be identified while analyzing the tweets text. It is assumed that individuals in traffic jams would have an increased proportion of negative sentiment, thus quantifying this and analyzing how such sentiment changes through the day seems to be an extension of the typical way to identify traffic events.

Semwal et al. (2015) integrated tweets, Facebook public page posts with other data sources such as weather data and news data to get updates about the traffic. It assumed that a negative sentiment



(a). A screenshot of TrafficWatch (Nguyen et al., 2016)

(b). A screenshot of Practical Hybrid Event Visualizer (Sakaki et al., 2012)

Fig. 4. Screenshots of some example traffic event visualization systems.

concerning an event/problem was likely to have an impact on traffic of corresponding location and time period. To be specific, once a spike in the number of tweets talking about an event/problem was detected at a location over a period of time, the tweets associated with the detected spike were then analyzed to evaluate their sentiments. Presence of a negative sentiment may lead to severe traffic issues. Further, apriori algorithm was used for viral event detection so that the frequently co-occurring problems and causes could be identified to predict the prevalence of a problem and provide suggestions for the detected problem. Similarly, (Salas, Georgakis, Nwagboso, Ammari, and Petalas (2017) classified the geocoded traffic tweets into positive, negative, and neutral class using sentiment analysis. A lexical approach was further applied to identify the level of stress and relaxation within the tweet to better understand user's emotions expressed in short messages. It assigned two values on a scale from 1 to 5 for relaxation and -1 to -5 for stress.

As mentioned in Section 1, some studies consider multiple events as causes to estimate traffic congestion. Daly et al. (2013) built a Dub-Star (Dublin's Semantic Traffic Annotator and Reasoner) system by exploring the underlying reasons behind traffic congestion based on semantic web technologies. Historical data regarding road conditions, congestion and events in the city were used to train an off-line diagnoser from both spatial and temporal dimensions. The links between events and congestion could be inferred using semantic matching technologies. Once the real-time event was identified as a potential cause to the congestion, the spatiotemporal relationship between them was investigated to evaluate the confidence level. As a result, it could present the best candidate responses to the user according to his geocoded requests sent from social media and SMS messages. A similar approach was proposed by Lécué et al. (2013) and Lécué et al. (2014) to make traffic status prediction based on exploring the connection between traffic congestion and its causes using semantic web technologies.

In addition, Wang et al. (2015) estimated traffic congestion considering the potential impacts of diverse events, such as parties, music shows, and sports, which were directly extracted from Twitter stream. The spatiotemporal correlations between events and congestion were investigated among the road segments. A two-dimensional Gaussian model was proposed to measure the impact intensity of the diverse events on the nearby road segments based on their Euclidean distance. Traffic congestion could be estimated using a coupled matrix and tensor factorization model, which integrated multi-typed traffic information and events together. The estimated traffic states were as follows: heavy congestion, medium-heavy congestion, medium, light, and flow

Table 6
Summary of the evaluation approaches and their performance metrics.

Articles	Baselines	Examples	Evaluation metrics			
			Precision	Recall	F_1-score	Accuracy
Sakaki et al. (2012)	Tweets tagged by humans	Heavy traffic	0.87	0.67	0.75	×
		Traffic restrictions	0.57	0.84	0.68	×
		Police checkpoints	0.74	0.93	0.83	×
		Rain	0.56	0.93	0.70	×
		Mist	0.56	0.83	0.67	×
Anantharam, Barnaghi, Thirunarayan, and Sheth (2015)	Incident reports from 511.org	N/A	×	×	×	0.40
D'Andrea et al. (2015)	70 traffic events with confirmation of official traffic news websites or local newspapers	In advance	×	×	×	0.44
		≤ 15 min	×	×	×	0.36
		15–50 min	×	×	×	0.20
Tejaswin et al. (2015)	Positive and negative samples	True traffic event	0.94	0.89	0.92	0.88
		False traffic event	0.75	0.86	0.80	
Gu et al. (2016)	RCRS and CFS data	30-min reporting time discrepancy and 1-mile distance	×	×	×	0.71
	HERE travel time data	The threshold for the significance level is 0.1	×	×	×	≥ 0.28

conditions, which were assigned with values in congestion matrix 1.0, 0.8, 0.6, 0.4, and 0.2, respectively.

9. Performance evaluation

In this section, we discuss the methods used to evaluate the performance of traffic event detection from a selected number of articles (summarized in Table 6), where the detected results are compared to the baseline available to their study with different metrics such as accuracy, precision, recall, and F_1-score. It is worth mentioning that all these evaluation approaches did not have a common baseline, which makes comparison of their results difficult. Ideally, if a baseline based on a simple approach can be found, it would make more sense to compare the results of these different approaches to that of the baseline approach. However, no such a baseline has been identified through our literature study.

As shown in Fig. 5, a confusion matrix (Kohavi & Provost, 1998) shows information about actual and predicted classes, both of which contain positive and negative instances. By comparing the predicted instances against the actual instances, four cases are obtained as follows: True Positive (TP), False Negative (FN), False Positive (FP) and True Negative (TN). To be specific, TP is the number of correct predictions that both predicted instance and actual instance are positive; FN is the number of incorrect predictions that the predicted instance is negative while the actual instance is positive; FP is the number of incorrect predictions that the predicted instance is positive while the actual instance is negative; and TN is the number of correct predictions that both predicted instance and actual instance are negative. They are the basis for calculating evaluation metrics in Table 6, which are accuracy, precision, recall, and F1-score.

Accuracy is the most intuitive metric to measure the performance. It is simply the proportion of correct predictions, which is calculated as Eq. (1). The value of accuracy is between 0 and 1. It seems that the higher the accuracy, the better and more useful the method is.

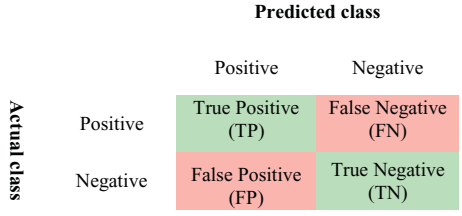


Fig. 5. Confusion matrix.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Precision is the ratio of the number of correct positive predictions to the total number of positive predictions, which is calculated as Eq. (2). The value of precision ranges from 0 to 1, where a perfect precision value of 1 for class A means that every instance labelled as belonging to class A does indeed belong to class A, but says nothing about the number of instances from class A that were not labelled correctly.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall is also known as sensitivity or true positive rate. It is the ratio of the correct positive predictions to the number of all actual positive instances, which is calculated as Eq. (3). The value of recall ranges from 0 to 1, where a perfect recall value of 1 for class A means that all instances from class A are labelled as belonging to class A, but says nothing about how many other instances are also incorrectly labelled as belonging to class A.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

Usually, there is an inverse relationship between precision and recall, where it is possible to increase one at the cost of reducing the other. Thus, precision and recall are not discussed in an isolated way. Instead, either values for one measure are compared for a fixed level at the other measure (e.g. precision at a recall level of 0.75) or both are combined into a single measure.

F1-score is a matter of balancing measure that combines precision and recall into a single value. It is the harmonic mean of precision and recall, which is extremely useful in most scenarios when we are working with imbalanced datasets (i.e., a dataset with a non-uniform distribution of class labels). It is calculated as Eq. (4), where the best F1 – score has its value at 1 and worst score at the value 0 In a word, different metrics give us different and valuable insights into how a proposed method performs.

$$F_1 = 2 \bullet \frac{\text{Precison} \bullet \text{Recall}}{\text{Precison} + \text{Recall}} = \frac{2TP}{2TP + FP + FN} \tag{4}$$

Sakaki et al. (2012) evaluated the performance of SVM classifier for each class by comparing to tweets tagged by humans with precision, recall and F-score. Reasonable values revealed that the proposed system made it possible to steadily assist in driving. Anantharam et al. (2015) evaluated the effectiveness of the extracted traffic events by investigating if they were corroborative, complementary, or timely. The incident reports from 511.org were selected as ground truth, and it

indicated that around 40% of them coexisted with spatiotemporal constraints. A lot more potential traffic events, which were not reported on 511.org, could also be discovered, but there was no ground truth for verification. D'Andrea et al. (2015) validated their approach using a meaningful set of traffic events confirmed by official traffic news websites or local newspapers. More precisely, 31 events were correctly detected in advance of traffic news websites or local newspapers, 25 events were detected within 15 min, and the remaining 14 events were detected beyond 15 min but within 50 min.

Tejaswin et al. (2015) created positive and negative samples as baselines in the city New Delhi, and gained a mean accuracy of 88.46%. Gu et al. (2016) used both incidents datasets (i.e., RCRS (Road Condition Report System) data and CFS (911 Call For Service (CFS) data) and HERE travel time data for validation. The location, time and category of traffic events extracted from geocoded tweets were compared with existing datasets to examine the number of Twitter based events that were reported by existing datasets, and the number of additional detected events that were not reported by existing datasets. When allowing 30-min reporting time discrepancy and 1-mile distance for both Twitter and incidents data to report the same incident, 71% of the entire set of RCRS + CFS incidents could be found in Twitter. HERE travel time data provided evidence whether the additional Twitter-reported events are likely to be true. When the threshold for the significance level was 0.1, it revealed that at least 28% of Twitter-reported incidents were true. The results demonstrated that mining geosocial media data held great potentials to complement incident report sources in a very efficient and cost-effective way.

10. Discussion

Considering the complexity of the real world, some approaches reviewed in this paper need to deal with practical issues when applied in a large scale. Detecting traffic events from geosocial media data in a real-time manner requires the capability of handling high volume of streaming data (Li et al., 2016). While some approaches are not capable of processing dynamic updates, and can only use a limited number of sample tweets to make offline analysis (Endarnoto et al., 2011). A list of challenges and opportunities in detecting traffic events using geosocial media data are summarized in Table 7.

Querying for the traffic related tweets containing one or more traffic related keywords or posted by traffic authorities is the basis for sensing and detecting traffic events. The quality of the queried dataset may directly influence the efficiency of following stages described in Sections 4–7. As is shown in Tables 1 and 2, different keywords are selected in different articles. Applying too many traffic related keywords will result in more noise and redundancy. In contrast, it will not be able to extract all the traffic related tweets. Kuflik et al. (2017) semi-automatically constructed a dictionary of traffic-related terms as a solution based on their presumed association with the traffic domain. However, this approach involves domain experts manually assessing the list of selected terms and assigning them a relevance score, which is time consuming and high cost.

There is a lack of traffic-specific dictionary and accompanying context analysis for automating text processing at the time of the research or to the author's best knowledge “Ontologies” serve as a methodological framework for representing contextual information as a networked structure of objects or concepts, with related items linked by labelled relationships. In terms of traffic domain, it allows an association of the text with transport categories at various granularities (e.g., transport mode and traffic condition) (Grant-Muller et al., 2014). Therefore, a transportation ontology defining a comprehensive set of traffic related terms needs to be developed by taking advantage of collaborative intelligence and drawing on contributions by non-experts. It will assist in training the classifiers and overcoming human annotator disagreement.

A classifier is built to identify the relevance of a tweet to a real-world traffic event based on feature extraction. Integration of specific keywords, temporal features, geospatial features, and social network specific to the user, which are listed in Tables 3 and 4, seems to be a potential way for features representation. With respect to the classifier mentioned in this paper, it is usually assumed that a static environment, and typically trained an offline classifier using a relatively small batch of geosocial media data labelled manually. Then the trained classifier is directly deployed to detect traffic event. Since the geosocial media sites keep updating constantly, it is likely to result in both false positive and negative errors with the offline classifier. Techniques such as incremental learning (Joshi & Kulkarni, 2012) and ensemble methods (Khreich, Granger, Miri, & Sabourin, 2012; Kuncheva, 2004; Polikar, 2006) may be employed to account for unseen events and adapt to changes that may occur over time.

Regarding the process of manually labeling the tweets to train a classifier, the limited datasets are unbalanced with positive and negative samples. Obtaining enough labelled tweets may ease this problem. The use of crowdsourcing for the manual labelling of examples based on the above-mentioned ontology is an available way to reduce the cost of labelling and achieve the agreement between human annotators. It is

Table 7
Challenges and opportunities of traffic events detection using geosocial media data.

Challenges	Opportunities
• There is a lack of traffic-specific dictionary and accompanying context analysis to query for the traffic related tweets.	• A transportation ontology defining a comprehensive set of traffic related terms needs to be developed by taking advantage of collaborative intelligence and drawing on contributions by non-experts.
• The offline classifier trained with a batch of manually labelled geosocial media data cannot meet the requirement of continuous updates. • The limited dataset used to train classifier is unbalanced with positive and negative samples.	• Techniques such as incremental learning and ensemble methods may be employed to account for unseen events and adapt to changes that may occur over time. • The classification is conducted with a larger number of labelled examples (e.g., crowdsourcing) and application of balancing techniques (e.g., resampling and model penalization).
• Some traffic related expressions are neglected by natural language processing techniques when extracting location information from tweets. • A common problem in most studies that utilize geosocial media data is how to ensure proper correlation between event location and posting location, which are sometimes different.	• More traffic specific rules should be considered to improve the performance of general natural language processing techniques. • Integrating location inference methods and fusing multiple data sources may be worth further investigations to develop better approach.
• Traffic events are passively presented to users based on their requests of specific location, time and keywords. • There is not enough resources of ground truth to evaluate the detected traffic events.	• Applying subscribe/publish mechanisms to notify users before they arrive in the potential emergency area. • Taking advantage of multiple social media sources such as Facebook Flickr, and local news websites will be a robust solution, as missing information from one source may be available in the other. • Combining social media data with traditional traffic sensors data is likely to be an efficient way for traffic anomalies detection and long-term traffic prediction.

suggested that the classification is conducted with a larger number of labelled examples and application of balancing techniques (Kuflik et al., 2017). Resampling technique and model penalization technique are good solutions to the sample imbalance problem.

Fu, Lu, et al. (2015), Fu, Nune, and Tao (2015), Fu, Zhong, et al. (2015), and D'Andrea et al. (2015) only focused on extracting traffic events by analyzing tweets semantics, but neglected the spatial dimension to identify the event location. Location information is more or less included in the traffic relevant tweet, since people tend to describe where the traffic event takes place when they observe it. Natural language processing techniques are widely used in pre-processing raw tweets and extracting location information from tweet texts. It enables normalization of the tweets as well as the removal of stop words. When it is used to extract location mentions, there commonly exist some abbreviations related to traffic domain, such as “av” is short for “avenue”, “st” is short for “street”. These expressions are likely to be neglected or cannot be annotated to the correct words with general natural language processing tools. Thus, more traffic specific rules should be considered to improve the performance of general natural language processing techniques.

Due to the fact that traffic related tweets may be posted at different locations from where traffic events actually occur, location deviation does exist when tagged GPS coordinates are used to extract location information for detected events. As indicated in Section 3, it is not a problem for location inference by using specific mentions of road names in tweet texts, which seems to be an available source to make up for this deficiency. In addition, information about one event is likely to be found in different sources since the way user reports the event varies with preference. For example, one user posts a traffic accident through Twitter while another user prefers to report the accident using WAZE application. Collecting multiple data sources provides more references for location inference. A hybrid model that balances the priority among different location inference methods and fuses multiple data sources can be trained to efficiently estimate the location for traffic events so that the location deviation would be reduced to the least.

After the traffic event is detected and geocoded, it is usually summarized and visualized on maps. It passively presents a general view of the traffic event based on user's request of specific location, time and keywords. To provide timely situational awareness and help users make right decisions, it is better to notify them before they arrive in the potential emergency area. The subscribe-publish mechanism can be a good choice, which enables the event detection system to actively push the real-time traffic information to users about their previous subscriptions, such as usual driving routes, places of interest, day of week, and time of day, through text messages or targeted mobile applications.

As summarized in Section 9, the proposed approaches are evaluated using baselines, such as manually labelled tweets and official traffic reports. Additional detected traffic events that are not reported in the baseline cannot be verified (Anantharam et al., 2015). The lack of a publicly available text corpus, along with limited resources of ground truth for a system performance evaluation, is a major obstacle for the traffic events detection techniques discussed in our survey, which mainly focused on detecting traffic events from Twitter stream. Taking advantage of multiple social media sources such as Facebook Flickr, Youtube, and local news websites will be a robust solution, as missing information from one source may be available in the other. In addition, combining social media data with traditional traffic sensors data is likely to be an efficient way for traffic anomalies detection (Pan, Zheng, Wilkie, & Shahabi, 2013) and long-term traffic prediction (He, Shen, Divakaruni, Wynter, & Lawrence, 2013).

11. Conclusion

Detecting traffic events provides valuable information for drivers and traffic management agencies to deal with abnormal traffic conditions, which were mainly conducted by using physical sensors and crowdsourcing data in the past. However, the high cost, sparsity, and limited spatial coverages of physical sensors, and data sharing limitations of crowdsourcing raise considerable challenges to effectively detect traffic events in a real-time manner. As a widely available and free online social networking service, geosocial media platforms allow users to share situational and actionable content with spatial and temporal information at a low cost, which could reveal real-world traffic events as users post content in real-time. An extensive review of noteworthy and recent traffic event detection techniques from geosocial media data streams was done to broadly organize the following approaches based on a typical processing workflow:

- Keywords and Accounts: Traffic related keywords and/or influential traffic accounts are mainly selected to query traffic related tweets from geosocial media data streams
- Natural Language Processing (NLP): NLP techniques are applied to pre-process the raw geosocial media messages since there exist a large amount of mundane information (e.g., meaningless, polluted, and rumor messages).
- Classification: Machine learning classification techniques have been adopted to distinguish real-world traffic events from the detected candidate traffic events since there exist messages containing traffic related keywords that do not actually refer to a real-world traffic event.
- Geocoding: Single or combinations of GPS coordinates, user profiles, and tweets texts create a solid foundation for the geocoding techniques to extract location information for the traffic events. Attempts to solve location deviation issues include identifying eyewitnesses, filtering spatial outliers and weighting spatial features of a sequence of tweets related to an event.
- Information Extraction: Traffic events are summarized by using text ranking and clustering methods rather than the simple tweets. The summarized results are visualized through mobile or web-based applications to help users intuitively understand the detected traffic event.

Finally, this paper highlights some major issues and open research improvements. A lack of traffic-specific dictionary and accompanying context analysis is likely to be relieved through ontology-based methods. The imbalance of positive/negative labelled samples poses great challenges to the machine learning classification process. Further work is required to propose effective measures (e.g., resampling technique and model penalization technique) to train a classifier so that real-world traffic events can be more accurately and efficiently extracted from geosocial media data. Publicly available ground truth data to evaluate the performance of different event detection approaches is needed to make comparisons between them. The introduction of subscribe/publish mechanisms in the event detection framework may potentially improve the efficiency of relevant responders and users in emergency situations. This paper provides a structured representation of the techniques applied in geosocial media-centric traffic events detection, which can be helpful for researchers working in this field directly or similar fields. The work done in the reviewed studies have also laid out foundation for detecting pre-and-post traffic patterns using social media content.

Conflict of interest statement

None.

Acknowledgement

This work has been funded by the Natural Science and Engineering Research Council of Canada (NSERC) (RGPIN-2017-05950). The authors would like to thank anonymous reviewers for their constructive comments, which helped improve the paper.

References

Abdelhaq, H., Gertz, M., & Armiti, A. (2017). Efficient online extraction of keywords for localized events in twitter. *GeoInformatica, 21*(2), 365–388. http://dx.doi.org/10.1007/s10707-016-0258-x.

Ajao, O., Hong, J., & Liu, W. (2015). A survey of location inference techniques on Twitter. *Journal of Information Science, 41*(6), 855–864. http://dx.doi.org/10.1177/0165551515602847.

Amer-Yahia, S., Anjum, S., Ghenai, A., Siddique, A., Abbar, S., Madden, S., ... El-Haddad, M. (2012). MAQSA: A system for social analytics on news. *Proceedings of the 2012 ACM SIGMOD international conference on management of data* (pp. 653–656). Scottsdale, Arizona, USA: ACM.

Anantharam, P., Barnaghi, P., Thirunarayan, K., & Sheth, A. (2015). Extracting city traffic events from social streams. *ACM Transactions on Intelligent Systems and Technology (TIST), 6*(4), 43.

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence, 31*(1), 132–164. http://dx.doi.org/10.1111/coin.12017.

Aziz, M. V. G., Prihatmanto, A. S., Henriyan, D., & Wljaya, R. (2015). Design and implementation of natural language processing with syntax and semantic analysis for extract traffic conditions from social media data. *2015 5th IEEE international conference on system engineering and technology (ICSET)* (pp. 43–48). Shah Alam, Malaysia: IEEE.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: A collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD international conference on management of data* (pp. 1247–1250). Vancouver, Canada: ACM. http://dx.doi.org/10.1145/1376616.1376746.

Bontcheva, K., & Rout, D. (2014). Making sense of social media streams through semantics: A survey. *Semantic Web, 5*(5), 373–403.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. *Proceedings of the 20th international conference on world wide web* (pp. 675–684). Hyderabad, India: ACM.

Chen, P.-T., Chen, F., & Qian, Z. (2014). Road traffic congestion monitoring in social media with hinge-loss Markov random fields. *2014 IEEE international conference on data mining* (pp. 80–89). http://dx.doi.org/10.1109/ICDM.2014.139.

Coifman, B., & Dhoorjaty, S. (2004). Event data-based traffic detector validation tests. *Journal of Transportation Engineering, 130*(3), 313–321.

Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). #Earthquake: Twitter as a distributed sensor system. *Transactions in GIS, 17*(1), 124–147. http://dx.doi.org/10.1111/j.1467-9671.2012.01359.x.

Cui, J., Fu, R., Dong, C., & Zhang, Z. (2014). Extraction of traffic information from social media interactions: Methods and experiments. *17th international IEEE conference on intelligent transportation systems (ITSC)* (pp. 1549–1554). http://dx.doi.org/10.1109/ITSC.2014.6957913.

Daly, E. M., Lécué, F., & Bicer, V. (2013). Westland row why so slow? Fusing social media and linked data sources for understanding real-time traffic conditions. *Proceedings of the 2013 international conference on intelligent user interfaces* (pp. 203–212). Santa Monica, California, USA: ACM.

D'Andrea, E., Ducange, P., Lazzerini, B., & Marcelloni, F. (2015). Real-time detection of traffic from twitter stream analysis. *IEEE Transactions on Intelligent Transportation Systems, 16*(4), 2269–2283. Retrieved from www.ijiset.com.

Demissie, M. G., de Almeida Correia, G. H., & Bento, C. (2013). Intelligent road traffic status detection system through cellular networks handover information: An exploratory study. *Transportation Research Part C: Emerging Technologies, 32*, 76–88. http://dx.doi.org/10.1016/j.trc.2013.03.010.

Dingli, A., Mercieca, L., Spina, R., & Galea, M. (2015). Event detection using social sensors. *2015 2nd international conference on information and communication technologies for disaster management (ICT-DM)* (pp. 35–41). Rennes, France: IEEE. http://dx.doi.org/10.1109/ICT-DM.2015.7402054.

Djahel, S., Doolan, R., Muntean, G.-M., & Murphy, J. (2015). A communications-oriented perspective on traffic Management Systems for Smart Cities: Challenges and innovative approaches. *IEEE Communication Surveys and Tutorials, 17*(1), 125–151. http://dx.doi.org/10.1109/COMST.2014.2339817.

Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM, 54*(4), 86–96. http://dx.doi.org/10.1145/1924421.1924442.

Doolan, R., & Muntean, G. M. (2017). EcoTrec-A novel VANET-based approach to reducing vehicle emissions. *IEEE Transactions on Intelligent Transportation Systems, 18*(3), 608–620. http://dx.doi.org/10.1109/TITS.2016.2585925.

Dunkel, J., Fernández, A., Ortiz, R., & Ossowski, S. (2011). Event-driven architecture for decision support in traffic management systems. *Expert Systems with Applications, 38*(6), 6530–6539. http://dx.doi.org/10.1016/j.eswa.2010.11.087.

Endarnoto, S. K., Pradipta, S., Nugroho, A. S., & Purnama, J. (2011). Traffic condition information extraction & visualization from social media twitter for android mobile application. *2011 international conference on electrical engineering and informatics (ICEEI)* (pp. 1–4). Bandung, Indonesia: IEEE. http://dx.doi.org/10.1109/ICEEI.2011.6021743.

Fu, K., Lu, C.-T., Nune, R., & Tao, J. X. (2015). Steds: Social media based transportation event detection with text summarization. *2015 IEEE 18th international conference on intelligent transportation systems (ITSC)* (pp. 1952–1957). Las Palmas, Spain: IEEE. http://dx.doi.org/10.1109/ITSC.2015.316.

Fu, K., Nune, R., & Tao, J. X. (2015). *Social media data analysis for traffic incident detection and management. Transportation research board 94th annual meeting (no. 15-4022)*.

Fu, K., Zhong, W., Lu, C., & Boedihardjo, A. P. (2015). Find the butterfly: A social media based arterial incidents detection and causality analysis system. *Proceedings of the 23nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* (pp. 99). Seattle, Washington: ACM. http://dx.doi.org/10.1145/2820783.2820797.

Gal-Tzur, A., Grant-Muller, S. M., Kuflik, T., Minkov, E., Nocera, S., & Shoor, I. (2014). The potential of social media in delivering transport policy goals. *Transport Policy, 32*, 115–123. http://dx.doi.org/10.1016/j.tranpol.2014.01.007.

Gal-Tzur, A., Grant-Muller, S. M., Minkov, E., & Nocera, S. (2014). The impact of social media usage on transport policy: Issues, challenges and recommendations. *Procedia - Social and Behavioral Sciences, 111*, 937–946. http://dx.doi.org/10.1016/j.sbspro.2014.01.128.

Garg, M., & Kumar, M. (2016). Review on event detection techniques in social multimedia. *Online Information Review, 40*(3), 347–361. http://dx.doi.org/10.1108/OIR-08-2015-0281.

Giridhar, P., Abdelzaher, T., George, J., & Kaplan, L. (2015). On quality of event localization from social network feeds. *2015 IEEE international conference on pervasive computing and communication workshops, PerCom workshops 2015* (pp. 75–80). http://dx.doi.org/10.1109/PERCOMW.2015.7133997.

Gong, Y., Deng, F., & Sinnott, R. O. (2015). Identification of (near) real-time traffic congestion in the cities of Australia through Twitter. *Proceedings of the ACM first international workshop on understanding the city with urban informatics* (pp. 7–12). Melbourne, Australia: ACM. http://dx.doi.org/10.1145/2811271.2811276.

Goswami, A., & Kumar, A. (2016). A survey of event detection techniques in online social networks. *Social Network Analysis and Mining, 6*(107), 1–25. http://dx.doi.org/10.1007/s13278-016-0414-1.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Kuflik, T., Nocera, S., & Shoor, I. (2015). Transport policy: Social media and user-generated content in a changing information paradigm. *Social media for government services. 2015. Social media for government services* (pp. 325–366). Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-27237-5.

Grant-Muller, S. M., Gal-Tzur, A., Minkov, E., Nocera, S., Ku, T., & Shoor, I. (2014). Enhancing transport data collection through social media sources: Methods, challenges and opportunities for textual data. *IET Intelligent Transport Systems, 9*(4), 407–417. http://dx.doi.org/10.1049/iet-its.2013.0214.

Gu, Y., Qian, Z.(. S.)., & Chen, F. (2016). From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies, 67*, 321–342. http://dx.doi.org/10.1016/j.trc.2016.02.011.

Guralnik, V., & Srivastava, J. (1999). Event detection from time series data. *Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 33–42). San Diego, California, USA: ACM.

Gutiérrez, C., Figuerias, P., Oliveira, P., Costa, R., & Jardim-Goncalves, R. (2015). Twitter mining for traffic events detection. *2015 science and information conference* (pp. 371–378). London, UK: IEEE. http://dx.doi.org/10.1109/SAI.2015.7237170.

Harrison, C., & Donnelly, I. A. (2011). A theory of smart cities. *Proceedings of the 55th annual meeting of the ISSS - 2011, Hull, UK, (Proceedings of the 55th annual meeting of the ISSS)* (pp. 1–15). http://dx.doi.org/10.1017/CBO9781107415324.004.

Hasan, M., Orgun, M. A., & Schwitter, R. (2017). A survey on real-time event detection from the Twitter data stream. *Journal of Information Science,* (November 2015), 1–21. http://dx.doi.org/10.1177/0165551517698564.

He, J., Shen, W., Divakaruni, P., Wynter, L., & Lawrence, R. (2013). Improving traffic prediction with tweet semantics. *Proceedings of the twenty-third international joint conference on artificial intelligence* (pp. 1387–1393).

Heravi, B. R., Morrison, D., Khare, P., & Marchand-Maillet, S. (2014). Where is the news breaking? Towards a location-based event detection framework for journalists. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8326 LNCS(PART 2)* (pp. 192–204). http://dx.doi.org/10.1007/978-3-319-04117-9_18.

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine, 14*(6), 1–5. http://dx.doi.org/10.1086/599595.

Hurlock, J., & Wilson, M. L. (2011). Searching Twitter: Separating the tweet from the chaff. *Proceedings of the fifth international AAAI conference on weblogs and social media searching* (pp. 161–168).

Ihler, A., Hutchins, J., & Smyth, P. (2006). Adaptive event detection with time-varying poisson processes. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 207–216). Philadelphia, PA, USA: ACM.

Joshi, P., & Kulkarni, P. (2012). Incremental learning: Areas and methods-a survey. *International Journal of Data Mining & Knowledge Management Process, 2*(5), 43–51.

Kamijo, S., Matsushita, Y., Ikeuchi, K., & Sakauchi, M. (2000). Traffic monitoring and accident detection at intersections. *IEEE Transactions on Intelligent Transportation Systems, 1*(2), 108–118.

Kamran, S., & Haas, O. (2007). A multilevel traffic incidents detection approach: Identifying traffic patterns and vehicle behaviours using GPS data. *Intelligent vehicles symposium, 2007 IEEE* (pp. 912–917). Istanbul, Turkey: IEEE. http://dx.doi.org/10.1109/IVS.2007.4290233.

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons, 53*(1), 59–68. http://dx.doi.org/10.1016/j.bushor.2009.09.003.

Kelley, M. J. (2013). The emergent urban imaginaries of geosocial media. *GeoJournal, 78*(1), 181–203. http://dx.doi.org/10.1007/s10708-011-9439-1.

Khreich, W., Granger, E., Miri, A., & Sabourin, R. (2012). Adaptive ROC-based ensembles of HMMs applied to anomaly detection. *Pattern Recognition, 45*(1), 208–230. http://dx.doi.org/10.1016/j.patcog.2011.06.014.

Kohavi, R., & Provost, F. (1998). Glossary of terms. *Machine Learning, 30*, 271–274.

Kosala, R., & Steven, E. A. (2012). Harvesting real time traffic information from Twitter. *Procedia Engineering, 50*, 1–11. http://dx.doi.org/10.1016/j.proeng.2012.10.001.

Kousiouris, G., Akbar, A., Sancho, J., Ta-Shma, P., Psychas, A., Kyriazis, D., & Varvarigou, T. (2018). An integrated information lifecycle management framework for exploiting social network data to identify dynamic large crowd concentration events in smart

cities applications. *Future Generation Computer Systems, 78*, 516–530. http://dx.doi.org/10.1016/j.future.2017.07.026.

Kryvasheyeu, Y., Chen, H., Obradovich, N., Moro, E., Van Hentenryck, P., Fowler, J., & Cebrian, M. (2016). Rapid assessment of disaster damage using social media activity rapid assessment of disaster damage using social media activity. *Science Advances, 2*(3), e1500779. http://dx.doi.org/10.1126/sciadv.1500779.

Kuflik, T., Minkov, E., Nocera, S., Grant-Muller, S., Gal-Tzur, A., & Shoor, I. (2017). Automating a framework to extract and analyse transport related social media content: The potential and the challenges. *Transportation Research Part C: Emerging Technologies, 77*, 275–291. http://dx.doi.org/10.1016/j.trc.2017.02.003.

Kumar, A., Jiang, M., & Fang, Y. (2014). Where not to go? Detecting road hazards using Twitter. *Proceedings of the 37th international ACM SIGIR conference on research & development in information retrieval* (pp. 1223–1226). Gold Coast, Queensland, Australia: ACM.

Kuncheva, L. I. (2004). Classifier ensembles for changing environments. *Multiple Classifier Systems, 3077*, 1–15.

Kurniawan, D. A., Wibirama, S., & Setiawan, N. A. (2016). Real-time traffic classification with Twitter data mining. *2016 8th international conference on information technology and electrical engineering (ICITEE)* (pp. 165–169). Yogyakarta, Indonesia: IEEE. http://dx.doi.org/10.1109/ICITEED.2016.7863251.

Kwon, J., Mauch, M., & Varaiya, P. (2006). Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record: Journal of the Transportation Research Board, 1959*, 84–91.

Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M., & Tommasi, P. (2013). Predicting severity of road traffic congestion using semantic web technologies. *ESWC 2014: The semantic web: Trends and challenges* (pp. 611–627). Cham: Springer. http://dx.doi.org/10.1007/978-3-319-07443-6_41.

Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M. L., & Tommasi, P. (2014). STAR-CITY: Semantic traffic analytics and reasoning for city. *Proceedings of the 19th international conference on intelligent user interfaces* (pp. 179–188). Haifa, Israel: ACM.

Leduc, G. (2008). Road traffic data: Collection methods and applications. *EUR number: Technical note: JRC 47967, JRC 47967. January 2008. EUR number: Technical note: JRC 47967, JRC 47967* (pp. 55–). JRC 47967–2008.

Lee, K., Eoff, B. D., & Caverlee, J. (2011). Seven months with the Devils: A long-term study of content polluters on Twitter. *Proceedings of the fifth international AAAI conference on weblogs and social media* (pp. 185–192).

Lee, R., & 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* (pp. 1–10). San Jose, California: ACM.

Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012). TEDAS: A twitter-based event detection and analysis system. *2012 IEEE 28th international conference on data engineering (ICDE)* (pp. 1273–1276). Washington, DC, USA: IEEE. http://dx.doi.org/10.1109/ICDE.2012.125.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., ... Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing, 115*, 119–133. http://dx.doi.org/10.1016/j.isprsjprs.2015.10.012.

Li, X., & Porikli, F. M. (2004). A hidden markov model framework for traffic event detection using video features. *International Conference on Image Processing, 2004. ICIP '04. 2004* (pp. 2901–2904). Singapore, Singapore: IEEE. http://dx.doi.org/10.1109/ICIP.2004.1421719.

Liu, M., Fu, K., Lu, C.-T., Chen, G., & Wang, H. (2014). A search and summary application for traffic events detection based on Twitter data. *Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 549–552). Dallas, Texas: ACM. http://dx.doi.org/10.1145/2666310.2666366.

Lv, Y., Chen, Y., Zhang, X., Duan, Y., & Li, N. (2017). Social media based transportation research: The state of the work and the networking. *IEEE/CAA Journal of Automatica Sinica, 4*(1), 19–26.

Mai, E., & Hranac, R. (2013). *Twitter interactions as a data source for transportation incidents. In transportation research board 92nd annual meeting (No. 13-1636)*DC, USA: Washington. Retrieved from http://docs.trb.org/prp/13-1636.pdf.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: System demonstrations* (pp. 55–60). Stroudsburg, PA, USA: Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/P14-5010.

Martchouk, M., Mannering, F., & Bullock, D. (2011). Analysis of freeway travel time variability using bluetooth detection. *Journal of Transportation Engineering, 137*(10), 697–704. http://dx.doi.org/10.1061/(ASCE)TE.1943-5436.0000253.

Metaxas, P. T., & Mustafaraj, E. (2012). Social media and the elections. *Science, 338*(6106), 472–473.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? *Comparing Data from Twitter's Streaming API with Twitter's Firehose* (pp. 400–408). http://dx.doi.org/10.1007/978-3-319-05579-4_10.

Nguyen, H., Liu, W., Rivera, P., & Chen, F. (2016). TrafficWatch: Real-time traffic incident detection and monitoring using social media. *PAKDD 2016: Advances in knowledge discovery and data mining* (pp. 540–551). Cham: Springer. http://dx.doi.org/10.1007/978-3-319-31753-3.

Novaco, R. W., & Gonzalez, O. I. (2009). Commuting and well-being. *Technology and Psychological Well-being*. http://dx.doi.org/10.1017/CBO9780511635373.008.

Nurwidyantoro, A., & Winarko, E. (2013). Event detection in social media: A survey. *2013 international conference on ICT for smart society (ICISS)* (pp. 1–5). Jakarta, Indonesia: IEEE.

Ozdikis, O., Oğuztüzün, H., & Karagoz, P. (2017). A survey on location estimation techniques for events detected in Twitter. *Knowledge and Information Systems, 52*(2), 291–339. http://dx.doi.org/10.1007/s10115-016-1007-z.

Pan, B., Zheng, Y., Wilkie, D., & Shahabi, C. (2013). Crowd sensing of traffic anomalies based on human mobility and social media. *Proceedings of the 21st ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 344–353). Orlando, Florida: ACM.

Park, H., & Haghani, A. (2016). Real-time prediction of secondary incident occurrences using vehicle probe data. *Transportation Research Part C, 70*, 69–85. http://dx.doi.org/10.1016/j.trc.2015.03.018.

Phuvipadawat, S., & Murata, T. (2010). Breaking news detection and tracking in twitter. *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (WI-IAT)* (pp. 120–123). Toronto, Canada: IEEE. http://dx.doi.org/10.1109/WI-IAT.2010.205.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine, 6*(3), 21–45.

Porikli, F., & Li, X. (2004). Traffic congestion estimation using HMM models without vehicle tracking. *Intelligent vehicles symposium, 2004 IEEE* (pp. 188–193). Parma, Italy: IEEE.

Rashidi, T. H., Abbasi, A., Maghrebi, M., Hasan, S., & Waller, T. S. (2017). Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C: Emerging Technologies, 75*, 197–211. http://dx.doi.org/10.1016/j.trc.2016.12.008.

Ribeiro, S. S., Jr., Davis, C. A., Jr., Oliveira, D. R. R., Meira, W., Jr., Gonçalves, T. S., & Pappa, G. L. (2012). Traffic observatory: A system to detect and locate traffic events and conditions using twitter. *Proceedings of the 5th ACM SIGSPATIAL international workshop on location-based social networks* (pp. 5–11). Redondo Beach, California: ACM. http://dx.doi.org/10.1145/2442796.2442800.

Sakaki, T., Matsuo, Y., Yanagihara, T., Chandrasiri, N. P., & Nawa, K. (2012). Real-time event extraction for driving information from social sensors. *2012 IEEE international conference on cyber technology in automation, control, and intelligent systems (CYBER)* (pp. 221–226). Bangkok, Thailand: IEEE. http://dx.doi.org/10.1109/CYBER.2012.6392557.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. *Proceedings of the 19th international conference on World wide web* (pp. 851–860). Raleigh, North Carolina, USA: ACM.

Salas, A., Georgakis, P., Nwagboso, C., Ammari, A., & Petalas, I. (2017). Traffic event detection framework using social media. *2017 IEEE international conference on smart grid and smart cities (ICSGSC)* (pp. 303–307). Singapore, Singapore: IEEE.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., & Sperling, J. (2009). TwitterStand: News in tweets. *Proceedings of the 17th ACM SIGSPATIAL international conference on advances in geographic information systems* (pp. 42–51). Seattle, Washington: ACM.

Schulz, A., & Ristoski, P. (2013). The car that hit the burning house: Understanding small scale incident related information in microblogs. *AAAI Technical Report/WS, 13–4*, 11–14. Retrieved from http://ub-madoc.bib.uni-mannheim.de/35450/1/The_Car_That_Hit_The_Burning_House_Understanding_Small_Scale_Incident_Related_Information_in_Microblogs.pdf.

Schulz, A., Ristoski, P., & Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. *Extended semantic web conference* (pp. 22–33). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-642-41242-4_3.

Semwal, D., Patil, S., Galhotra, S., Arora, A., & Unny, N. (2015). STAR: Real-time spatio-temporal analysis and prediction of traffic insights using social media. *Proceedings of the 2nd IKDD conference on data sciences* (pp. 7). Bangalore, India: ACM. http://dx.doi.org/10.1145/2778865.2778872.

Tejaswin, P., Kumar, R., & Gupta, S. (2015). Tweeting traffic: Analyzing twitter for generating real-time city traffic insights and predictions. *Proceedings of the 2nd IKDD conference on data sciences* (pp. 9). Bangalore, India: ACM. http://dx.doi.org/10.1145/2778865.2778874.

Terroso-Sáenz, F., Valdés-Vela, M., Sotomayor-Martínez, C., Toledo-Moreo, R., & Gómez-Skarmeta, A. F. (2012). A cooperative approach to traffic congestion detection with complex event processing and VANET. *IEEE Transactions on Intelligent Transportation Systems, 13*(2), 914–929.

Veeraraghavan, H., Schrater, P., & Papanikolopoulos, N. (2005). Switching kalman filter-based approach for tracking and event detection at traffic intersections. *Proceedings of the 2005 IEEE international symposium on intelligent control, 2005. Mediterrean conference on control and automation* (pp. 1167–1172). Limassol, Cyprus: IEEE.

Wang, S., He, L., Stenneth, L., Yu, P. S., & Li, Z. (2015). Citywide traffic congestion estimation with social media. *Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems* (pp. 34). Seattle, Washington: ACM.

Wang, X., Gerber, M. S., & Brown, D. E. (2012). Automatic crime prediction using events extracted from twitter posts. *5th international conference on social computing, behavioral-cultural modeling, and prediction (SBP)* (pp. 231–238). College Park, MD, USA: Springer International Publishing.

Wanichayapong, N., Pruthipunyaskul, W., Pattara-Atikom, W., & Chaovalit, P. (2011). Social-based traffic information extraction and classification. *2011 11th international conference on ITS telecommunications (ITST)* (pp. 107–112). St. Petersburg, Russia: IEEE. http://dx.doi.org/10.1109/ITST.2011.6060036.

Watanabe, K., Ochi, M., Okabe, M., & Onai, R. (2011). Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. *Proceedings of the 20th ACM international conference on information and knowledge management* (pp. 2541–2544). Glasgow, Scotland, UK: ACM. http://dx.doi.org/10.1145/2063576.2064014.

Weiler, A., Grossniklaus, M., & Scholl, M. H. (2015). Evaluation measures for event detection techniques on twitter data streams. In S. Maneth (Vol. Ed.), *Data science.*

BICOD 2015. Lecture notes in computer science. vol. 9147. Data science. BICOD 2015. Lecture notes in computer science (pp. 108–119). Edinburgh, UK: Springer, Cham. http://dx.doi.org/10.1007/978-3-319-20424-6.

White, J., Thompson, C., Turner, H., Dougherty, B., & Schmidt, D. C. (2011). WreckWatch: Automatic traffic accident detection and notification with smartphones. *Mobile Networks and Applications, 16*(3), 285–303. http://dx.doi.org/10.1007/s11036-011-0304-8.

Yazici, M. A., Mudigonda, S., & Kamga, C. (2017). *Incident detection through Twitter organization vs. personal accounts (no. 17-03884)*.

Zandbergen, P. A. (2009). Accuracy of iPhone locations: A Comparison of assisted GPS, WiFi and cellular positioning. *Transactions in GIS, 13*(s1), 5–25. http://dx.doi.org/10.1111/j.1467-9671.2009.01152.x.

Zhang, S., Tang, J., Wang, H., & Wang, Y. (2015). Enhancing traffic incident detection by using spatial point pattern analysis on social media. *Transportation Research Record: Journal of the Transportation Research Board, 2528*, 69–77. http://dx.doi.org/10.3141/2528-08.

Zhang, Z., Ni, M., He, Q., Gao, J., Gou, J., & Li, X. (2016). An exploratory study on the correlation between Twitter concentration and traffic surge. *Transportation Research Record, 35*, 36 Retrieved from the possibility of detect traffic from Twitter.

Zhao, L., Chen, F., Lu, C.-T., & Ramakrishnan, N. (2015). *Spatiotemporal event forecasting in social media.* http://dx.doi.org/10.1137/1.9781611974010.108.

Zheng, X., Chen, W., Wang, P., Shen, D., Chen, S., Wang, X., ... Yang, L. (2016). Big data for social transportation. *IEEE Transactions on Intelligent Transportation Systems, 17*(3), 620–630. http://dx.doi.org/10.1109/TITS.2015.2480157.