Traffic Accident Detection with Both Traffic and Social Media Data

Zhenhua Zhang
Department of Civil, Structural and Environmental Engineering
The State University of New York, Buffalo, NY 14260
Email: zhenhuaz@buffalo.edu

Ming Ni
Department of Industrial and Systems Engineering
The State University of New York, Buffalo, NY 14260
Email: mingni@buffalo.edu

Qing He
Department of Civil, Structural and Environmental Engineering
Department of Industrial and Systems Engineering
The State University of New York, Buffalo, NY 14260
Email: qinghe@buffalo.edu

Jing Gao
Department of Computer Science and Engineering
The State University of New York, Buffalo, NY 14260
Email: jing@buffalo.edu

Jizhan Gou
Serco Inc., Reston VA 20190
Email: Jizhan.Gou@serco-na.com

Xiaoling Li
Northern Region Operations,
Virginia Department of Transportation, Fairfax, VA 22030
Email: Ling.li@vdot.virginia.gov

1 Corresponding Author
Abstract
Social media receives increasing attention as crowdsourced information for traffic operations and management. One recent trending study is to use social media to detect on-site traffic accidents. However, it remains unknown how effective the social media based detection methods is as compared with traditional loop detector based method. In this paper, we first explore the features of keywords and their association rules inherent in the accident-related tweets and explore the potentials of tweets in accident detection. Combining the traffic flow and occupancy data, our prediction results show that tweets can sometimes respond to traffic accident much more quickly than traditional methods and can even find some un-documented accidents which make up for the deficiencies of VDOT records. Also, the limitations and disadvantages are also discussed which provide insights in utilizing social media data to assist accurate on-site traffic accident detection.

Keyword: traffic accident detection; tweet; social media; association rules; traffic signature;

1 Introduction
Major traffic accidents, which cause irreparable damages, injuries, and even fatalities, are considered as one of the most important urban problems worldwide. National Highway Traffic Safety Administration (NHTSA), which publishes yearly reports on traffic safety facts, states that since 1988 more than 5,000,000 car crashes occur in the States each year and about 30% of them bring fatalities and injuries (NHTSA, 2015). After years of research, it has been widely accepted that significant reductions in accident impact can be achieved through effective detection methods and corresponding management strategies. Accurate monitoring of traffic and effective detection of traffic accidents are critical to modern transportation management.

Due to the fact that major traffic accidents potentially bring severe breakdown in the traffic flow, traditional attempts in traffic accident detection focus mainly on monitoring fluctuations and changes of one or more traffic-related metrics such as the traffic flow, occupancy, speed, etc. Traditional methods leverage the time-of-day characteristics and geographic features to identify the anomalies that may indicate a traffic accident. For example, Payne et al. (1978) used freeway traffic flow data for the detection of accidents and other lane blockage incidents that temporarily disrupt traffic flow; Samant et al. (2000) developed an effective traffic incident detection algorithm to extract incident-related features from traffic patterns. Amin et al. (2012) proposed to utilize the capability of a GPS receiver to monitor the speed of a vehicle and detect accident based on monitored speed. Jin et al. (2009) proposed an incident decision-making algorithm to detect traffic incidents on the basis of traffic flow-occupancy relationships.

Despite the adaptabilities of these studies, the improvement in the accuracy of detection with only traffic data still meets certain challenges. First, most of the previous research, which utilized the field data to detect the traffic accidents, build on the implicit assumption that the data is reliable. However, requiring real-time data from traffic
detectors is very expensive in maintenance and operations. Detector failures or data
errors are perennial problems in traffic operations. For example, Illinois Department of
Transportation (IDOT) in Chicago reported that no more than 5 percent of their loops
(detectors) are inoperative at any given time (Kell et al., 1990). The percentage is not low
enough as compared to the rate of a traffic accident. The problem of malfunctioned
sensors cause even more troubles in incident detection in large regions, say, an area with
more than 10,000 signalized intersections. Second, the uncertainty nature of traffic
patterns and non-recurrent social activities may undermine the potential of traffic metrics
in justifying the traffic accidents. Besides traffic accidents, daily traffic operations may
suffer breakdowns by other factors such as parades, road constructions, running races,
etc. Thus, the metrics including the traffic flow and occupancy inherently perform as an
indirect support for traffic accidents instead of a direct proof. To address these
challenges, there are efforts in applying clustering or classification methodologies such as
K-means (Münz et al., 2007) on large data collections to diminish the errors. Other
tendencies lie in incorporating more facts that relate to the real-time interaction of
accidents.

Different from data sources from loop detectors, Twitter, the microblogging service that
has received increasing attentions in recent years, has been gradually accepted as a user-
contributed data source in event detection. Compared to other information diffusion
channels, Twitter creates an online environment where content is created, consumed,
promoted, distributed, discovered or shared for purposes that are primarily related to
communities and social activities, rather than functional task-oriented objectives (Gal-
Tzur et al., 2014). Thus, it founds a perfect stage of “We Media” and this makes possible
the wide-range information retrieval from the broad masses of the people. Retrieving
useful information from Twitter has been studied widely in many applications including
detecting earthquake (Sakaki et al., 2010), bird flu (Aramaki et al., 2011), politic events
(Shirky, 2011), etc. These approaches focus on identifying the keywords pair that occur
disproportionally frequently at the current time (Giridhar et al., 2014). There are some
studies on the analysis of the correlation between tweets and traffic-related problems. For
example, Schulz et al. (2013) used microblogs to detect the small scale incidents; Gal-
Tzur et al. (2014) conducted a corridor study on the correlation between tweet and traffic
jam; Mai et al. (Mai and Hranac, 2013) compared incident records with Twitter messages
and argued the potentials for information from Twitter to add context to other traffic
measurements as a supplemental data source. Our preliminary examinations also have
demonstrated the potentials of tweets in traffic accident detection such as:

- “major accident next to the sunoco near the parkway a car got flipped over”
- “the worst car accident possible just happened in front of me”

However, the shortcomings of using tweets to detect traffic accidents are almost as
obvious as its merits. There are two major challenges to be addressed before the use of
tweets in traffic accident detection. First, as compared to events that arouse enormous
public concerns such as key basketball games, extreme weathers or traditional festivals,
the influence of traffic accidents are comparably a “midget”. From our observation,
tweets related to traffic accidents are thus in small quantity. What’s more, most of them are confined to a small area and limited to a relatively short time interval and some researchers call them small-scale events (Schulz et al., 2013). Second, the challenge in tweets lies in its inherent complexity and unstructured nature of data: language ambiguity (Chen et al., 2014). The common methods in detecting the traffic-related events include support vector machine (D’Andrea et al., 2015; Schulz et al., 2013), natural language processing (Li et al., 2012; Wanichayapong et al., 2011), etc. which explore the semantic features in the keywords. However, as the context of tweet is limited to 140 words and the tweet contents try to be concise, keyword detection is sometimes not sufficient for accurate automatic language processing. For example, “internet traffic is slow” and “internet shows traffic is slow” may deliver totally different information. Thus, the association rules in the tweet contents should be explored and implemented in the traffic accident detection.

To address above challenges, we employ a novel regression model to automatically detect the accident-related tweets. We investigate the features of both keywords and their association rules in the tweet contents and study the potentials of traffic data of flow and occupancy in increasing the predicting accuracy. There are two major contributions:
First, in addition to analyze the word or token separately, we reveal the association rules between words in each Twitter post and include the association features in our detection model; Second, explore a more accurate detection of on-site traffic accidents by combining the traffic-related metrics and tweet information. In principle, the fusion of multi-source data provides significant advantages over single source data (Hall and Llinas, 1997). In sum, the integrations of association features inherent in the tweet contents and other data sources are expected to produce more synthetic and informative results.

The paper is organized in the following steps: In Section 2, the study area and the raw data sources are introduced. In Section 3, the models and the selected features are detailed. In Section 4, we present the results of using different features and compare with the ground truth to reveal the pros and cons of accident detection based on tweets. In Section 5, we conclude this paper with a few empirical findings and generalizations together with some thoughtful discussions.

2 Data description
The study area, shown in Figure 1 (a), is located in the vast road network of Northern Virginia (NOVA). With 2.8 million residents (about a third of the state), NOVA is the most populous region of Virginia and the Washington D.C. Metropolitan Area. It has long been known for its heavy traffic (Cervero, 1994). The road network is a 50 square-kilometer (31 miles) area with more than 1,200 signalized intersections. In our study, we mainly include three categories of data:
The tweet data were collected through Twitter Streaming API with geo-location filter. Filtering by the coordinates, we extracted tweets posted only from NOVA region. There are more than 584,000 tweets from January 2014 to December 2014.. Each tweet posts are coupled with specific date, time and location information. The tweets are the
reflection of what people are interested at the specific time and location. Thus, they can justify the traffic accident if the text content has a clear expression of it. The location information is the paired latitude and longitude where the tweets are posted. The resolution of the location can be as high as 100 meters. Automatic extraction of accident-related tweets can be of great use in traffic management. The effectiveness of the detection is the major topic in our study.

The traffic data are collected by loop detectors equipped at the approach of the intersection. The detectors amount to nearly 15,000 in NOVA. These loop detectors keep recording the traffic flow and occupancy at an interval of 15 minutes. With these traffic detectors, the access to real-time traffic information in our study area is becoming routine as under growing pressure for improving traffic management (Leduc, 2008).

**Figure 1** (a) Geographic districts of the study area, (b) road network map of one sample region and (c) locations of the detectors on approaches of the intersections
To test the validity of our results, we can refer to the traffic management log maintained by Virginia Department of Transportation (VDOT). The traffic management log is an accident database recording the historical accidents in NOVA in the past few years. The accidents include “collision”, “disabled vehicle”, “vehicle on fire” etc. There are about 52,496 accidents happen in our study area throughout Year 2014. Each accident database is paired with detailed information of latitude, longitude, date, time and corresponding incident description. Such data can be the verification of the accident-related tweets and are taken as the ground truth in our classification model.

3 Model and feature selection

3.1 Regression model

We employ the support vector machines (SVMs) as our classification model. SVMs, first introduced in (Cortes and Vapnik, 1995), is a supervised learning method used for regression analysis. In our study, we first train the model by the labeled tweets and then employ the model to automatic classify the tweets whether they are related to traffic accidents.

In training the model, we first give manual labels for the tweets and this manual label is used as the expected output of the model. Then we select the proper features from both the tweet contents and corresponding traffic-related data and these features are the covariates for the model. In training the regression model, we further implement 5-fold cross validation (Geisser, 1993) to increase the accuracy of the predicted model. Cross-validation can give insight on how the model will generalize to an independent dataset. Directed by this method, the dataset is randomly partitioned into 5 folds. The classification model is trained on 4 folds, and the remaining fold is used for testing the trained model. This procedure is repeated 5 times and each fold is used exactly once as a test data. We finally obtained an overall estimation by averaging 5 test results.

Besides the manual labelling of tweets, the feature selection is detailed in the following sections.

3.2 Feature extraction from tweet information

3.2.1 Tweet filtering and labelling

Tweet filtering fetches the candidate tweets that can be used to train the accident detection model. The candidate tweets should be those that possibly describe the on-site traffic accidents. We can assume that people describe traffic accidents by accident-related words and these words can be decided their semantics as well as their semantics. We turn to the web news that broadcast the traffic accident and several words are summarized in Table 1.

<table>
<thead>
<tr>
<th>Table 1 Accident-related words</th>
</tr>
</thead>
</table>

We apply the filter based on keywords to get the accident-related tweets. To ensure both the accuracy and sample size, certain rules must be followed:
- Include the words that are relevant to accidents but apparently misspelled or personally modified including “acident”, “incident”, etc.
- Include other variations of accident-related words such as the word pairs that have a hyphen in word pair such as roll-over, etc.
- Exclude the words related to transportation authority or news media.

Finally, we obtained more than 3500 eligible tweets. According to the label \( L \), more than 400 tweets are taken as accident related. For these accident-related tweets, we randomly pick tweets of the same size that are not related to accidents. We combined the tweets that related to accident and selected those that are unrelated. These tweets symbolizes as \( T \) and \( i \)th tweet is \( T_i \). They will be used for training our regression model.

### 3.2.2 Token filtering and stemming

To fetch the proper features, each tweet is further decomposed into several tokens including words, characters, numbers or even Latinized symbols. There are more than 10000 tokens from all eligible tweets. Each token can potentially convey the instantaneous ideas and feelings of the tweeters and they will be selected as the features of the regression models after necessary filtering and stemming. The steps of the process can be illustrated as Figure 2.

#### Figure 2 Steps of token filtering and stemming

First, the punctuation marks convey almost no meanings and should be discarded and all other words should be converted into lower case. Meanwhile, some of the words or characters that have no apparent linguistic meanings or significant event indications should be filtered out before the processing. These words are referred as stop-words. Stop-word filtering is a prevailing method in page analyzer and article analyzer in preprocessing of natural language (Rajaraman et al., 2012). The stop-word list we used refer to (Ranks-NL, 2015).

Second, some of the words have different writing expressions due to the grammatical reasons but convey almost the same meanings such as “accidents” and “accident”. The
token stemming is necessary to reduce these inflected (or sometimes derived) words to their word stem, base or root form. In this study, we employ the Porter stemming algorithm (Porter, 1980) for the token stemming and each token is grouped into the proper stemmed token.

After token filtering and stemming, each tweet \( T_i \) can be summarized several stemmed tokens. Of all the tweets \( T \), there are more than 3000 stemmed tokens symbolized as \( \{ t_1, t_2, ... t_j \} \). The stemmed tokens are the features for each tweet \( T_i \) and each tweet has different token features. If the tweet contains a stemmed token, the corresponding token features are labeled as 1 otherwise 0. Thus, the token features and the tweets \( T \) form our binary database \( D_S \) and it will be used for the feature selection.

### 3.2.3 Feature selection from each token

This section describes the steps of selecting features from each token in the feature-selection database \( D_S \). As the stemmed tokens are in large amounts and only a few of them can be related to the traffic accident, we proceed to filter out the proper features from these tokens. In this section, we focus on correlation between each token and our manual label. It is a prevailing method of employing selected tokens as the features in studies such as (Abel et al., 2012; D'Andrea et al., 2015).

The correlation benchmark we choose is phi coefficient (Cramér, 1999), which is widely accepted as a measure of association between two binary variables. The coefficient (usually denoted as \( \phi \)) between two variables \( x \) and \( y \) is calculated as:

\[
\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1}n_{01}n_{00}n_{10}}}
\]

Where all notations are defined in the following table:

<table>
<thead>
<tr>
<th></th>
<th>( y = 1 )</th>
<th>( y = 0 )</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 1 )</td>
<td>( n_{11} )</td>
<td>( n_{10} )</td>
<td>( n_{1} )</td>
</tr>
<tr>
<td>( x = 0 )</td>
<td>( n_{01} )</td>
<td>( n_{00} )</td>
<td>( n_{0} )</td>
</tr>
<tr>
<td>Total</td>
<td>( n_{1} )</td>
<td>( n_{0} )</td>
<td>( n )</td>
</tr>
</tbody>
</table>

Those tokens whose \( |\phi| \) is higher than 0.1 are selected. Following this rule, 27 tokens are selected and five of them are shown in Table 2.

**Table 2** some of the tokens and their correlations with accident label

<table>
<thead>
<tr>
<th>Features</th>
<th>Correlation</th>
<th>Features</th>
<th>Correlation</th>
<th>Features</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>accident</td>
<td>0.371</td>
<td>66</td>
<td>0.186</td>
<td>mile</td>
<td>0.143</td>
</tr>
<tr>
<td>car</td>
<td>0.282</td>
<td>damage</td>
<td>-0.169</td>
<td>loop</td>
<td>0.143</td>
</tr>
<tr>
<td>lane</td>
<td>0.252</td>
<td>95</td>
<td>0.159</td>
<td>vehicle</td>
<td>0.143</td>
</tr>
<tr>
<td>traffic</td>
<td>0.242</td>
<td>rd</td>
<td>0.157</td>
<td>accidently</td>
<td>-0.142</td>
</tr>
</tbody>
</table>
From Table 2, some of the tokens may be accounted by the geographic uniqueness such as “66”, “95”, and “495” which indicates the route number, and this means tweeter inclined to report the traffic accidents with route name; some may direct to the relevant topics “traffic”, “accident”, etc.; other words such as “damage” or “accidently” are too general in our daily lives and thus lose the uniqueness in describing the traffic accident.

3.2.4 Feature selection from paired tokens

In Section 3.3, we select the features from tokens that contain only one word. However, these may overlook the interconnections between tokens and sometimes the associations between words can have much more significant indications than single ones. For example, in a tweet post, the occurrence of token “car” conditioned by “accident” may increase the accident-related probability. Conversely, the occurrence of token “car” conditioned by “maintenance” or “repair” may undermine the potentials of indicating a traffic accident. Potentially, the co-occurrence of certain word pairs in a tweet may indicate the existence of traffic accident.

In this section, we select the features from paired tokens by study the association rules between the manual label and the stemmed tokens in the binary database $D_S$. The association rules can be unveiled by the Apriori algorithm (Agrawal and Srikant, 1994).

Apriori algorithm can find the regularities in large-scale binary data by two major probabilities: support and confidence.

Support calculates the proportion of tweets that contain certain tokens (the token feature equal to 1) in the database:

$$supp(t_j) = \frac{sizef([{T_i}, t_j \subseteq T_i])}{sizef([{T_i}])}$$

(2)

Given a threshold $supp(t_j)$, there will be a limited number of qualified $t_j$. We can even calculate the support of more than one tokens:

$$supp(t_{j_1} \cap t_{j_2} \cap \ldots \ldots \cap t_{j_m}) = \frac{sizef([{T_i}, t_{j_1} \cap t_{j_2} \cap \ldots \ldots \cap t_{j_m} \subseteq T_i})}{sizef([{T_i}])}$$

(3)

Where $j_1, j_2, \ldots, j_m$ are random numbers that are not correlated to each other. Noting that support deals mainly with the frequencies of paired tokens, we still need the confidence which calculates proportions of tweets with manual label $L_i$ which also contains a set of tokens: $t_{j_1} \cap t_{j_2} \cap \ldots \ldots \cap t_{j_m}$:

$$conf(L_i \Rightarrow t_{j_1} \cap t_{j_2} \cap \ldots \ldots \cap t_{j_m}) = \frac{supp(L_i \cap t_{j_1} \cap t_{j_2} \cap \ldots \ldots \cap t_{j_m})}{supp(t_{j_1} \cap t_{j_2} \cap \ldots \ldots \cap t_{j_m})}$$

(4)

In the confidence calculation, the binary manual label $L_i$ can be equal to 1. This means we focus more on the paired tokens that related to traffic accident. The size of the paired tokens is theoretically equal to the total counts of tokens, but due to the limited size of the
tweet posts, larger values will be of no use. By setting the support equal to 0.1 and confidence equal to 0.5, our results show that most paired tokens contain “accident”.

### Table 3 Paired tokens by Apriori algorithm

<table>
<thead>
<tr>
<th>loop</th>
<th>accident</th>
<th>mile</th>
<th>accident</th>
<th>road</th>
</tr>
</thead>
<tbody>
<tr>
<td>exit</td>
<td>accident</td>
<td>major</td>
<td>accident</td>
<td>close</td>
</tr>
<tr>
<td>south</td>
<td>accident</td>
<td>left</td>
<td>accident</td>
<td>bad</td>
</tr>
<tr>
<td>accident</td>
<td>involve</td>
<td>accident</td>
<td>block</td>
<td>lane</td>
</tr>
<tr>
<td>495</td>
<td>accident</td>
<td>95</td>
<td>accident</td>
<td>66</td>
</tr>
<tr>
<td>car</td>
<td>accident</td>
<td>right</td>
<td>accident</td>
<td>rd</td>
</tr>
</tbody>
</table>
| traffic | near | accident | traffic-related information

As the traffic accident potentially influence the road traffic operations, the abnormal patterns of traffic-related information are potential features. It is also a viable method of monitoring the traffic operations in traditional studies (Coifman et al., 1998; Oh et al., 2001). Two major problems exist: first, the impact of traffic accident to its surround areas is unknown both in time and geographic scale; second, the traffic patterns are difficult to identify given the large volumes of historical data. Here we employ a systematic method to extract the traffic-related features.

#### 3.3 Feature extraction from traffic-related information

The recurrent traffic pattern of each detector can be unveiled by studying the historical traffic volume and occupancy data. For each detector, we evenly divide the traffic occupancy into \( N \) separate groups. For each traffic occupancy group, we take the median of the corresponding traffic flow values as the traffic signature. We use the median because it is less affected by outliers than mean. The traffic signature of a detector \( d \) is defined as the vector of these traffic flow values. That is:

\[
\mathbf{F}^d = (F_1^d, F_2^d, \ldots, F_0^d, \ldots, F_N^d).
\]

Where \( F_0^d \) is the median value of traffic flow given a range of occupancy \( o \) in detector \( d \). One can see that for each detector, the traffic pattern is a vector of \( N \) traffic flow values. If there is no traffic flow record over a certain occupancy, we employed the linear interpolation of traffic flow median of adjacent occupancies. We can finally obtain the traffic signatures of more than 15,000 detectors in over 1,250 signalized intersections. In this paper, \( N \) is set as 50.

According to the thorough study of the fundamental diagram (Jin and Ran, 2009), it is widely accepted that there exists a relationship between the traffic flow and occupancy (or density). However, the hypothesis of this relationship is diverse (e.g. triangle, parabola, trapezoid, broken-line, etc.). In our study, we do not make assumptions about
this relationship between $F^d$ and its corresponding occupancy. Instead, we assume the unchanged nature of the relationship:

- **Assumption 1**: there exists an unchanged traffic signature in a given location. The traffic flow corresponding to a certain occupancy interval will mostly fall into a reasonable range, and those that deviate from the feasible range are traffic outliers.

To validate the assumption, we employ the K-means algorithm without pre-defining the clustering centers and the number of clusters to reveal the relationship. K-means clustering algorithm can partition the traffic signatures into finite groups of similar patterns. The inputs are the traffic signatures of all detectors, and the outputs are the collection of cluster centers and the cluster IDs that detectors belong to. We employ Akaike information criterion (AIC) (Akaike, 1998) to find the proper number of clusters. AIC measures the relative quality of the clustering results, shown in the following equation.

$$AIC = \sum_{i}^{k} \sum_{d \in dom(i)} d(F^d_i, C^i) + k \cdot N$$

Where $F^d_i$ denotes the traffic signature of the $d$th detector that belongs to $i$th cluster. $C^i$ is clustering center of the $i$th cluster. $d(F^d_i, C^i)$ is the Euclidean distance between traffic signature $F^d_i$ and its clustering center $C^i$. $dom(i)$ is the domain (collection) of all detectors in $i$th cluster. $k$ is the current number of clusters. $N$ is the count of elements in a traffic signature, which equals to 50 in our study.

Our algorithm starts with the lower bound of the number of clusters and iterates the K-means clustering by increasing the cluster number. We calculate the AIC difference between the current iteration and the previous one. The iteration ends until the AIC difference is less than $\epsilon$. The algorithm is as follows:

**Algorithm**:

**Input**: The maximum number of clusters $K$, and traffic signature $F^d$ for all detectors. (in this study, our data has more than 15,000 rows and 50 columns. Each row represents the traffic signature of a detector.)

**Output**: Centers of clusters ($C^1, ..., C^i, ..., C^k$);
   Cluster IDs detectors belong to.

Assign the initial number of clusters $k=2$, initialize $AIC=+\infty$

Repeat

  Implement K-means clustering algorithm with $k$ clusters:

  Pick randomly the cluster centers ($C^1, ..., C^i, ..., C^k$);

  Repeat

  Cluster each traffic signature $F$ to the nearest cluster center $C^i$ with

  $\min(d(F^d_i, C^i))$;
Replace $C^t$ by $\text{mean}(F^{dt})$;

Until none of the detectors switch clusters;

Calculate the AIC difference between each cycle;

Until AIC difference $\leq \epsilon$ or $k=\kappa$

The AIC values will theoretically decrease with the increase of $k$. In this paper, we set $\epsilon$ as 3%. When $k=15$, the change in AIC goes lower than 3%, as shown in Figure 3.

![AIC values for different number of clusters](image3.png)

**Figure 3** AIC values for different number of clusters

Thus, we finally cluster nearly 15,000 detectors into 15 different groups. The centers of clusters are shown in Figure 4. From the shape of our clustering results, it is not
surprising that the relationships between traffic flow and occupancy differ greatly from each other. Unlike a predefined relationship, this method has certain advantages:

- The method is totally driven by the analysis of large-scale data. The aggregation analysis of large-scale data can lead to reduced noise in the results.
- The method clusters the traffic signatures with similar traffic patterns and potentially identifies the location of detectors that hold similar characteristics in the road network.
- The method excludes the influences of daily differences or time-of-day differences inherited in the traffic data.

### 3.3.2 Abnormal pattern identification and abnormal probabilities

The output cluster centers represent the relationship between traffic flow and occupancy. One can intuitively figure out the possible traffic outliers by comparing the clustered center to the original data as shown in Figure 5.

![Figure 5](image)

**Figure 5** Comparisons between clustered centers and the original traffic flow and occupancy data in two sample detectors

For each cluster, the traffic flows over a specified occupancy interval are distributed around their cluster centers. Further, the outliers can be quantified by a probabilistic method that measures its deviation degree. Our empirical examinations show that the distributions of the traffic flow in a particular cluster and occupancy interval follows a Gaussian distribution shown in Figure 6. The traffic outliers can be intuitively identified in the distribution tail.
Thus for each detector, the abnormal degree of traffic-related data can be quantified by the cumulative probability of the distribution.

\[ P_{dt} \quad \Phi \left( \frac{F_{o}^{dt} - C_{o}^{i}}{\sigma_{o}^{i}} \right) \]  

(5)

Where \( P_{dt} \) is the probability for detector \( d \) over time period \( t \). \( i \) indicates the \( i \)th cluster of \( d \); \( F_{o}^{dt} \) is the traffic flow data over traffic occupancy interval \( o \); \( \sigma_{o}^{i} \) and \( C_{o}^{i} \) is the standard deviation and center of traffic flow in Cluster \( i \) over occupancy interval \( o \). \( P_{dt} \) quantifies abnormal probability for the deviation of traffic data from its cluster centers. The larger \( P_{dt} \) is, the worse the traffic operations should be and the more likely the traffic is influenced by traffic accident.
As the geographic impact of the traffic accident may vary, the abnormal traffic data may exist in one or even more intersections nearby; thus, we can expect the overall traffic probabilities around the traffic accident site may increase. Besides, the traffic accident may happen either before or after when the tweet is blogged, the abnormal traffic data may be more apparent over a certain time period as compared to others; then we can also assume the abnormal probabilities may be much larger over this time period. Thus, to better quantify the traffic influence as to a tweet post, we mainly study the traffic related information within certain spatial and temporal ranges. The temporal ranges are set to be before and after one hour when the tweet is blogged. The spatial ranges are set to be 100m around where a tweet is blogged. Two features are then generated for our regression model for each tweet:

\[
p_{\text{traffic}} = \frac{1}{NUM} \sum_{t \in \text{dom}(t)} \sum_{d \in \text{dom}(d)} p_{dt} \quad (6)
\]

\[
q_{\text{traffic}} = Q3\{ P_{dt}, d \in \text{dom}(d) \cap t \in \text{dom}(t) \} \quad (7)
\]

Where \( t \) is the hour period; \( d \) is the detector ID and \( i \) is the cluster ID; \( \text{dom}(d) \) is the domain of all the detectors within the geo-scale of the tweets and \( \text{dom}(t) \) is the domain of all time periods within the time-scale of the tweets; \( Q3() \) is the operator of 75th percentile; \( NUM \) is the total number of traffic data related to a tweet.

4 Results

4.1 Regression results

Based on the model introduced in Sections 3, a model is built given joint prediction results from both the real-time traffic and tweet related data. The structure of the input data should be:

- **Time and Location**: Date and time, Geographic coordinate;
- **Tweet data**: Word features.
- **Traffic data**: Traffic flow, Traffic occupancy.

The “Time and location” information are obtained from the tweet data. The “Traffic data” refers to the data collection of traffic flow and occupancy in the detectors. Major features include those extracted from each token of the tweets, paired tokens of the tweets and the related traffic data. We further make a simple comparison between the prediction results with different features. To evaluate the achieved results in different models, we employed statistical metrics: accuracy and precision:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (8)
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \quad \frac{TN}{TN + FN} \quad (9)
\]

Where

| \( P_{\text{Model}} > 0.5 \) | \( P_{\text{Model}} \leq 0.5 \) |
The results shown in Figure 7 indicate that including both the association rules and traffic data can improve the detection rate of traffic accidents.

**Figure 7** Comparisons of accuracy and precision using different features; precision 1 is for accident, precision 2 is for non-accident.

It is worth mentioning that Figure 7 is built on the historical traffic and tweet data throughout the calendar year of 2014. One can see that the association rules within the tweet can apparently improve accuracy of the traffic accident detection. This means that some word combinations can predict the traffic accident more accurately than single ones. In comparison, the improvement for traffic related data may not be so significant as that of the paired tokens. There are mainly two reasons: first, there are much less traffic data features than the token features and the influence of traffic related features are not so significant as compared to other token features; second, the location and time information of the tweet may have some deviations. We can explore that by comparing the prediction results with the incident database from Virginia Department of Transportation (VDOT).

**4.2 Comparison with ground truth**

We further compare our model results with the traffic management log from Virginia Department of Transportation (VDOT). We can pair the accident-related tweets with the corresponding accident records by the latitude, longitude and time information in both database. The comparison results are insightful in studying the potentials of tweets in traffic accident detection.

We employ our model with features of single and paired tokens to predict the accident label of all tweets in NOVA. Given the accuracy of around 80%, we can generally estimate the geo-tagged tweets that are accident-related in NOVA area. According to our results, there are about 626 out of 584,000 tweets that are accident-related. Even though each of them directs to a different on-site accident, their amounts are still very small as compared to 52,496 accident records obtained from VDOT. Thus, it is worth mentioning that in our study, we only consider the geo-tagged tweets which take no more than 5% of
all tweets posted online. For those tweets without specific latitude and longitude
messages, we can possibly infer their locations according to their tweet messages (Ikawa
et al., 2012). Even though, the coverage of the tweets on traffic accidents cannot be high
enough to overtake other automatic accident detection methods. This is mainly because
the they are relatively small-scaled incidents (Schulz et al., 2013) and seldom arouse
public attentions. The influence of them may not be as high as that of earthquake or
festival parades, not all travelers are willing to leave a corresponding messages online.
Also, when passing by the site of traffic accident, most of the drivers cannot tweet about
it just for their own traffic safety.

Even though geo-tagged tweets have relatively low coverage, sometimes they still have
significance in traffic operations and management. We pair each accident-related tweet
with the accident records from management log within 6 km before and after 1 hour the
tweet is posted. Of more than 600 labeled accident tweets, there are about 300 of them
can be traced to an accident record by VDOT. The time when the tweets are posted can
be either earlier or later than the starting time of the traffic accident records as shown in
Figure 8(a). If the starting time in the traffic management log is the time when the police
arrives the accident site, about half of the accident-related tweets are posted earlier than
the traffic accident. This coincides with the findings in (D’Andrea et al., 2015) that tweets
detect traffic accidents more than 1 hour earlier than traditional media. If so, detecting the
accident-related tweets online can sometimes significantly reduce the response time.

![Figure 8](a) Time and (b) space difference between the accident-related tweets and the
accident records by VDOT.

The shortcomings are also obvious that the space differences are sometimes too large and
it is hard for traffic operators to pinpoint the accident site solely with the latitude and
longitude information of geo-tagged tweets. This also explains the reasons why traffic-
related information cannot improve the prediction accuracy because the travelers usually
tweet where they are far away from the accident sites. As compared with some events where tweeters express their instantaneous feelings on site, accident-related tweets are more probable to be a post-event recall. It is worth mentioning that these shortcomings can possibly be overcome by hinting the tweet contents.

Another interesting issues are some accident-related tweets express explicit meanings about the traffic accidents but cannot be traced by VDOT accident records. After our examinations, more than one third of these tweets are from the media channels such as “wtop”, “wtoptraffic”, etc. The locations for these tweets may not provide useful locations for accident detection. Other tweets may possibly be accounted by several reasons: Compared with the traffic management log maintained by VDOT, it is entirely possible that the tweets can capture the unexpected small events happened in our daily life. These events may include those “mild” accidents that do not incur the attention of traffic police and thus may not be included in the management log. The consequences of these events such as the road lanes blocking or cars slowing down may not last long and the corresponding affairs may come with a proper ending. If so, the tweets may act as a complement of the current accident detection system. Such as:

- “woooo got rear ended on i495 going to md great way to start a monday morning”
- “holy shit i just crashed my dads car”

Other reasons possibly exist: some of the accident-related tweets may be posted too far away from the accident site; some tweeters retweet about an accident instead of seeing in person; some tweeters may misjudge the situations and their inferences are from the jammed conditions of the roadways. For example:

- “sooo the car just said attention there is a car accident 12 miles ahead wtf kin of car does that”
- “major vehicle accident southbound i95 near lorton va traffic dmv”

After comparing with the ground truth, one can say that the tweets labeled by our model can possibly identify the existence of potential traffic accidents. This identifications may be faster than the traditional methods. The locations of the traffic accident may not be just exactly the latitude and longitude where the tweets are posted and the traffic operators should incorporate more information sources pinpoint the locations. In sum, it is entirely possible to increase the efficiency of traffic accident detection by monitoring the geo-tagged tweets.

6 Conclusions and discussions

In this paper, we investigate traffic accident detection models based on traffic and tweet data separately, and generate three important features: single token, paired token and traffic-related data to achieve a more accurate and effective on-site traffic accident detection. Our findings can be summarized as follows:

First, we thoroughly investigate the tweet contents related to traffic accidents. We found token features: single tokens and paired tokens that may correlate with the traffic accident labels. Our results show that paired tokens can possibly capture the association rules
inherent in the accident-related tweets and increase the accuracy of the traffic accident
detection.

Second, we unveil the relationships between traffic flow and occupancy based on the
fundamental diagram using large-scale data and point out that these relationships vary
different locations. We employ the K-means clustering algorithm to cluster the detectors
into different patterns of fundamental diagrams. The traffic flows over a certain range of
occupancy in a given cluster are observed to follow a Gaussian distribution. The derived
traffic-related information may provide limited improvement for accident prediction.

Third, the comparison between the prediction results and the traffic management log
maintained by VDOT provides insights in the studying the accident-related tweets: First,
sometimes the tweet reflection on the traffic accident is much faster than the traditional
methods and detecting the accident-related tweets online can sometimes significantly
reduce the response time. Second, tweets can sometimes capture those “mild” accidents
that do not incur the attention of traffic police and this indicates possibility of tweets
making up for the deficiencies of traffic management log. Third, some accident-related
tweets, include those posted by traditional media, are more probable to be a post-event
recall rather than an expressions of instantaneous feelings. These tweets cannot give an
exact location of the accident site and precise location detection should involve more data
sources.

Finally, it is concluded that integrating social media data into the traffic-related study
opens up a wide range of possibilities for research in on-site traffic accident detection.
The results show that social media data are very noisy and even unreliable, so solely
relying on social media data is still not a perfect option. Further studies can focus on the
data fusion of different data sources to better realize the purposes of other research such
as traffic jam detection, traffic emergency evacuation, etc. The spatial-temporal features
of traffic data are also worth studying for regional traffic operations. Note that our tweet
data and traffic data are labeled by both time and locations. It would be an interesting
extension to detect traffic event with non-geotagged tweets.

Acknowledgement

This research is supported by Transportation Informatics (TransInfo) University
Transportation Center at University at Buffalo, The State University of New York.
Reference


