TRAFFIC EVENT DETECTION USING TWITTER DATA BASED ON ASSOCIATION RULES

Shishuo Xu [1, 2], Songnian Li [1, *], Richard Wen [1], Wei Huang [3]

[1] Department of Civil Engineering, Ryerson University, 350 Victoria St., Toronto, ON M5B 2K3, Canada – (shishuo.xu, snli, rwen)@ryerson.ca
[2] School of Environment Science and Spatial Informatics, China University of Mining and Technology, No. 1 Daxue Road, Xuzhou, Jiangsu 221116, China
[3] Ministry of Transportation, Ontario, 777 Bay St., Toronto, ON M5B 2H7, Canada – (huangweibuct@gmail.com)

KEY WORDS: Traffic event, Event detection, Geosocial media data, Twitter, Natural language processing, Association rules

ABSTRACT:

Social media platforms allow millions of people worldwide to instantly share their thoughts online. Many people use social media to share traffic related experiences and events with online posts. A large amount of traffic related data can be obtained from these online posts – especially geosocial media data, where posts are tagged with geolocation information such as coordinates or place names. By extracting traffic events from geosocial media data, drivers can adapt to changing traffic conditions, while traffic management departments can propose timely and effective plans to improve traffic conditions. Most of the existing studies query traffic-related information based on a list of single keywords, which result in large amounts of noisy data – negative data containing one or more traffic-related keywords, but do not actually represent real-world traffic events. This paper aims to filter noisy data by mining association rules among words in positive data containing messages representing traffic events. Messages are more likely to be true traffic events if they follow the co-occurrence pattern of words mined from positive samples. A case study was conducted in Toronto, Canada using Twitter data. The tweets queried by the association rules were classified into non-traffic event, traffic accidents, roadwork, severe weather conditions, and special events with an 85% accuracy based on supervised machine learning methods. Compared with hourly average travel speed data, 81% of detected events were identified as real-world traffic events. This research sheds light on traffic condition monitoring in smart transportation platforms, which plays an important role for smart cities.

1. INTRODUCTION

Traffic events (such as roadworks, traffic accidents, bad weather conditions, and special events) are likely to result in non-recurrent traffic congestion (Gutiérrez et al., 2015). Detecting traffic events in a real-time manner can help traffic authorities and drivers make responsive plans to improve traffic conditions (Fu et al., 2015a, 2015b; Gu et al., 2016). Geosocial media platforms provide a free technology for users to express their experiences through messages tagged with timestamps and geolocations. The large spatiotemporal coverage and widespread use of social media platforms make geosocial media data a potential source to extract useful information for traffic event detection (Kaplan and Haenlein, 2010). This approach shows advantages over traditional methods that suffer from high financial costs and limited spatiotemporal coverage (e.g., physical sensors installed along major roads), and other crowdsourced data that have open data sharing challenges (e.g., traffic data collected by specific mobile applications belonging to private companies).

Twitter is one of the most frequently used geosocial media platforms. A twitter message can be posted with a limited number of words, a timestamp, and a pair of coordinates (i.e., longitude and latitude) if location services are turned on. Normally, around 1% of posted tweets are geotagged with GPS coordinates (Morstatter et al., 2013). It is free to collect public tweets through open Application Programming Interfaces (APIs) including Representational State Transfer (REST) APIs and Streaming APIs (Twitter Inc., 2018). Recently, traffic related topics have brought more and more attention to our daily lives. People have a tendency to post traffic related information on social media platforms whenever they see car accidents, road constructions, or road closures. Mai and Hranac (2013) illustrated that compared to real-world traffic events, the relevant tweets were usually posted within 5 hours and between 10 to 25 miles of the actual location of traffic events. In Seattle downtown area, it was revealed that most traffic related tweets were located within 800 meters around the actual traffic event (Zhang et al., 2015). These studies demonstrate the capability of Twitter data in detecting traffic events.

In recent years, a number of studies concerning the extraction of traffic events from Twitter data have been conducted. Liu et al. (2014) detected traffic events by combining spatiotemporal analysis models with wavelet analysis models. Semwal et al. (2015) trained a classifier to predict traffic anomalies for the next day by mining the relationship between co-occurrence of certain problems and causes. Most of the current studies follow a general process of filtering traffic related tweets through keywords or specific accounts, preprocessing them with natural language processing techniques, identifying traffic events using classification methods, and geocoding them to real-world locations (Nguyen et al., 2016; Ribeiro Jr. et al., 2012; Li et al., 2012).

Noise exists in initially queried tweets when a single keyword such as “crash” or “street” is separately used as queries. Tweets queried by single selected keywords usually contain much more tweets that are negatively related to real-world traffic event (negative tweets) than tweets that are positively related to real-world traffic event (positive tweets). In other words, a large amount of tweets may not refer to a real-world traffic event even though they include one or more traffic related keywords. For example, the following two tweets: “Quality is never an *accident*. It is always the result of intelligent effort.” and “The accident

https://doi.org/10.5194/isprs-annals-IV-2-W5-543-2019 | © Authors 2019. CC BY 4.0 License.

occurred around 9:45pm at Nkawkaw, when our bus ran into a stationary vehicle." can be queried by the traffic related keyword "accident". The second tweet refers to a traffic accident, while "accident" means coincidence in the first tweet, which is a negative tweet. The imbalance between the number of positive tweets and positive tweets negatively influences the performance of text classification processes. Cui et al. (2014) proposed an n-gram based approach to solve the problem. This method requires predetermining the parameter n in n-gram, which limits processing to a fixed number n of words (grams) in sequence.

This research solves these issues by mining the co-occurring patterns (association rules) between words in positive tweets, and then constructing Twitter queries using a combination of a few words (called a "wordset") rather than single words to extract positive tweets, and to discard negative tweets as much as possible. The remainder of the paper is organized as follows: Section 2 introduces the overall workflow and main technology adopted in each step, Section 3 presents the result of a case study conducted in Toronto, Canada, and Section 4 outlines the conclusions and future work.

2. METHODOLOGY

The general workflow of detecting traffic event using geosocial media data is summarized in Figure 1, where a number of stages and techniques applied in each stage are illustrated. Twitter provides several APIs to the public for obtaining free data. Streaming APIs and REST APIs are mostly used for research purposes, where a list of preselected keywords or a geospatial bounding box is used to crawl raw tweets in JSON format. In this study, a geospatial bounding box was applied to collect Twitter data through Streaming APIs. Tweets geotagged with GPS coordinates or a place (e.g. Eaton Center, Toronto, Canada) falling in the area defined by the bounding box were included. Only tweets tagged with GPS coordinates were used in this study.

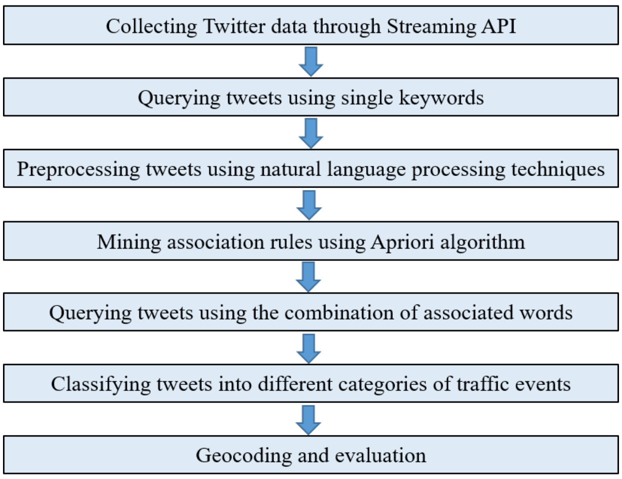

Figure 1. The overall workflow of detecting traffic events using geosocial media data

2.1 Preprocessing Twitter data

Traffic related tweets were queried from raw tweets using the keyword-based method. In this study, 59 words and phrases that appeared twice or more in the studies reviewed by Xu et al. (2018) were selected as keywords for the initial query. The top 10 keywords and their frequency counted in reviewed papers are summarized in Table 1.

Table 1. Top 10 traffic related keywords

Keywords	Frequency	Keywords	Frequency
Accident	10	Street	7
Traffic	9	Congestion	6
Crash	9	Delay	6
Road	8	Incident	5
Blocked	7	Closed	5

The initial queried tweets were further processed based on Natural Language Processing (NLP) techniques. Stanford NLP tools were used to tokenize, lowercase, and lemmatize words as well as to remove stop words. An example showing the detailed procedure is presented in Figure 2.

2.2 Mining association rules

The association rules between words in positive tweets were mined by Apriori algorithm (Rakesh and Ramakrishnan, 1994). A new wordset containing the association rules was composed for further queries of positive tweets. For instance, there exist an association rule among words "*crash*", "*street*" and "*left*", which are joined together for a complicated query, namely "*crash*" and "*street*" and "*left*" rather than "*crash*" or "*street*" or "*left*".

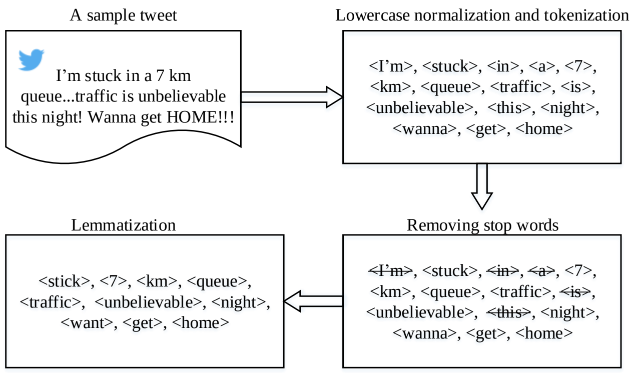

Figure 2. An example of preprocessing tweets

The Apriori algorithm is widely used in the field of transaction mining to mine frequent itemsets for establishing Boolean association rules. The frequent itemsets refer to sets of items that have minimum support, which can be iteratively found with cardinality from 1 to k – namely ranging from 1-itemset frequent patterns to k-itemset frequent patterns. Association rules are generated based on the identified frequent itemsets. Two parameters, namely support and confidence, are required to be estimated to perform this process. In this study, each word token (w) was viewed as an item, and the whole itemset (W) was the tweets collection containing all word tokens. The support of the i^{th} word token (w_i) refers to the proportion of tweets containing w_i ($T(w_i|w_i \in W)$) in all tweets $T(W)$, which can be calculated as follows:

$$Support(w_i) = \frac{T(w_i|w_i \in W)}{T(W)} \tag{1}$$

Similarly, the support of n associated words ($w_1 \cap w_2 \cap \ldots \cap w_n$) can be calculated as:

$$Support(w_i) = \frac{T(w_1 \cap w_2 \cap \ldots \cap w_n | w_1, w_2, \ldots, w_n \in W)}{T(W)} \tag{2}$$

The confidence refers the likelihood that word token i (w_i) also appears when word token j (w_j) appears. It can be calculated as the number of tweets containing w_i and w_j ($T(w_i \cap w_j | w_i, w_j \in W)$) divided by the number of tweets containing w_i ($T(w_i | w_i \in W)$):

$$Confidence\ (w_i \rightarrow w_j) = \frac{T(w_i \cap w_j | w_i, w_j \in W)}{T(w_i | w_i \in W)} \quad (3)$$

2.3 Text classification

As mentioned in Section 1, traffic events can be potential causes for non-recurrent traffic congestion. In this study, we further classified the queried tweets into five categories including non-traffic, traffic accidents, roadwork, severe weather conditions, and special events based on supervised machine learning classification methods. The Term frequency-inverse document frequency (tf-idf) method (Hand, 2006) was adopted to select features for the multi-class text classification. Accordingly, Naïve Bayes and Logistic Regression methods were both tested to train classifiers.

A tweet is represented as a feature vector (x_1, x_2, ..., x_n). With respect to Naïve Bayes method (Maron, 1961), the probability of a tweet belonging to class y is calculated as:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)P(x_1, x_2, \dots, x_n|y)}{P(x_1, x_2, \dots, x_n)} \quad (4)$$

By adopting the naïve conditional independence assumption that:

$$P(x_i|y, x_1, x_2, \dots, x_i, x_{i+1}, \dots, x_n) = P(x_i|y) \quad (5)$$

Equation (4) can be simplified as:

$$P(y|x_1, x_2, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, x_2, \dots, x_n)} \quad (6)$$

where $P(x_i|y)$ refers to the probability of feature i (x_i) appearing in a tweet belonging to class y. It can be estimated as:

$$P(x_i|y) = \frac{\sum_{x \in T} x_i + \alpha}{\sum_{i=1}^{n} \sum_{x \in T} x_i + \alpha n} \quad (7)$$

where $\sum_{x \in T} x_i$ refers to the number of times that feature i (x_i) appears in a tweet of class y in the training set T, α refers to the smoothing priors, and $\sum_{i=1}^{n} \sum_{x \in T} x_i$ refers to the total count of all features for class y.

The Logistic Regression method (Walker and Duncan, 1967) estimates the probability of a tweet belonging to a class y using a logistic/sigmoid function. It can be calculated as follows:

$$P(y|x_1, x_2, \dots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad (8)$$

where $\beta_0, \beta_1, \dots, \beta_n$ are coefficients parameters for the model.

3. CASE STUDY

Twitter data posted in Toronto, Ontario, Canada was collected for a case study (Figure 3) based on a geo-bounding box through Twitter Streaming APIs. As shown in Figure 3, Toronto is located in south Ontario on the northern shore of Lake Ontario, and more tweets are located in downtown areas than in suburbs. As a result, a total of 17,170,543 tweets tagged with GPS coordinates or places were obtained from April 1, 2014 to March 31, 2015. A total of 4,413,821 tweets tagged with GPS coordinates, namely around 25.7% of the obtained tweets, were used for analysis in this work.

Based on the 59 traffic related keywords, 160,747 tweets were initially queried. Considering both severe weather conditions (e.g., snowstorm) and special events (e.g., Christmas parade) are more likely to happen during winter, we manually labelled positive tweets posted in November, 2014 as training data to mine the association rules among words. With regard to adopting Apriori algorithm, the support and confidence were empirically set as 0.01 and 0.1, respectively (Zhang et al., 2018). A total of 53 association rules were extracted to make up the wordset, ten of which are listed in Table 2 as an example. By further querying based on the wordset, a total of 3,594 tweets were left for the following categorical classification.

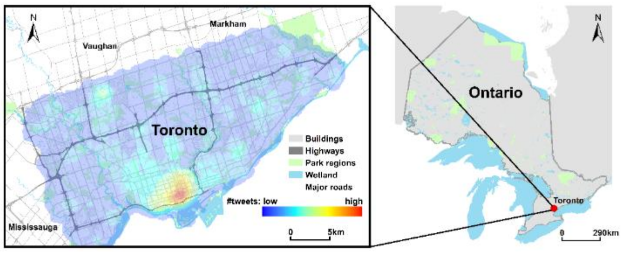

Figure 3. The geography of Toronto, Ontario, Canada

Table 2. A sample of association rules

close & highway	serious & injury
collision & condition	strike & vehicle
involve & crash	injure & crash
crash & hwy	snow & drive
safe & drive	northbound & close

The tools from the Python library scikit-learn (Pedregosa et al., 2011) was adopted to conduct machine learning classification using the Naïve Bayes and Logistic Regression methods. We manually label all of the queried tweets, namely above mentioned 3,594 tweets, where 80% were randomly selected as training data to train the classifiers, and the other 20% were used for prediction according to Pedregosa et al. (2011). Based on the training data, the performance of the two classifiers were evaluated and the results are presented in Table 3. The classifier generated by the Logistic Regression method outperforms the classifier generated by the Naïve Bayes method. Therefore, we used the Logistic Regression classifier to predict the 20% of the queried tweets, which resulted in an overall accuracy of 0.85, as well as precision of 0.85, recall of 0.83, and F_1-score of 0.84 by comparing using the manual labels. Four types of traffic events (traffic accidents, roadwork, severe weather conditions, and special events) were then geocoded to real world locations by referring to their GPS coordinates.

Table 3. Performance of two classifiers

Method	Accuracy	Precision	Recall	F_1-score
Naïve Bayes	0.83	0.76	0.83	0.78
Logistic Regression	0.85	0.83	0.85	0.83

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-2-W5-543-2019

The traffic event detection results were validated with the travel speed data of trucks collected in the Toronto area during the year of 2014 and 2015, which were provided by the Ontario Ministry of Transportation, Canada. This dataset includes hourly average travel speed of each road link for each day. Considering there may be certain location deviations between the geotagged coordinates of real-world traffic event and road links, the geocoded events were further located to the nearest road link using the "Near" tool in ArcGIS. A Z-test was performed for evaluation with the null hypothesis that the actual travel speed (s_i), when a traffic event i occurs, follows the Gaussian distribution of the typical travel speed $(T_1, T_2, T_3, \dots T_i, \dots, T_n)$ without traffic anomalies, given the same road link, hour, and day of the week. With the standardized travel speed described by Equations (9) and (10), the P-value of traffic event i was generated by Equation (11). Finally, a significance level L was required to reject or accept the null hypothesis.

$$s_i' = \frac{s_i - Mean(T_1, T_2, T_3, \dots T_i, \dots, T_n)}{Standard\ deviation(T_1, T_2, T_3, \dots T_i, \dots, T_n)} \quad (9)$$

$$T_i' = \frac{T_i - Mean(T_1, T_2, T_3, \dots T_i, \dots, T_n)}{Standard\ deviation(T_1, T_2, T_3, \dots T_i, \dots, T_n)} \quad (10)$$

$$P_i = Probability \begin{cases} P_I \geq L, & accept\ the\ null\ hypothesis \\ P_I < L, & reject\ the\ null\ hypothesis \end{cases} \quad (11)$$

In this work, accepting the null hypothesis indicated that the detected traffic events refer to real-world traffic events. As shown in Figure 4, a sensitivity analysis was done to evaluate the effects of the significance level on event detection accuracy. When the significance level was defined as 0.1, 81% of the detected traffic events are likely to be real-world traffic events, while at least 15% of events were actually detected when the significance level was set to be 0.9.

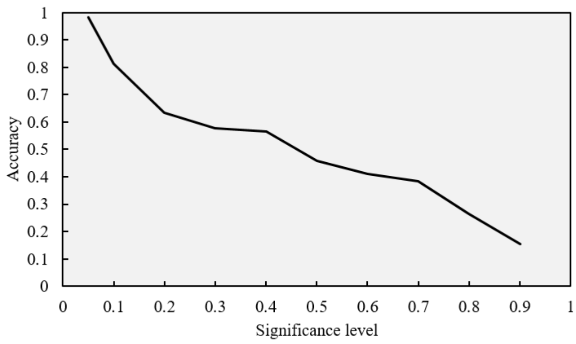

Figure 4. The relationship between significance level and traffic event detection accuracy

4 CONCLUSIONS

Detecting traffic events in a real-time manner can help drivers and traffic managers take effective measures to alleviate traffic congestion. Geosocial media platforms provide cost-effective and large sources of traffic event data by allowing people to report traffic anomalies at any time and at any location with internet. A case study using Twitter data was conducted in Toronto, Canada. We mined the association rules between words in positive tweets for a secondary query to further filter the noisy tweets introduced by the initial single-keyword query from raw Twitter streams. The Logistic Regression method was then selected to classify the queried tweets into non-traffic events, traffic accidents, roadwork, severe weather conditions, and special events, which was able to achieve an accuracy of 85%. By validating with the hourly average travel speed data through a Z-test, 81% of the detected events can be identified as actual real-world traffic events when the significance level was set as 0.1. In addition, this work can help control traffic flow and improve road safety – resulting in the reduction of air pollution severity caused by traffic congestion, and the improvement in the quality of life for citizens. This work also promotes the development of smart transportation platforms and smart cities.

In the future, we will take advantage of multiple sources of geosocial media data to extract as many real-world traffic event as possible. People have different geosocial media platform preferences, and taking advantage of multiple sources will account for these user preferences, where traffic related information absent in Twitter may be possibly filled by other platforms. Cross validation methods will be applied to automatically determine the optimal values for support and confidence in the Apriori algorithm. Considering that users may be far away from the exact location of traffic events when they post, the location information present in the text of traffic related tweets (e.g., street name) can be further extracted for location inference.

ACKNOWLEGEMENTS

This work has been funded by the Natural Science and Engineering Research Council of Canada (NSERC).

REFERENCES

Cui, J., Fu, R., Dong, C., Zhang, Z., 2014. Extraction of traffic information from social media interactions: Methods and experiments. 17th Int. IEEE Conf. Intell. Transp. Syst. 1549–1554. https://doi.org/10.1109/ITSC.2014.6957913

Fu, K., Lu, C.-T., Nune, R., Tao, J.X., 2015a. Steds: Social media based transportation event detection with text summarization, in: 2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC). IEEE, Las Palmas, Spain, pp. 1952–1957. https://doi.org/10.1109/ITSC.2015.316

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

Gu, Y., Qian, Z. (Sean), Chen, F., 2016. From Twitter to detector: Real-time traffic incident detection using social media data. Transp. Res. part C Emerg. Technol. 67, 321–342. https://doi.org/10.1016/j.trc.2016.02.011

Gutiérrez, C., Figuerias, P., Oliveira, P., Costa, R., Jardim-Goncalves, R., 2015. Twitter mining for traffic events detection, in: 2015 Science and Information Conference. IEEE, London, UK, pp. 371–378. https://doi.org/10.1109/SAI.2015.7237170

Hand, D.J., 2006. Data mining. Data Mining. Encycl. Environmetrics 2. https://doi.org/10.1007/978-3-319-67008-9_20

Kaplan, A.M., Haenlein, M., 2010. Users of the world, unite! The challenges and opportunities of Social Media. Bus. Horiz. 53, 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Li, R., Lei, K.H., Khadiwala, R., Chang, K.C.C., 2012. TEDAS: A twitter-based event detection and analysis system, in: 2012

IEEE 28th International Conference on Data Engineering (ICDE). IEEE, Washington, DC, USA, pp. 1273–1276. https://doi.org/10.1109/ICDE.2012.125

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, in: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM, Dallas, Texas, pp. 549–552. https://doi.org/10.1145/2666310.2666366

Mai, E., Hranac, R., 2013. Twitter Interactions as a Data Source for Transportation Incidents. Transp. Res. Board 92nd Annu. … 078, 1–11.

Maron, M.E., 1961. Automatic indexing: an experimental inquiry. J. ACM 8, 404–417. https://doi.org/10.1145/321075.321084

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, in: ICWSM. pp. 400–408. https://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, in: PAKDD 2016: Advances in Knowledge Discovery and Data Mining. Springer, Cham, pp. 540–551. https://doi.org/10.1007/978-3-319-31753-3

Pedregosa, F., Michel, V., Grisel OLIVIER, Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Rakesh, A., Ramakrishnan, S., 1994. Fast Algorithms for Mining Association Rules, in: Proc. 20th Int. Conf. Very Large Data Bases, VLDB. pp. 487–499. https://doi.org/10.1.1.20.1600

Ribeiro Jr., S.S., Davis Jr., C.A., Oliveira, D.R.R., Meira Jr., W., Gonçalves, T.S., Pappa, G.L., 2012. Traffic Observatory: A system to detect and locate traffic events and conditions using Twitter, in: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks. ACM, Redondo Beach, California, pp. 5–11. https://doi.org/10.1145/2442796.2442800

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 1–4.

Twitter Inc., 2018. Avaialble at URL: https://developer.twitter.com/en/docs.html (Last Accessed March 12, 2019)

Walker, S.H., Duncan, D.B., 1967. Estimation of the probability of an event as a function of several variables independent. Biometrika 54, 167–179.

Xu, S., Li, S., Wen, R., 2018. Sensing and detecting traffic events using geosocial media data: A review. Comput. Environ. Urban Syst. 72, 146–160. https://doi.org/10.1016/j.compenvurbsys.2018.06.006

Zhang, S., Tang, J., Wang, H., Wang, Y., 2015. Enhancing traffic incident detection by using spatial point pattern analysis on social media. Transp. Res. Rec. J. Transp. Res. Board 2528, 69–77. https://doi.org/10.3141/2528-08

Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transp. Res. Part C Emerg. Technol. 86, 580–596. https://doi.org/10.1016/j.trc.2017.11.027