# Tracking traffic congestion and accidents using social media data
## A case study of Shanghai

Chang, Haoliang; Li, Lishuai; Huang, Jianxiang; Zhang, Qingpeng; Chin, Kwai Sang

**DOI**
10.1016/j.aap.2022.106618

**Publication date**
2022

**Document Version**
Final published version

**Published in**
Accident Analysis and Prevention

Accident Analysis and Prevention 169 (2022) 106618

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Tracking traffic congestion and accidents using social media data: A case study of Shanghai

Haoliang Chang [a,*], Lishuai Li [b,c], Jianxiang Huang [d], Qingpeng Zhang [c], Kwai-Sang Chin [a]

[a] Department of Advanced Design and Systems Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[b] Faculty of Aerospace Engineering, Delft University of Technology, Postbus 5, 2600 AA Delft, Netherlands
[c] School of Data Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[d] Department of Urban Planning and Design, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

ARTICLE INFO

*Keywords:*
Traffic accident
Traffic congestion
Kernel density estimation
Social media data
Geographic information science
Natural language processing

ABSTRACT

Traffic congestion and accidents take a toll on commuters' daily experiences and society. Locating the venues prone to congestion and accidents and capturing their perception by public members is invaluable for transport policy-makers. However, few previous methods consider user perception toward the accidents and congestion in finding and profiling the accident- and congestion-prone areas, leaving decision-makers unaware of the subsequent behavior responses and priorities of retrofitting measures. This study develops a framework to identify and characterize the accident- and congestion-prone areas heatedly discussed on social media. First, we use natural language processing and deep learning to detect the accident- and congestion-relevant Chinese microblogs posted on Sina Weibo, a Chinese social media platform. Then a modified Kernel Density Estimation method considering the sentiment of microblogs is employed to find the accident- and congestion-prone regions. The results show that the 'congestion-prone areas' discussed on social media are mainly distributed throughout the historical urban core and the Northwest of Pudong New Area, in reasonably good agreements with actual congestion records. In contrast, the 'accident-prone areas' are primarily found in locations with severe accidents. Finally, the above venues are characterized in spatio-temporal and semantic aspects to understand the nature of the incidents and assess the priority level for mitigation measures. The outcomes can provide a reference for traffic authorities to inform resource allocation and prioritize mitigation measures in future traffic management.

## 1. Introduction

Road traffic is a major burden and source of public complaints in a large metropolis. Road accidents and congestion pose significant disturbances to drivers, commuters, and traffic management (D'Andrea et al., 2015). In China, road traffic accidents were accountable for 63,194 deaths and a direct economic loss of 1.38 billion RMB in 2018, according to the National Bureau of Statistics of China (2019). Moreover, traffic congestion also significantly restricts the improvement of urbanization of Chinese cities with different scales (Han et al., 2018). Therefore, finding and profiling the regions prone to accidents and congestion are crucial in road safety management.

Conventionally, these places were identified using historical accident records collected by surveys and sensors (Bíl et al., 2017; Tao et al., 2015). Based on statistical analysis, researchers characterized the spatio-temporal profiles of risky areas (Bíl et al., 2019) and identified factors contributing to the emergence of accident-prone areas (Wang et al., 2020). However, few studies consider the public perception of traffic accidents and congestion in identifying and characterizing the accident- and congestion-prone areas.

Recently, social media platforms, such as Twitter, Facebook, and the Chinese microblog platform Sina Weibo, have developed rapidly. In March 2011, the Sina Weibo service had gained tremendous popularity, with over 300 million registered Weibo users and over 100 million microblogs per day (Liu et al., 2012). By the middle of 2012, the number of registered Weibo users had reached 368 million (Millward, 2012). In 2015, over 300 million Twitter users globally posted about 500 million tweets each day (Lansley and Longley, 2016). Social media sites are information channels through which real-time traffic information is posted, a practice increasingly adopted by the transport authorities (Xu et al., 2018). The breadth, timeliness, and relatively low cost of traffic information accessible on social media (Serna et al., 2017) constitute

* Corresponding author.
*E-mail address:* hlchang4-c@my.cityu.edu.hk (H. Chang).

https://doi.org/10.1016/j.aap.2022.106618
Received 24 August 2021; Received in revised form 20 January 2022; Accepted 15 February 2022
Available online 26 February 2022

many advantages compared with the traditional traffic data obtained from physical sensors, which are costly to maintain and limited to venues with sensors (Cao et al., 2018). The time, text, and location information embedded in the social media posts can infer the time, location, and cause of traffic accidents or congestion. Nevertheless, this people-generated data has not been fully utilized in finding and profiling the accident- and congestion-prone areas in the city. A framework is necessary to transform the unstructured social media data into a clear presentation of accident- and congestion-prone regions for transport decision-makers.

Hence, this study aims to find and profile the accident- and congestion-prone areas heatedly discussed in the social network. A people-centric framework is built to identify and characterize these regions using eight months, over 14 million Chinese microblogs posted in Shanghai, China. Results depict the accident- and congestion-prone regions and provide suggestions for future traffic management in Shanghai. The main contributions of this study are as follows:

1. We developed a modified Kernel Density Estimation (KDE*) method, which incorporates the sentiment of traffic-related messages to identify the accident- and congestion-prone areas discussed on social media.
2. We conducted extensive experiments to show the relationship between the identified areas based on traffic-related Weibos and actual traffic records by employing multiple evaluation metrics.
3. We used social media factors (sentiment index and the number of accident- or congestion-related Weibos in each accident- and congestion-prone area) to characterize and prioritize the regions for traffic treatment. We further provided suggestions for future traffic management in Shanghai.

The remainder of this study is organized as follows. In Section 2, we mainly review the recent studies about traffic hotspot analysis and the application of social media data analysis in transportation research. Section 3 introduces the study area, Shanghai, and various datasets used in this study. Next, Section 4 presents the proposed framework, including detecting the traffic-related information, finding, and profiling the accident- and congestion-prone areas. Section 5 provides the analysis and results. Finally, Section 6 concludes this study with future works.

## 2. Literature review

Social media data plays a vital role in transportation information management research. However, this data source has not been fully utilized in finding and profiling the accident- and congestion-prone areas in the city. This section first discusses the previous studies about identifying and characterizing the high-density accident and congestion zones. Next, we review the application of social media data analysis in traffic-related studies.

### 2.1. Review of traffic hotspot analysis

The research field that is most relevant to our study is the traffic hotspot analysis, of which the target is to identify road segments or areas where the crash density is relatively higher compared to other parts of the studied area (Bíl et al., 2019; Harirforoush and Bellalite, 2019). The research works relevant to traffic hotspot analysis can be divided into 1) Identification and validation; 2) Understanding and characterization.

The first body of literature aims to identify the accident hotspots based on historical accident records. Compared to the methods merely based on the crash frequencies and crash rates, clustering is a more popular approach to find traffic hot zones in recent studies. The unsupervised clustering methods could be applied to cluster the locations into different groups and find the traffic hotspots based on the features such as the number of traffic accidents, economic loss, and the number of deaths and injuries. Xu and Tao (2018) used the Principal Component Analysis and K-means clustering approach to find the hotspots in the Anhui section of the G50 Hu-Yu highway. Moreover, Holmgren et al. (2020) run the DBSCAN algorithm to identify the unsafe areas for bicyclists in Lund, Sweden. Besides the clustering approach, another popular method is the Kernel Density Estimation (KDE), which computes the traffic accident point density over the studying space and highlights the areas with relatively high-density values compared to other parts of the study area. These areas are always regarded as the traffic hotspot. Yao et al. (2018) used KDE to identify the vehicle–pedestrian risk areas in Changning District, Shanghai. Al-Aamri et al. (2021) applied the KDE to find the traffic hotspots based on crash records of the Sultanate of Oman.

The second body of works aims to build a holistic understanding of traffic hotspots and inform traffic management, including more efficient patrols and more intelligent traffic information dissemination (Xie and Yan, 2008). Xie et al. (2017) studied the relationship between pedestrians-involved traffic crash cost and various indicators, including transportation, land use, sociodemographic, and social media usage. Results showed that transportation-related variables such as vehicle miles traveled (VMT), land use indicators such as the residential ratio, population, and the average number of tweets per year have a significant positive correlation with the crash cost. Wang et al. (2020) investigated the relationship between built environment attributes (e.g., road design, road safety, pedestrian safety, and traffic density) on road traffic crashes. The authors found that high-risk areas were associated with poor road conditions such as uneven road surface and slippery. Moreover, the pedestrian-oriented built environment, such as crosswalks and pedestrian traffic lights, also negatively impacted traffic safety in the study area.

The existing traffic hotspot studies primarily focus on accident hotspot identification and profiling. However, the congestion-prone areas have not been fully studied. Moreover, most analyses were based on the historical accident records collected by sensors, which neglected the commuter's perception of accidents. Commuters might be more concerned about accidents in a particular region, and specific traffic treatment is more urgently needed.

### 2.2. Review of social media data analysis in traffic-related studies

Social media data provides rich information from users' perspectives. Previous works tried to use the information embedded in social media data, such as time, text content, and geoinformation, to solve the problems in transportation studies, including traffic-related information detection (Ali et al., 2021; Chen et al., 2019; D'Andrea et al., 2015; Dabiri and Heaslip, 2019), traffic flow prediction (Ni et al., 2014), transport policy evaluation (Cao et al., 2014; Casas and Delmelle, 2017; Haghighi et al., 2018), and traffic condition analysis and information management (Ali et al., 2019b, 2019a; Cottrill et al., 2017). The research fields more relevant to this study are traffic information detection and traffic information management.

For traffic information detection in social media, researchers regard this problem as a text classification task and build machine learning models to detect traffic-related information. Researchers first use the bag-of-words feature to conduct traffic-relevant message detection. D'Andrea et al. (2015) manually labeled 1,330 tweets as traffic or non-traffic. They used the Support Vector Machine (SVM) (Cortes and Vapnik, 1995) as a classifier to determine whether a tweet is traffic-related or not. However, using the bag-of-words features cannot capture word meaning and the relationship between words (Mikolov et al., 2013). Recently, with the help of dense representation of words and deep learning, researchers in computer science have built text classification models by considering both the word meaning and the context (Kim, 2014). In the transportation-related studies, Convolutional Neural Network (CNN) (LeCun et al., 1995) and Long Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997) based on the

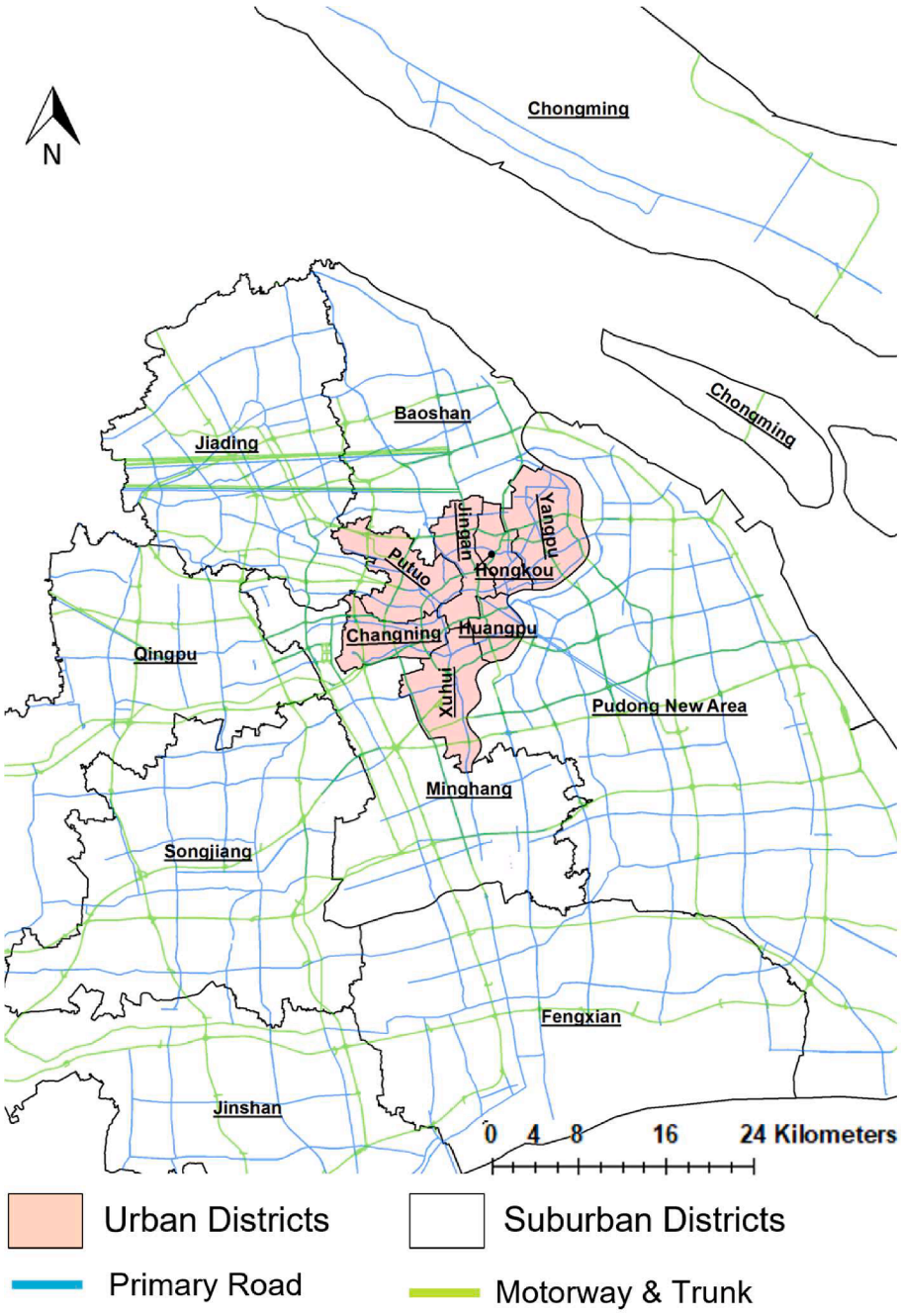


**Fig. 1.** Study Area: Shanghai.

dense vector representation of words are popular architectures to build the traffic information detection module (Chen et al., 2019; Dabiri and Heaslip, 2019). The developed identification models have reached over 90 percent accuracy in determining whether a social media message is traffic-related. However, these studies did not fully utilize spatial and temporal information of traffic-related social media data to reflect the traffic conditions. Ali et al. (2021) developed a real-time traffic monitoring framework based on accident-related social media data. The authors used an ontology-based LDA model to label the collected social media data and built the deep learning model to detect traffic-relevant social media messages. Combined with the sentiment analysis of detected information, the results could help the traffic authorities understand the city's traffic conditions.

Furthermore, social media data has become a new source to monitor traffic conditions in the traffic management domain. Wang et al. (2014) found that the most frequent congestion areas in a city have more traffic-related tweets than other areas. This finding can help researchers to detect the congestion-prone areas in a city. Ali et al. (2017) introduced a traffic information management framework incorporating the fuzzy ontology-based sentiment analysis and semantic web rule language (SWRL) rule-based decision-making to support safe traveling. The created opinion summary can help the traffic authorities find the traffic-related problems and provide guidance for travelers. Cao et al. (2018) used the sentiment of social media messages as an indicator to estimate the real-time traffic condition. The author found that microblogs with negative sentiment could help find places with traffic accidents or congestion.

The previous works show that analyzing traffic-relevant social media data can support urban traffic management. However, the repost behavior is a unique feature of social networks nowadays (Xu et al.,

**Table 1**
Road Traffic-relevant Dictionary.

| Chinese Word | English Translation | Chinese Word | English Translation |
|---|---|---|---|
| 堵, 拥堵, 阻塞, 拥挤 | Congestion | 快速路 | Expressway |
| 塞车 | Jam | 大桥 | Bridge |
| 车祸, 事故 | Accident | 路线 | Route |
| 剐蹭 | Sideswipe | 隧道 | Tunnel |
| 撞 | Bumping | 避让 | Avoiding |
| 追尾, 相撞 | Car Crash | 车距 | Car Spacing |
| 路况 | Real-time Traffic | 绕行 | Detour |
| 封道, 封路 | Road Closure | 畅通 | Smooth |
| 驾驶, 行车 | Driving | 立交 | Flyover |
| 路段 | Road Section | 高架 | Elevated Road |

2018) but was neglected by most previous traffic-related studies. Social media messages that ever reposted the traffic-related information, which might not contain any traffic-related keywords, reflect people's perception towards the discussed traffic event. Furthermore, social media data has not been utilized to find and profile the city's accident- and congestion-prone areas.

## 3. Study area and data description

In this section, we first introduce the study area, Shanghai. Then we describe three datasets used in this study, including the real-world accident and congestion records obtained from Sina Weibo account Lexing Shanghai, social media data collected from Sina Weibo, and administrative map and road network of Shanghai.

### 3.1. Study area: Shanghai

This study focuses on Shanghai, a major industrial and commercial center in China. Fig. 1 presents an overview of Shanghai districts and the spatial distribution of primary roads, motorways, and trunks[1]. According to the data released by Shanghai Urban and Rural Construction and Traffic Committee (2010), the modernized area of Shanghai expanded from 1,505 $km^2$ to 2,288 $km^2$. The rapid expanding of city is always accompanied with great challenge in traffic safety. In 2012, 2,256 traffic accidents happened in Shanghai, which caused 916 deaths (Shanghai Municipal Bureau of Statistics, 2015). Based on this situation, the traffic authorities in Shanghai have been continuously building platforms for better traffic information dissemination. In July 2011, the Traffic Command Center of Shanghai Transportation Commission registered a Sina Weibo account, Lexing Shanghai, to send real-time traffic information to the public. Citizens in Shanghai can access real-time traffic information immediately by following the Lexing Shanghai Sina Weibo account.

### 3.2. Accident and congestion records in Shanghai

Getting real-world traffic information from the China Public Security Bureau is very difficult. Instead, we chose to crawl the road traffic-related Weibos posted by Lexing Shanghai. We collected 108 accident-relevant records and 3,197 congestion-relevant records in this social media account from June 1 to November 30, 2012. Each datum includes the Weibo id, Weibo text, and posting time. These collected Weibos do not contain the location information of each road traffic event. Hence, we geocoded the location information embedded in the Weibo text. A more detailed description of how we conducted the Weibo text geocoding is given in Section 4.1.

[1] We use the urban district setting of Shanghai defined by Chang and Murakami, 2019.

### 3.3. Weibo data

#### 3.3.1. Weibo data collection and preprocessing

In this study, a Weibo crawler was built to fetch the Weibo data following Chen et al. (2019). We first set the search criteria (e.g., the time range and geographical coordinate boundaries). Then a Hypertext Transfer Protocol (HTTP) request was sent to the Sina Weibo platform to access the designated microblogs. The retrieved data include each Weibo's time, user ID, Weibo ID, text contents, and self-reported latitude and longitude information (if shared by the user). Moreover, to consider the repost behavior in Sina Weibo, we also collected the Weibos' repost information. If one Weibo user reposted one Weibo, we also saved basic information about the Weibo that this user reposted, including this Weibo's id and text. Finally, we collected 14,550,726 Weibos posted in Shanghai from April 1, 2012, to November 30, 2012. We further split the collected microblogs into two parts, one for building the traffic information detection module (3,673,714 Weibos posted between April 1 and May 30, 2012) and another for identifying and characterizing the accident- and congestion-prone areas (10,877,012 Weibos posted between June 1 and November 30, 2012).

After collecting the Weibo data, we cleaned the text of Weibos by removing the stop words, URLs, usernames, numbers, and punctuations. Secondly, we segmented the cleaned Weibo text into Chinese words, which can be used later to generate Weibo representation. Finally, we transformed the temporal information of each Weibo to the structured Python datetime object for the following temporal analysis.

#### 3.3.2. Building labeled data to train traffic information detection module

We started labeling the Weibo data for building the traffic Weibo detection module. We first expanded the keywords used by Chen et al. (2019) by adding the most frequent 20 traffic-related words found in historical Weibos posted in Lexing Shanghai. Next, we filtered the Weibos posted in April and May 2012 and obtained the candidate traffic-related microblogs. Each candidate microblog contains at least one traffic-related keyword. Table 1 shows the final keyword set we used to find the candidate traffic-relevant Weibos. Moreover, the accident-relevant words (车祸, 事故, 剐蹭, 撞, 追尾, 相撞) and congestion-relevant words (堵, 拥堵, 阻塞, 拥挤, 塞车) would be used later to find the corresponding type of Weibos.

After confirming the traffic-relevant keywords, we randomly sampled 6,000 candidate traffic-relevant Weibo messages posted in April and May 2012 and invited two annotators to label the candidate Weibos. Unlike previous works that only focused on deciding whether a Weibo is traffic-relevant or not (Chen et al., 2019; Dabiri and Heaslip, 2019), we wanted our detection module to find Weibos that can support the following spatial traffic-relevant analysis. Hence, we wanted our detection module to decide automatically:

- Whether a Weibo is traffic-relevant or not.
- If a Weibo is traffic-relevant, does it contain the location information of the described traffic event in the text?

Hence, we assumed any candidate traffic Weibo falling in one of the following categories:

- Road traffic-irrelevant.
- Road traffic-relevant but not Having Location Information.
- Road traffic-relevant and Having Location Information.

The detailed labeling process and some annotated Weibos are given in Appendix A. After the manually labeling process, we collected 3,589 road traffic irrelevant Weibos, 1,414 road traffic-related Weibos without location information in text, and 997 road traffic-related Weibos with location information in the text. The annotated Weibos would be used later to train the traffic information detection module.

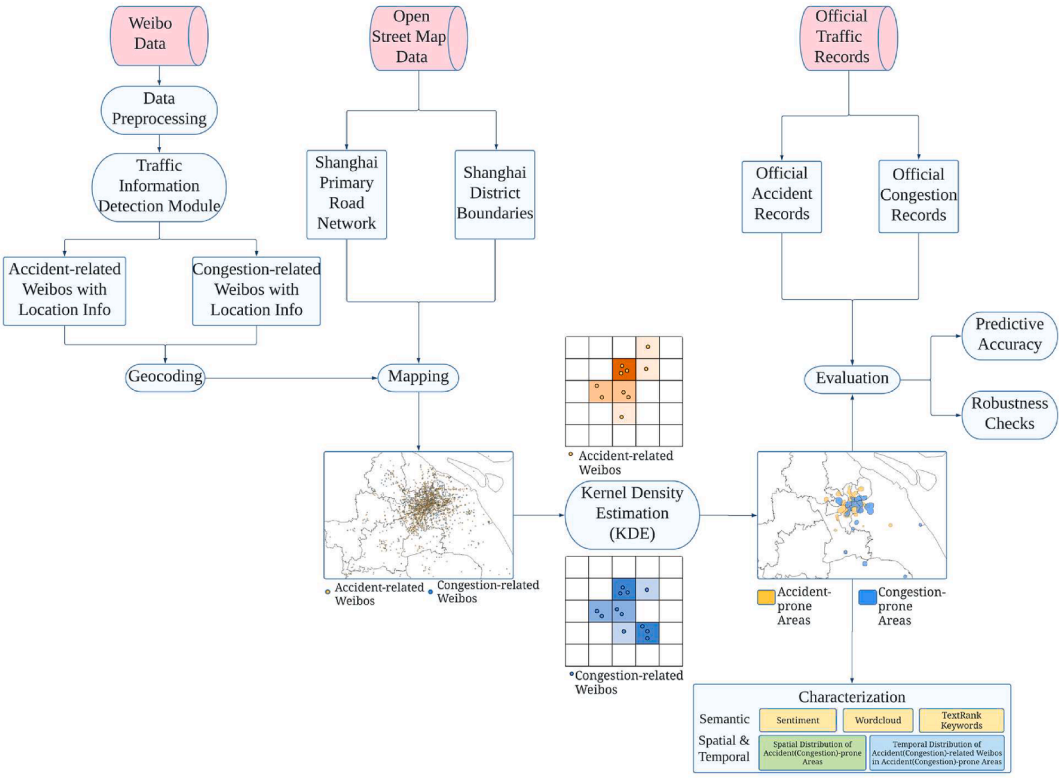


**Fig. 2.** Overview of the Proposed Framework for Finding and Profiling the Accident- and Congestion-prone Areas Based on Chinese Microblogs.

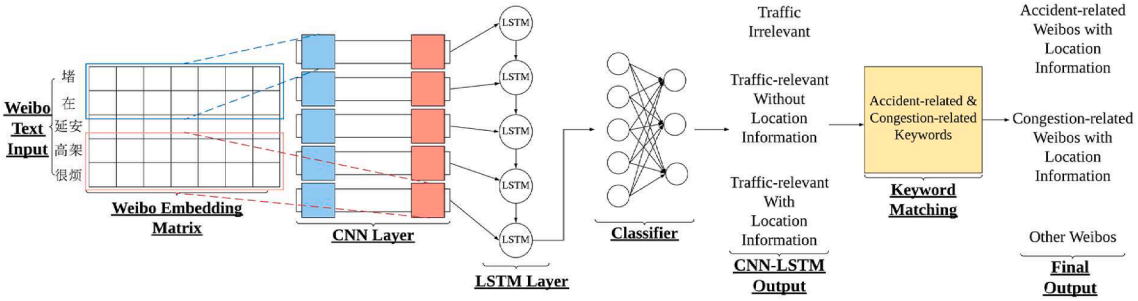


**Fig. 3.** Overview of the Traffic Information Detection Module.

### 3.4. Geospatial data

For spatial analysis, the map of Shanghai with district boundaries was downloaded from GaryBikini (2020). Moreover, to better characterize the spatial distribution of identified accident- and congestion-prone areas, we also downloaded Shanghai's primary road and motorway data from OpenStreetMap (Haklay and Weber, 2008). We transformed all the shapefiles into the Project Coordinate System UTM zone 51 N for later spatial analysis.

## 4. Methodology

The overview of our proposed framework for identifying and characterizing the accident- and congestion-prone areas based on Weibos is given in Fig. 2. The proposed framework first uses deep learning and natural language processing techniques to detect the accident-relevant and congestion-relevant Weibos with location information. Then the KDE* is employed to identify the corresponding accident- and congestion-prone areas based on the detected accident- and congestion-related Weibos. Finally, characterization is performed based on the spatial, temporal, and semantic information of accident-related or congestion-related Weibos posted in the detected high-density accident and congestion zones. The accident and congestion zones with relatively worse sentiment and the larger number of accident-related or congestion-related Weibos will be prioritized for future traffic treatment.

### 4.1. Accident and congestion Weibo detection and geocoding

This section aims to find the Weibos ever discussed the accidents and congestion with location description. Fig. 3 shows the overview of the detection module. The developed module will be applied to any candidate traffic-relevant microblogs and the reposted candidate traffic-relevant microblogs to find the Weibos discussing the accidents and congestion in Shanghai.

More specifically, we first trained the CNN-LSTM network based on the dense vector representation of Chinese words to find the traffic-relevant Weibos with location information. The CNN-LSTM module first inputs the vectors of words in Weibos. Then the CNN-LSTM module can be trained to detect the traffic-related Weibos with location infor-

**Table 2**
Performance Evaluation Metrics for Traffic Information Detection Module.

| Evaluation Index | Definition |
|---|---|
| Precision | $Precision = \frac{TP}{TP + FP}$ |
| Recall | $Recall = \frac{TP}{TP + FN}$ |
| $F_1$ score | $F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ |

mation. During training, we used the average $F_1$ score (Powers, 2020) to evaluate and compare CNN-LSTM models with different parameter settings. The definition of $F_1$ score is given in Table 2:

It is computed based on *TP* (true positive rate), *FP* (false positive rate), and *FN* (false negative rate). The CNN-LSTM model with the highest average $F_1$ score across the labels would be considered as the best-performed model. A detailed description of training the CNN-LSTM model is given in Appendix B.

Next, we used the best-performed CNN-LSTM module to find the traffic-related Weibos with location information. Then by matching the accident- and congestion-relevant keywords, we could detect the accident- and congestion-relevant Weibos with location description. We applied the following keyword search method to identify Weibos describing accidents or congestion:

- If a traffic-relevant Weibo contains at least one accident-related keyword but does not contain any congestion-related keyword, we regard this Weibo as an accident-related Weibo.
- If a traffic-relevant Weibo contains at least one congestion-related keyword but does not contain any accident-related keyword, we consider this Weibo as a congestion-related Weibo.
- If a traffic-relevant Weibo contains both accident-related and congestion-related keywords, we manually review the Weibo text. Suppose this Weibo text contains a detailed description of a traffic accident and corresponding location information. In that case, we regard this Weibo as accident-related since the accidents are usually the leading cause of the described congestion. Otherwise, we define this Weibo as a congestion-related Weibo.

After finding the accident-relevant and congestion-relevant microblogs with location information, Weibo geocoding was finally conducted. Some Weibo users shared their latitude and longitude information when posting their Weibos. However, most identified accident-related and congestion-related Weibos did not offer this precise location information. Table 3 shows some random sampled geocoding results:

We employed the following strategy to geocode the accident-relevant and congestion-relevant Weibos:

- If a traffic-relevant Weibo has self-reported latitude and longitude information, we utilized the latitude and longitude information shared by the users as the location information of the traffic accident or congestion described in this Weibo.
- If a traffic-relevant Weibo does not contain self-reported latitude and longitude information, we used the Google Geocoding API to geocode the text describing the accident or congestion with location information (Wu and Cui, 2018). The Geocoding API can find the location-related Chinese string in Weibos and convert them to the corresponding geographic location with coordinates[2]. We neglected the Weibos if only city-level location information was provided in the text.

### 4.2. KDE* for measuring the density of Accident- and Congestion-related Weibos

Once the geocoding process was completed, we started using the KDE* method, which considers the sentiment of traffic-related Weibos and identifies accident- and congestion-prone areas heatedly discussed in Weibos. More specifically, we first input the geocoded accident- and congestion-relevant Weibos and the boundary of Shanghai districts into ArcGIS 10.4.1 (ESRI, 2017). Afterward, we split the study area Shanghai into multiple spatial square units stored in a set *S*, given a unit size. Next, we started computing the kernel density values. For each spatial unit in the set *S* with its centroid locating at $(x, y)$, density was computed based on the Eq. (1), using the Planar Kernel Density toolbox released by ArcGIS 10.4.1 following the Luo and He (2021) and Ouni and Belloumi (2018):

$$Density_{(x,y)} = \begin{cases} \frac{1}{h^2}\sum_{i=1}^{n}\left[\frac{3}{\pi} \bullet sent_i\left(1 - \left(\frac{dist_{i,(x,y)}}{h}\right)^2\right)^2\right] & dist_{i,(x,y)} < h \\ 0 & dist_{i,(x,y)} \geq h \end{cases} \quad (1)$$

where the parameters are:

- $i = 1 \cdots n$: the accident-relevant or congestion-relevant Weibos.
- $h$: the search radius from the centroid $(x, y)$ using Euclidean distance, also known as the bandwidth.
- $sent_i$: The sentiment of the accident-relevant or congestion-relevant Weibos. We let $sent_i = 1$ if and only if the sentiment of this accident or congestion Weibo is negative. Otherwise, we set $sent_i = 0$. This setting lets us focus on the events that negatively influenced Weibo users.
- $dist_{i,(x,y)}$: The Euclidean distance between the accident- or congestion-relevant Weibo *i* and the centroid $(x, y)$ of the spatial unit.

Following Eq. (1), we calculated the KDE density values for all the spatial units in *S*, denoted as $Density_S$. Next, we set the threshold in Eq.

**Table 3**
Random Sampled Weibo Geocoding Results.

| Chinese Weibos | English Translation | Have Self-reported Latitude and Longitude or Not | Location Information |
|---|---|---|---|
| 暴雨, 到处都是积水, 到处都是塞车, 车祸也多! | Heavy rain, standing water everywhere, traffic jams everywhere, and many car accidents! | Yes | (22.536940, 114.054482) |
| 早上7点不到 延安高架就堵了15分钟! | Before 7 a.m., the Yan'an elevated highway was blocked for 15 min | No | Yan An Gao Jia Road, Shanghai (31.20985, 121.424534) |
| 偶堵在华浜新村了 | I am stuck in Huabang New Village | No | Changjiang Road, Baoshan District, Shanghai (31.349624, 121.492826) |

[2] The structure of geocoding results generated by Google Geocoding API can be found in: https://developers.google.com/maps/documentation/geocoding/intro#Results

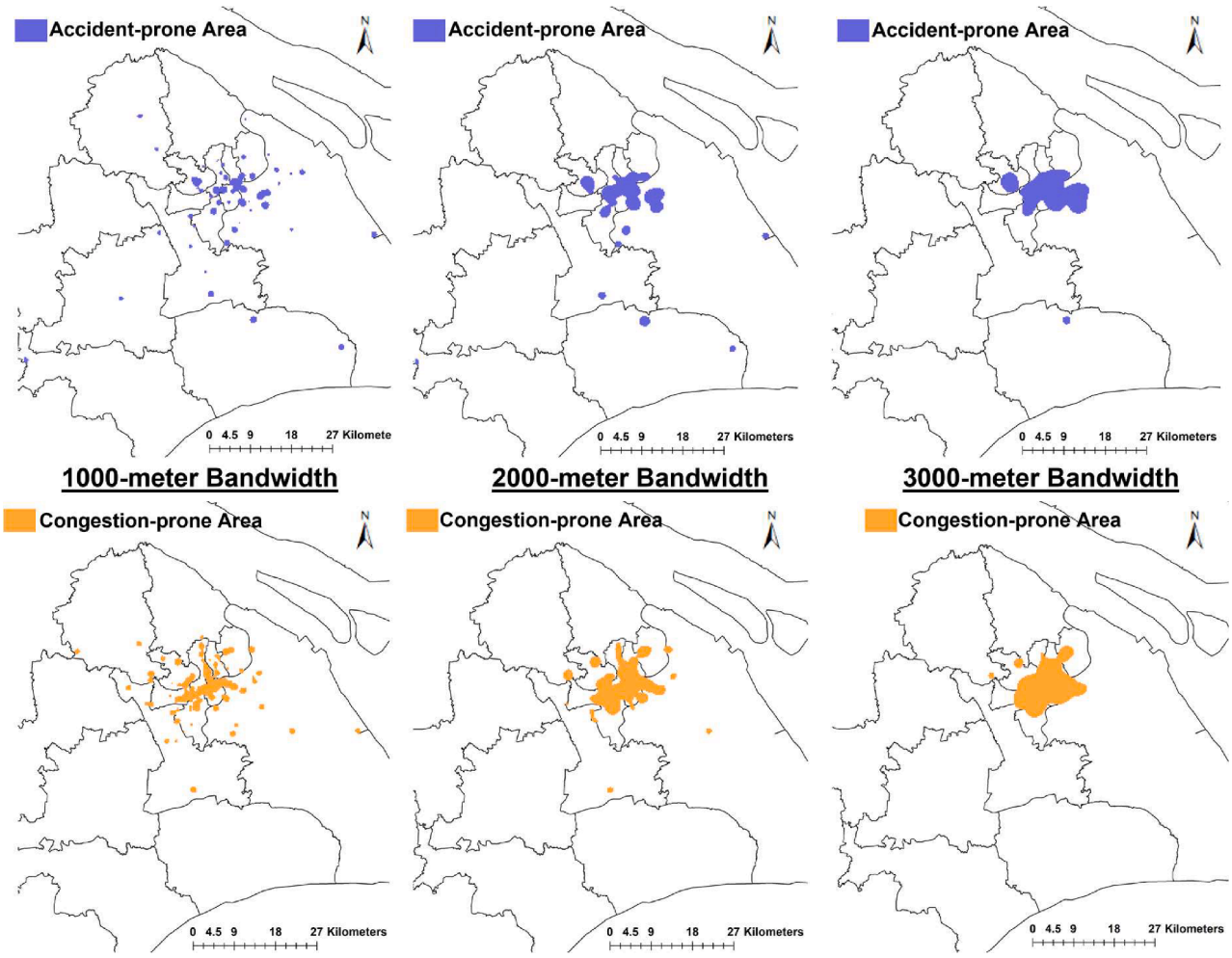


**Fig. 4.** Effect of Bandwidth in Finding the Accident- and Congestion-prone Areas.

**Table 4**
Spatial Unit Size and Sentiment Parameter Settings for KDE Modules.

| KDE Parameters | Parameter Settings |
|---|---|
| Spatial Unit Size (meter) | 200, 260, 300 |
| Consider Sentiment or Not | Regarding the Eq. (1):<br>• Consider sentiment: Let $sent_i = 1$ if and only if the sentiment of this accident-relevant or congestion-relevant Weibo is negative. Otherwise, let $sent_i = 0$.<br>• Do not consider sentiment: Let $sent_i = 1$ for all the accident-relevant- or congestion-relevant Weibos. |

(2) to identify the accident- and congestion-prone areas discussed in Weibos (Harirforoush and Bellalite, 2019):

$$threshold = Mean(Density_S) + 3 \times std(Density_S) \quad (2)$$

Where $Mean(Density_S)$ and $std(Density_S)$ represent the mean and standard deviation of all kernel density values. Finally, we picked the spatial units in $S$ which satisfied Eq. (3) and spatially combined them as the high-density accident or congestion zones:

$$Density_{(x,y)} \geq threshold \quad (3)$$

However, various KDE parameters will influence the density values, particularly bandwidth and spatial unit size (Ouni and Belloumi, 2018). This study also wants to check whether considering the sentiment information can help find the desired high-density accident and congestion zones. The best parameter settings should be used to find the accident- and congestion-prone areas. We first tried different bandwidth settings, and the effect is illustrated in Fig. 4. In this case, we set the spatial unit as a 260 m × 260 m block since the average block size of Shanghai from 2000 to 2015 was 6.8 ha (Angel et al., 2012). We made $sent_i = 1$ if and only if the sentiment of an accident or congestion Weibo is negative when computing the kernel density using Eq. (1). Otherwise, we made $sent_i = 0$. As the bandwidth increases from 1 *km* to 3 *km*, the hazardous locations begin to mix with their neighbors. The 1 *km* and 3 *km* are either too small or too large to make an informative interpretation of the detected high-density areas. Hence, a bandwidth of 2 *km* was chosen to find the accident- and congestion-prone regions.

For the other two parameters, spatial unit size and the sentiment settings, we explored the values shown in Table 4. The KDE modules with six different combinations of parameters would be compared based on the metrics presented in Section 4.3.

### 4.3. Performance evaluation of KDE methods

We compared the performance of KDE methods from two perspectives: predictive accuracy and robustness. To compare the predictive accuracy of KDE modules with different parameter settings, we employed the following metrics:

- Prediction Accuracy Index (PAI) (Chainey et al., 2008): Measure the ability of the KDE identification module to find areas with a high-density traffic-related Weibos.

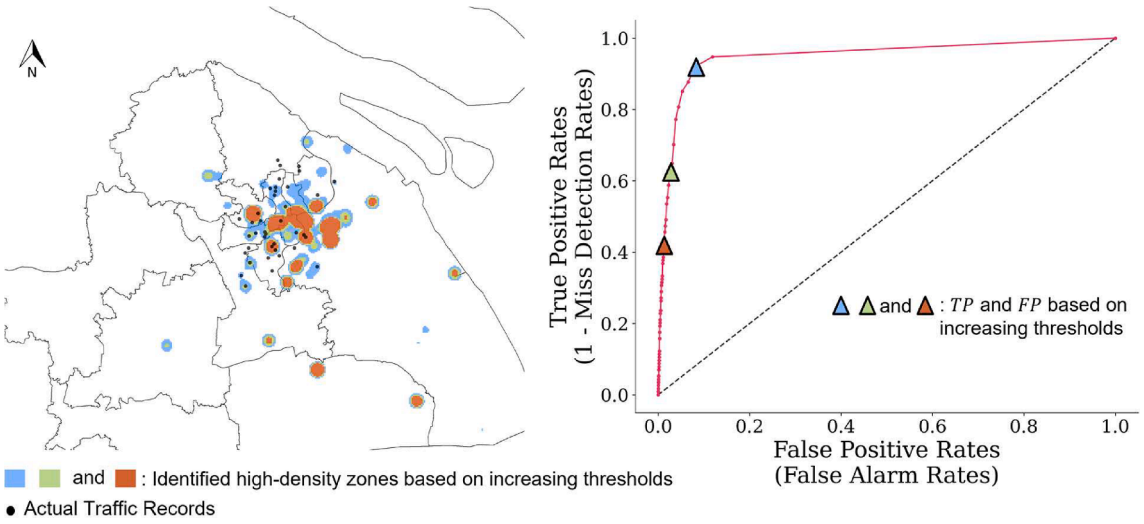


**Fig. 5.** Robustness Check of KDE Modules Given Different Thresholds.

**Table 5**
Parameter Settings of the Best-performed CNN-LSTM Module.

| Hyperparameter | Value |
|---|---|
| CNN Kernel Size | (2, 300) |
| Batch Size | 16 |
| Adam Learning Rate | 0.001 |

- Precision rate (Xia et al., 2019): Compute the overlapping rate between the high-density areas identified by traffic-related Weibos and the reference high-density areas based on actual traffic records.
- Miss detection rate: For the spatial units containing the actual traffic records, the miss detection rate is the proportion of units the identified accident- or congestion-prone areas do not cover.
- False alarm rate: For the spatial units not containing the actual traffic records, the false alarm rate is the proportion of units the identified accident- or congestion-prone areas falsely cover.

For the robustness checks, we used Receiver Operating Characteristic Curve (ROC Curve) (Powers, 2020) given different threshold values to check the robustness of the KDE modules in distinguishing the spatial units containing actual traffic records from the other spatial units.

Firstly, The PAI is defined in Eq. (4). Larger PAI values mean a better ability of one KDE module to locate the areas with high-density point features.

$$PAI_{weibo} = \frac{\frac{n}{N} \times 100}{\frac{m}{M} \times 100} \tag{4}$$

Where:

- $n$: Number of accident-relevant or congestion-relevant Weibos in the high-density accident and congestion zones
- $N$: Total number of accident-relevant or congestion-relevant Weibos in Shanghai
- $m$: The area of the high-density accident and congestion zones, in square kilometers
- $M$: Total area of Shanghai, in square kilometers

Secondly, the precision rate is computed as in Eq. (5):

$$Precision = \frac{Area_{identical}}{Area_{zone}} \tag{5}$$

**Table 6**
Confusion Matrix of Best-performed CNN-LSTM Module on the Test Data.

| Predicted / Actual | Traffic-irrelevant | Traffic-relevant but not Having Location Info | Traffic-relevant and Having Location Info |
|---|---|---|---|
| Traffic-irrelevant | 622 | 63 | 26 |
| Traffic-relevant but not Having Location Info | 79 | 189 | 13 |
| Traffic-relevant and Having Location Info | 42 | 25 | 141 |

Where $Area_{zone}$ is the area computed by the KDE method based on traffic-related Weibos. $Area_{identical}$ denotes the identical area between high-density zones identified by traffic-relevant Weibos and reference high-density zones detected by actual traffic records. We ensured that we used the same spatial unit size to identify high-density areas based on traffic-related Weibos and actual traffic records.

Thirdly, the miss detection and false alarm rates measure how effectively the KDE modules capture the spatial units containing the actual traffic records. For each spatial unit $i$, we created the following hypothesis test:

$H_{i0}$(Null Hypothesis): Spatial Unit $i$ is not a component of the accident- or congestion-prone areas

$H_{i1}$ (Alternative Hypothesis): Spatial Unit $i$ is a component of the accident- or congestion-prone areas

Then the miss detection rate can be calculated as in Eq. (6):

$$\begin{aligned} Miss\ Detection\ Rate &= \frac{\begin{array}{c}Num\ of\ Units\ Containing\\ Actual\ Traffic\ Records\ but\ not\\ Covered\ by\ Identified\ Zone\end{array}}{\begin{array}{c}Num\ of\ Units\ Containing\\ Actual\ Traffic\ Records\end{array}} \\ &= 1 - TP \end{aligned} \tag{6}$$

The false alarm rate is measured as in Eq. (7):

**Table 7**
Some Prediction Results Generated by CNN-LSTM Module.

| Weibo | English Translation | Prediction |
|---|---|---|
| 【路况信息:部分快速路匝道地面积水, 司机请绕行】南北东侧徐家汇路至永兴路、共和立交至临汾路等路段较堵 | [Traffic information: some expressway ramps are watery, drivers please detour] **The road sections from Xujiahui Road to Yongxing Road, Gonghe Interchange to Linfen Road on the east side of the North-South Elevated Road** are quite congested | Traffic-relevant and Having Location Info |
| 每到周一总是要迟到! 讨厌的堵车![怒][怒] | Always be late every Monday! Disgusting traffic jam! [Anger] [anger] | Traffic-relevant but not Having Location Info |
| 我在顶点撞球俱乐部。吃饱了, 来运动运动 | I am at the Apex Pool Club. I am full, come to exercise | Traffic-irrelevant |

**Table 8**
Number of Accident- and Congestion-relevant Weibos with Location Information from June 1 to November 30, 2012.

| | Accident-related | Congestion-related | Total |
|---|---|---|---|
| Having Self-reported Latitude and Longitude Information | 341 | 1,474 | 1,815 |
| Having Location Information in the Weibo or Reposted Weibo Text | 2,985 | 2,416 | 5,401 |
| Total | 3,326 | 3,890 | 7,216 |

$$\begin{aligned} \textit{False Alarm Rate} &= \frac{\begin{array}{c}\textit{Num of Units not Containing}\\ \textit{Actual Traffic Records}\\ \textit{but Covered by Identified Zone}\end{array}}{\begin{array}{c}\textit{Num of Units not Containing}\\ \textit{Actual Traffic Records}\end{array}} \\ &= FP \end{aligned} \tag{7}$$

Finally, the general process of robustness check is illustrated in Fig. 5.

A larger threshold would make the accident- or congestion-prone areas less likely to capture enough actual traffic records. However, a smaller threshold would make the identified area too broad. In this study, we tried 70 different threshold values, ranging from 0 to the maximum density values of all the spatial units, to evaluate the robustness of the KDE modules in separating the spatial units containing actual traffic records from others.

### 4.4. Profiling and mitigation prioritization

To characterize and prioritize the accident- and congestion-prone regions, we considered the spatial, temporal, and semantic aspects of Weibos with location information posted in the identified areas to answer the following questions:

1. What kind of traffic events were people discussing in the accident- and congestion-prone areas?
2. Which accident and congestion zones should be prioritized for traffic treatment?

To answer the above questions, we conducted the following analyses:

1. For the first question, we manually reviewed the Weibos posted in the identified areas with a combination of word cloud and keywords generated from the Weibo text. A more detailed description of Weibo text analysis is in Section 4.5.2.
2. For the second question, we considered the sentiment index and the number of accident-related or congestion-related Weibos posted in the accident- or congestion-prone areas, respectively. We would prioritize the areas with sentiment indices below the 50th percentile and the total numbers of accident-relevant or congestion-relevant Weibos above the 50th percentile. Section 4.5.1 presents the details about the Weibo sentiment analysis and sentiment index.

### 4.5. Social media data analysis

#### 4.5.1. Weibo sentiment analysis and sentiment index

To determine the sentiment of Weibo, we used the Baidu Sentiment Analysis API (Tian et al., 2020) to estimate the sentiment of each Weibo. The API can output the probability of any text string being positive or negative, based on the word's meaning and sequence. To make the sentiment analysis API assign Weibo with the neutral sentiment, we randomly sampled some collected Chinese microblogs, ran the sentiment analysis API, and set a threshold for sentiment labeling. If Positive probability – Negative probability $> 0.4$, we regard this Weibo as positive; If Negative probability – Positive probability $> 0.4$, the Weibo is negative; Otherwise, the sentiment of Weibo is neutral.

Moreover, to obtain the overall sentiment towards one identified accident- or congestion-prone area discussed on Weibo, the sentiment index was measured by subtracting the percentage of negative accident- or congestion-related Weibos from the positive accident- or congestion-relevant Weibos posted in this area (Durahim and Coşkun, 2015).

#### 4.5.2. Weibo text content analysis

This study first used the word cloud to reflect the semantic information in the accident- and congestion-related Weibos. The frequency of a word is related to its size in the word cloud. However, the word cloud cannot present the high-level information of a corpus. Hence, we also applied the TextRank (Mihalcea and Tarau, 2004) algorithm, a popular approach extracting keywords from a text corpus, and summarized accident- or congestion-related Weibos in the accident- and congestion-prone areas.

**Table 9**
Performance Comparisons of KDE Modules in Finding the Accident-prone Areas.

| Unit Size (meter) | Sent or not | Area ($km^2$) | Num of Actual Accident Records | Num of Accident Weibos | $PAI_{weibo}$ | Precision % | Miss % | False % |
|---|---|---|---|---|---|---|---|---|
| 200 | Yes | 96.88 | 20 | 1,965 | 42.78 | 23.91 | 79.73 | 1.37 |
| 260 | Yes | 97.28 | 21 | 1,969 | 42.69 | 24.32 | 78.08 | 1.37 |
| 300 | Yes | 97.20 | 21 | 1,976 | 42.88 | 25.09 | 78.08 | 1.37 |
| 200 | No | 62.52 | 14 | 1,685 | 56.84 | 27.13 | 86.49 | 0.89 |
| 260 | No | 62.39 | 15 | 1,685 | 56.96 | 28.06 | 84.93 | 0.88 |
| 300 | No | 62.19 | 16 | 1,719 | 58.30 | 28.65 | 83.56 | 0.87 |

**Note**: Sent or not means considering the sentiment or not in the KDE analysis. Precision % means the precision rate. Miss % stands for the miss detection rate. False % represents the false alarm rate.

**Table 10**
Performance Comparisons of KDE Modules in Finding the Congestion-prone Areas.

| Unit Size (meter) | Sent or not | Area ($km^2$) | Num of Actual Congestion Records | Num of Congestion Weibos | $PAI_{weibo}$ | Precision % | Miss % | False % |
|---|---|---|---|---|---|---|---|---|
| 200 | Yes | 128.64 | 2,067 | 2,181 | 30.57 | 51.34 | 46.88 | 1.79 |
| 260 | Yes | 128.51 | 2,163 | 2,186 | 30.68 | 51.66 | 48.39 | 1.76 |
| 300 | Yes | 128.25 | 2,156 | 2,179 | 30.64 | 52.14 | 47.40 | 1.73 |
| 200 | No | 115.92 | 2,186 | 2,126 | 33.07 | 55.00 | 45.63 | 1.60 |
| 260 | No | 114.45 | 2,054 | 2,107 | 33.20 | 56.17 | 47.74 | 1.56 |
| 300 | No | 116.46 | 2,160 | 2,129 | 32.97 | 55.41 | 47.40 | 1.56 |

**Note**: Sent or not means considering the sentiment or not in the KDE analysis. Precision % means the precision rate. Miss % stands for the miss detection rate. False % represents the false alarm rate.

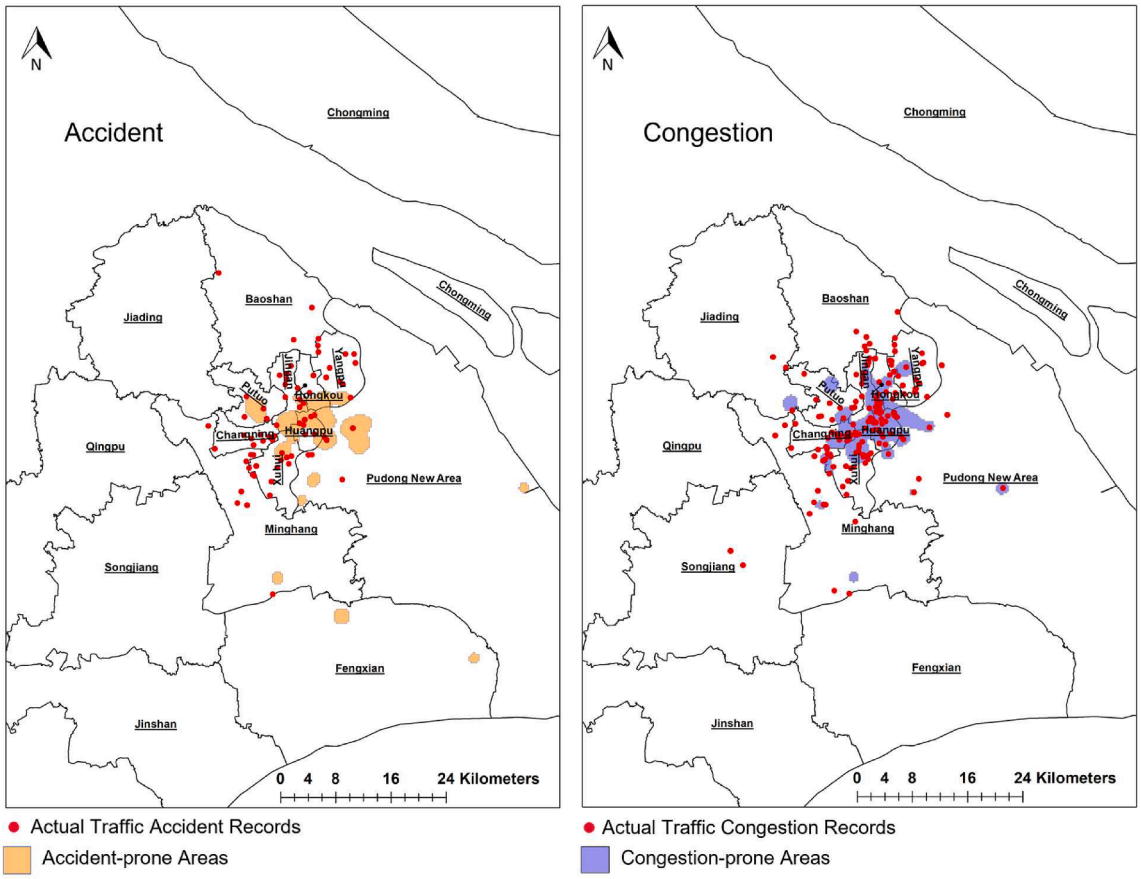


**Fig. 6.** Spatial Distribution of Accident- and Congestion-prone Areas and Actual Accident and Congestion Records.

## 5. Analysis and results

### 5.1. Accident- and Congestion-related Weibo detection

In this section, we present the traffic information detection results. Table 5 shows the hyperparameter setting of the best-performed CNN-LSTM module, which reaches a 0.76 average $F_1$ score on the test data.

The confusion matrix showing the performance of this module on test data is given in Table 6. The performance is promising as the detection module performs well on each type of label. Some random sampled Weibo in the test data with the model's prediction is given in Table 7. The location information has been highlighted.

Finally, we ran the CNN-LSTM module on all the Weibos posted between June 1 and November 30, 2012. Then by matching the accident-related and congestion-related keywords, we detected the accident-relevant and congestion-relevant Weibos with location information. The number of Weibos regarding the traffic type and the type of offered geo-information is given in Table 8.

### 5.2. Identification of Accident- and Congestion-prone areas

To select the most appropriate parameter settings of KDE modules in finding the accident- and congestion-prone areas, Tables 9 and 10 compare the performance of KDE modules with different unit sizes and sentiment settings based on evaluation metrics described in Section 4.3.

Current results indicate that sentiment setting significantly impacts the shape of accident- and congestion-prone areas. When we fixed the unit size and checked the sentiment's role in the identification process, the high-density accident zones detected by KDE with sentiment were about 50% larger than their counterparts. On the contrary, there was no substantial difference between the identified high-density congestion zones in terms of covered area and the number of captured congestion-related actual records and Weibos. Even though the accident zones identified by KDE with sentiment had lower PAI values, lower precision rates, and higher false alarm rates, they reached relatively lower miss detection rates. The high-density accident zones identified by KDE with sentiment cover more accident-related Weibos and actual accident records. However, the sentiment information did not contribute to finding

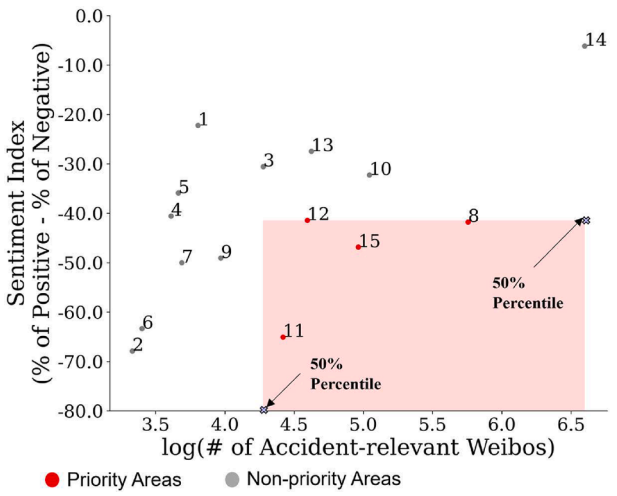


**Fig. 7.** Scatter Plot of Accident-prone Areas Showing Sentiment Index Against log (Number of Accident-relevant Weibos).

the congestion-prone areas.

On the other hand, the spatial unit size had limited influence on identified high-density locations. Little difference was found in the number of captured accident- and congestion-related official records and Weibos, precision rates, false alarm rates, and miss detection rates. Larger unit size is more appropriate when identifying the accident-prone areas from current results. Moreover, we used the 260-meter spatial units to find the congestion-prone areas.

Based on the above analysis, we used the following parameter settings to find the accident- and congestion-prone areas:

- Accident-prone area: spatial unit size: 300-meter; sentiment setting: considering sentiment.
- Congestion-prone area: spatial unit size: 260-meter; sentiment setting: not considering sentiment.

The robustness checks of KDE modules are presented in Appendix C. There is little difference between different KDE modules in separating the spatial units containing the actual accident or congestion records from other units.

Fig. 6 presents the spatial distribution of the accident-prone areas, congestion-prone areas, actual accident records, and actual congestion records. We found that the accident-prone areas and actual accident records did not overlap much. It might be because people like to discuss road accidents with severe consequences on social media. The accident-prone areas detected by social media are more event-driven and more sensitive to locations where severe traffic accidents happen. Moreover, the scale of some traffic accidents reported in social media is small (such as some minor rear-end collisions). Official traffic Weibo accounts did not record these accidents. On the other hand, many real-world traffic congestion records fell in the identified congestion-prone areas, mainly spreading the central urban districts. The urban districts of Shanghai are known for their high congestion level. Meanwhile, when Weibo users encounter traffic congestion on the road, some might also post complaints about the traffic congestion they encounter. Hence, congestion-relevant Weibos are more location-driven and tend to be spatially distributed in places where traffic congestion always happens.

### 5.3. Characterization of Accident- and Congestion-prone areas

After finding the high-density accident and congestion zones, the final step is to characterize these areas for future traffic treatment. We found that:

- From the spatial perspective, the junction region between the Pudong New Area, Hongkou, Jing'an, and Huangpu Districts was the area Weibo users worried about most based on the spatial distribution of accident- and congestion-prone areas. The possible reason is that this junction region covers essential roads such as Nanpu Bridge and Shanghai Bund Tunnel, which encompass large traffic volumes between the urban districts and Pudong New Area.
- From the temporal perspective, the accident- and congestion-related Weibos were mainly posted in days when severe traffic accidents and congestion happened.
- From the semantic perspective, primary roads, including Nanpu Bridge, Shanghai Bund Tunnel, North-South Elevated Road, and Yan'an Elevated Road, were frequently mentioned in both the accident- and congestion-related Weibos in the prioritized accident and congestion zones. Moreover, abnormal weather, poor traffic control under extreme weather and before special events, and driving over

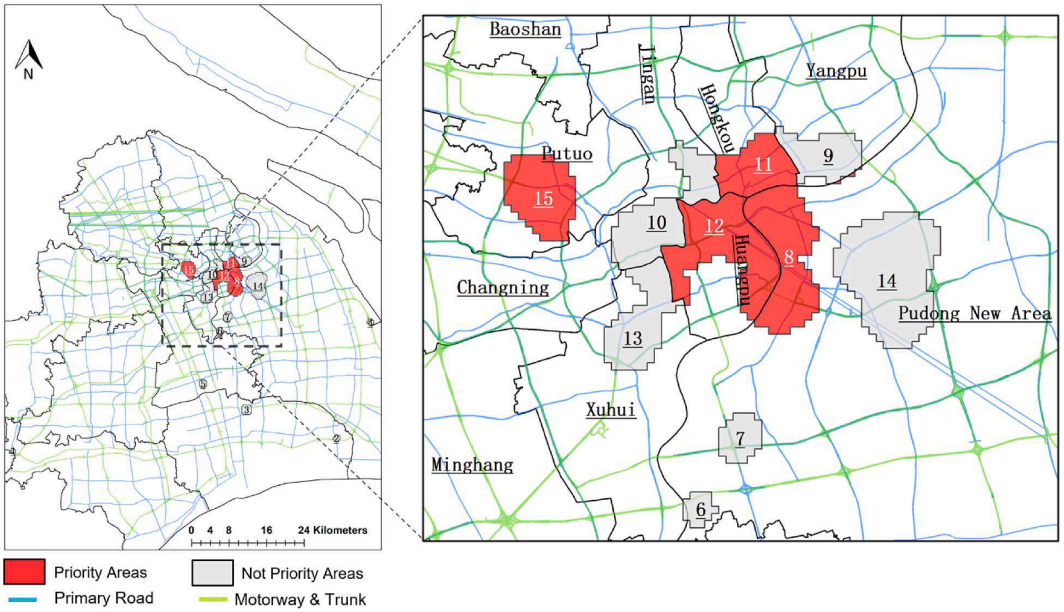


**Fig. 8.** Spatial Distribution of Priority and Non-priority Accident-prone Areas.

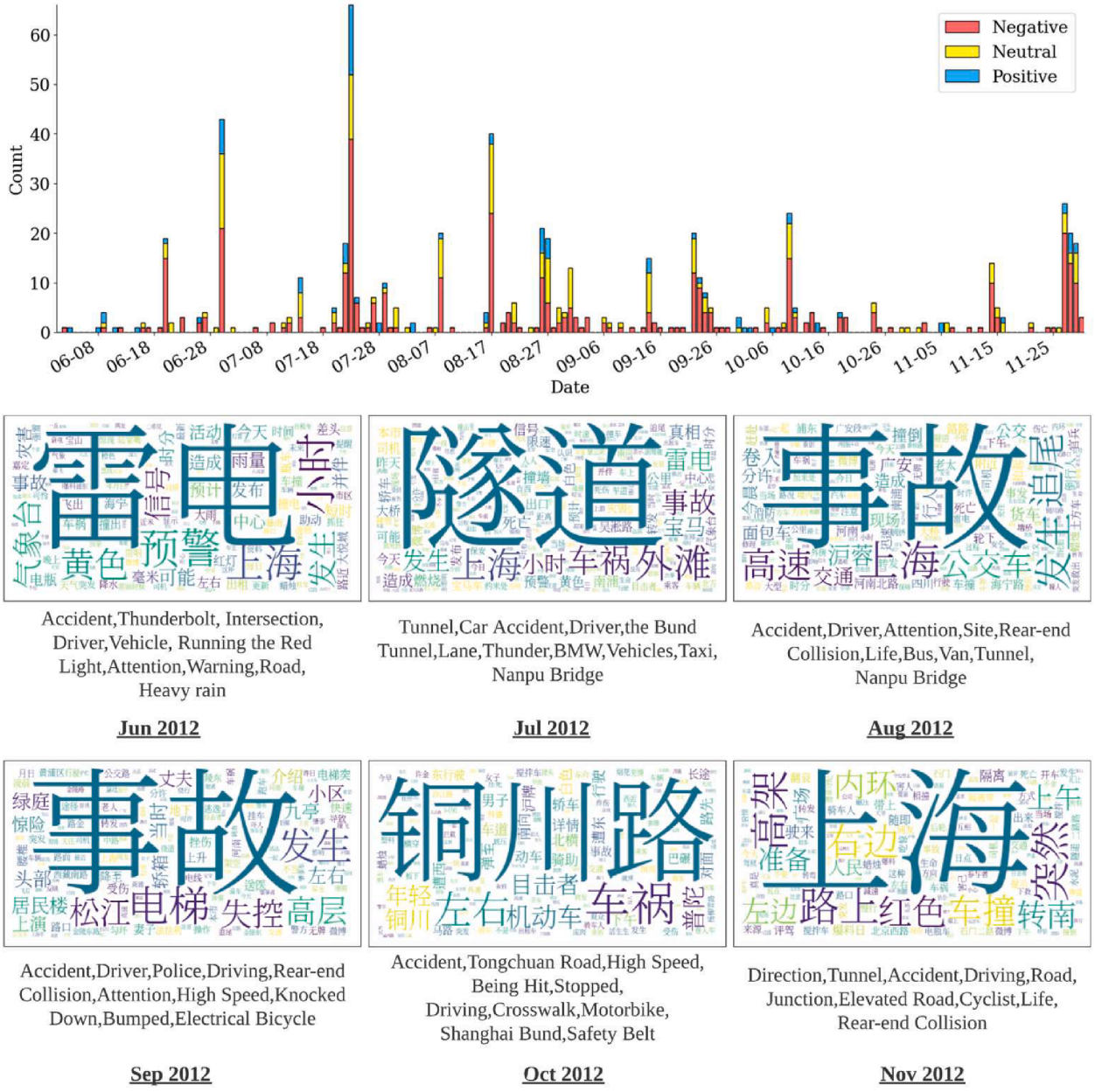


**Fig. 9.** Temporal and Semantic Characterization of the Prioritized Accident-prone Areas.

the speed limit were the leading causes of accidents and congestion in the prioritized accident and congestion zones.

For the accident-prone areas, we identified 15 accident zones with a concentration of accident-related Weibos[3]. Fig. 7 shows the sentiment index against the natural log of the number of accident-relevant Weibos for each identified accident-prone area. Accident zones 8, 11, 12, 15 would be prioritized for future traffic management, since their sentiment indices were below the 50th percentile and the number of accident-relevant Weibos exceeded the 50th percentile.

Furthermore, Fig. 8 presents the spatial distribution of priority and non-priority accident zones based on detected accident-related Weibos. As Fig. 8 shows, from the geographic perspective, most of these prioritized accident zones spread the junction region between the Pudong, Hongkou, Jing'an, and Huangpu Districts. Accident zone 15 was in the center of the Putuo District.

Fig. 9 presents the number of positive, neutral, and negative Weibos posted per day in the prioritized accident zones. Meanwhile, the word cloud and the traffic-related keywords extracted from the accident-related Weibos each month were also given. We discovered that accidents were more frequently discussed in the first three months (June, July, and August) compared to the last three months (September, October, and November). Furthermore, places such as Pudong New Area, Shanghai Bund Tunnel connecting the accident zone 8 and 12, and Tongchuan Road in accident zone 15 were mentioned many times in the accident-relevant Weibos. We further manually read the Weibo text to

[3] We intersected all the identified high-density accident zones with the Shanghai district boundaries, spatially joined the adjacent polygons, and regarded each part as an individual area. We conducted the same spatial processing in finding the priority and non-priority congestion zones.

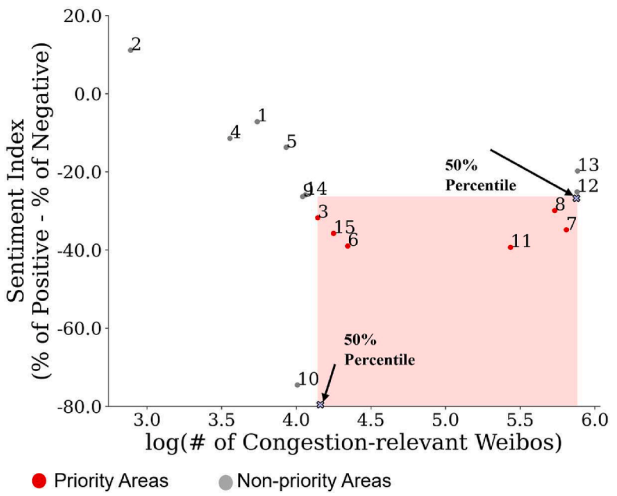


**Fig. 10.** Scatter Plot of Congestion-prone Areas Showing Sentiment Index Against log (Number of Congestion-relevant Weibos).

summarize the cause of accidents in the prioritized accident zones. The main reasons for the accidents in these prioritized accident zones were poor weather, driving over the speed limit, pedestrians not following the traffic rules, and illegal parking of large trucks.

For future traffic management, placing the speed limit signs on the roadside of the major roads in the prioritized accident-prone areas, especially the Shanghai Bund Tunnel in Accident-prone area 8, should be considered. Moreover, more effective traffic management on primary roads of these regions under poor weather needs to be adopted.

For the congestion-prone areas, 15 high-density congestion zones were identified. Fig. 10 presents the sentiment index, natural log of the number of congestion-relevant Weibos in each detected congestion zone. Based on the sentiment index and the number of congestion-relevant Weibos in each identified area, we chose to prioritize the congestion zones 3, 6, 7, 8, 11, and 15 in the following congestion characterization since their sentiment indices were below the 50th percentile and total numbers of congestion-relevant Weibos exceeded the 50th percentile.

Moreover, the spatial distribution of priority and non-priority congestion zones is shown in Fig. 11. Regarding the prioritized congestion zones, like the priority accident zones shown in Fig. 8, the priority congestion zones also covered the Northwest corner of Pudong New Area. Congestion zone 8 and 11 spread across the Jing'an and Changning Districts. Congestion zone 15 emerged at the North of Putuo District. While congestion zone 6 was in the west of Yangpu District, congestion zone 3 was located at the intersection region between Minhang and Xuhui Districts.

Fig. 12 further characterized the congestion zones on priority from the temporal and semantic perspectives. Compared to the temporal distribution of accident-related Weibos presented in Fig. 9, congestion-related Weibos in priority congestion zones were mainly concentrated on specific days, including August 8 (Typhoon Haikui hit Shanghai) and September 29 (two days before the China National Day). Moreover, places including the Nanpu Bridge in congestion zone 7, North-South Elevated Road in congestion zone 8, Yan'an Elevated Road in congestion zone 11, Zhen Bei Road and Wu Wei Road in congestion zone 15, and Pudong New Area were frequently mentioned. After manually reviewing the congestion-relevant Weibos posted in the prioritized congestion-related locations, the priority congestion zones emerged because of the enormous traffic volume, sudden worsening of traffic conditions caused by typhoons, and poor traffic management before the China National Day.

In future traffic management, improving the ability of the current traffic system to deal with extreme weather, such as upgrading the water drainage system of the Zhenbei Road and the Wuwei road in congestion zone 15, should be considered. Moreover, more effective traffic management should be implemented on the North-South Elevated Road in congestion zone 8 before the special event (like the China National Day) to ensure travel efficiency. Nanpu Bridge in congestion zone 7, which connects the Pudong New Area and the Shanghai urban districts, was

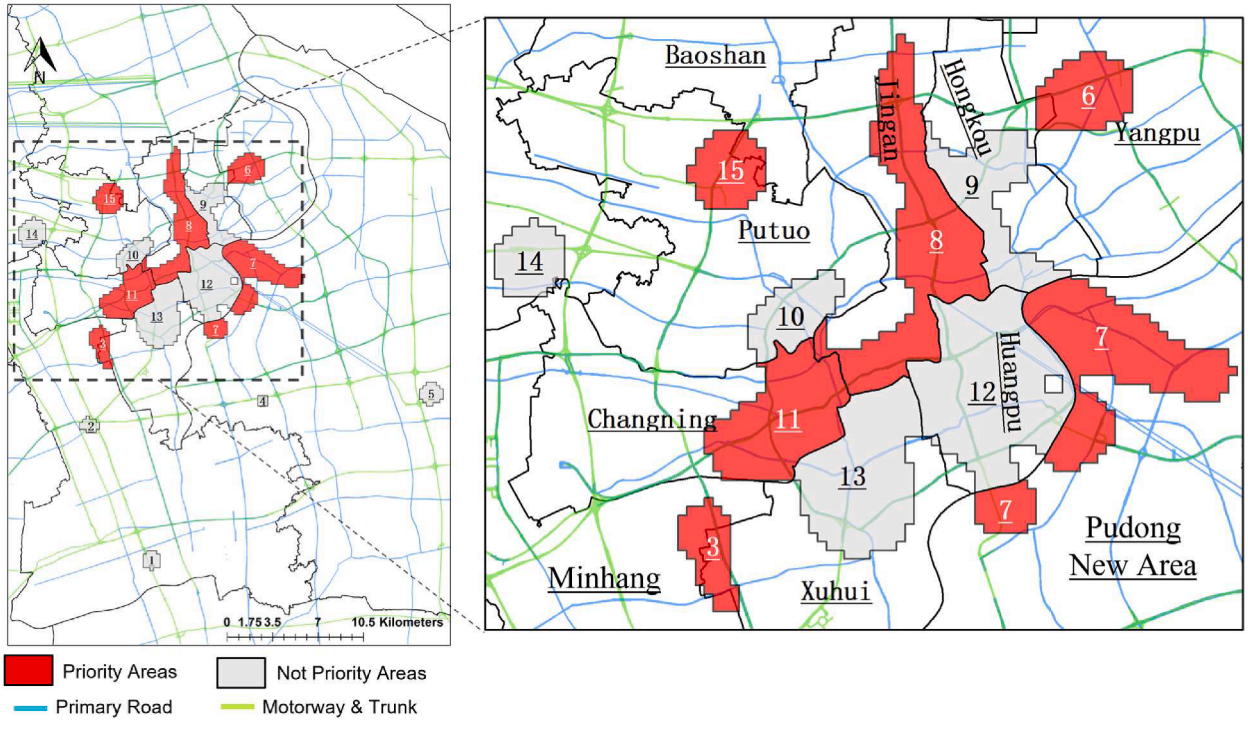


**Fig. 11.** Spatial Distribution of Priority and Non-priority Congestion-prone Areas.

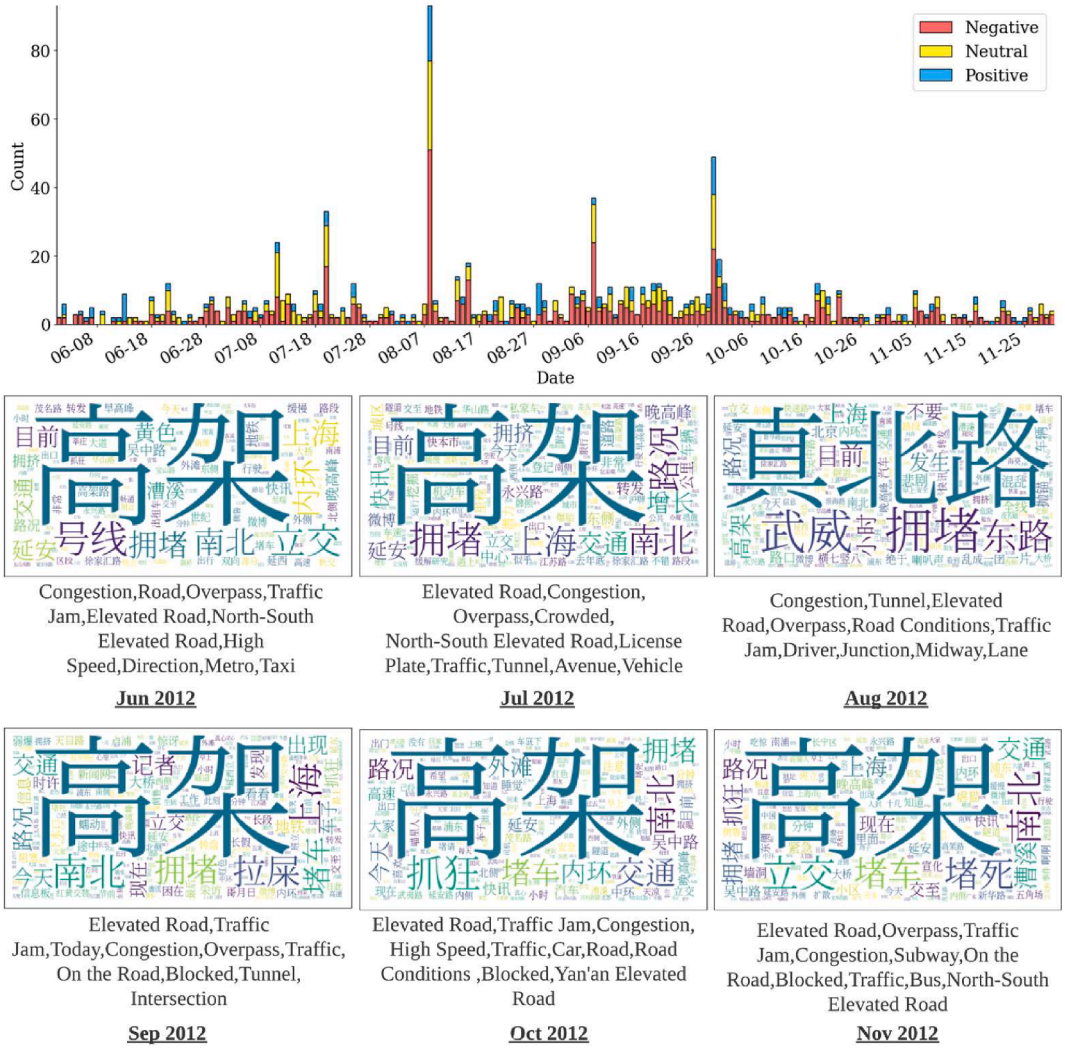


**Fig. 12.** Temporal and Semantic Characterization of the Prioritized Congestion-prone Areas.

always under heavy traffic pressure. It would be imperative to improve the traffic efficiency of Nanpu Bridge in the future traffic management in Shanghai.

## 6. Conclusions and future works

This study develops a framework to integrate a modified KDE analysis with traffic-relevant social media data to identify and characterize accident- and congestion-prone areas heatedly discussed in social media. Shanghai, a major city in China undergoing rapid transportation development, is used as a case study to demonstrate the applicability of the proposed system.

Current results show that sentiment information can help decrease the miss detection probability when detecting the accident-prone areas. However, the sentiment information does not help to detect the congestion-prone areas. High-density congestion zones based on congestion-relevant Weibos are more location-driven and mainly spatially distributed in the urban districts where traffic congestion always happens. However, the high-density accident zones are more event-driven and mainly spatially distributed in areas where severe accidents are frequently discussed in social media. Finally, based on the sentiment index and the number of accident- or congestion-related Weibos in each identified accident- and congestion-prone area, 4 out of 15 accident-prone areas and 6 out of 15 congestion-prone areas are regarded as the prioritized regions in the following traffic management. Both prioritized accident- and congestion-prone areas highlight the Northwest of Pudong New Area and Shanghai urban districts.

Unlike previous studies based on the sensor-based data, we developed a framework to find and profile the accident- and congestion-prone areas where people heatedly discussed in the social media platform. The advantages of the proposed framework are the following:

- The proposed framework identifies the accident- and congestion-prone areas from people's perspectives by analyzing the accident and congestion events reported in the social media messages, many of which are not recorded by the traffic survey data.
- By utilizing the time, text, and geo-information attached to the social media data, the proposed framework can characterize the identified accident- and congestion-prone areas and help the traffic authorities design countermeasures to improve the urban traffic experience.

- Compared to the methods based on survey data, the proposed framework has a lower maintenance cost in tracking the accidents and congestion in the city.

The results confirm that the proposed framework can help the traffic authorities improve the road environment and potentially reduce the number of accidents and congestion in the city. However, the developed framework also contains the following limitations.

- This developed framework might not work in areas with low social media penetration rates.
- The location information of the traffic accidents and traffic congestion reported by social media users is not as accurate as the location information collected by sensors or surveys.

In the future, we would like to put more effort into building more powerful traffic information detection modules, which can help the traffic authorities detect the traffic information with different consequences (e.g., with or without injuries, fatal traffic crashes or not). Moreover, the number of traffic-related social media data across the city can be an indicator to estimate people's attention toward locations regarding the traffic condition in the city. A safety model can be built between the traffic-related social media exposure, actual accident records, traffic volume, and other built environment variables to understand the emergence and contributing factors of the accident-prone areas.

## CRediT authorship contribution statement

**Haoliang Chang:** Methodology, Investigation, Software, Writing – original draft. **Lishuai Li:** Conceptualization, Validation, Resources. **Jianxiang Huang:** Writing – review & editing. **Qingpeng Zhang:** Data curation. **Kwai-Sang Chin:** Project administration.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This work was supported by the Hong Kong Research Grant Council Theme-Based Research Scheme (Grant No. T32-101/15-R) and the Hong Kong Research Grant Council General Research Fund (Project No. 11209717).

## Appendix A:. Traffic Weibo labeling procedure

We invited two human annotators to label 6,000 candidate traffic-relevant Weibos manually. The whole labeling process consisted of the following three rounds:

- Round 1: Each annotator labeled all the 6,000 candidate traffic-relevant Weibos.
- Round 2: The annotators picked the Weibos labeled differently and repeated the annotation process
- Round 3: The annotators selected Weibos still labeled differently. They discussed these Weibos together and made a final decision.

Table A1 shows some sampled Weibos, their English translation, and labels produced by the annotators. The location information of Weibos labeled as "Traffic-relevant and Having Location Information" has been highlighted.

**Table A1**
Labeled Samples of Candidate Traffic-relevant Weibos.

| Sample Weibo Content | English Translation | Label |
|---|---|---|
| 清明时节, 未必雨纷纷。但一定会堵在扫墓的路上。早上六点二十分, 已经堵到**虹桥路高架**了。还没出**上海**呢。 | During the Qingming Festival, it may not be rainy. But it will be stuck in the way of tomb sweeping. At 6:20 in the morning, it was blocked to **the elevated Hongqiao Road**. I have not left **Shanghai** yet | Traffic-relevant and Having Location Information |
| 经过班机延误、**北京**大堵车, 我终于到达了酒店。收发一下邮件就去祭五脏庙了。晚上继续勘察活动现场。 | After flight delays and traffic jams in **Beijing**, I finally arrived at the hotel. After sending and receiving emails, I worshiped the five internal organs temple. Continue to survey the activity site in the evening. | |
| 路上好堵… | The road is jammed | Traffic-relevant but not Having Location Information |
| 你也在现场啦~~~最近车祸好多咩~~~ | You are here too~~~ There have been many car accidents recently~~~ | |
| //@我们都震惊了: 世界上最惨烈的车祸镜头集合, 太惨烈, 太劲爆了啊! | @We are all shocked: The world's most tragic car accident lens collection, too tragic, too crazy! | Traffic-irrelevant |
| 那我算运气好的…正常上了二号线, 只是小挤……//@汪懿俊Milo: 难得早出门, 结果堵在2号线里, 各种汗味和劣质香水味[衰][衰][衰] | Then I am lucky… I got on line 2 normally, just a tiny squeeze…… It is rare to go out early, but I am stuck in line 2 with sweat and cheap perfume. | |

## Appendix B:. Training details of the CNN-LSTM module

To assign each word with a representation, we used the word embeddings released by S. Li et al., 2018 for the static Chinese word representation, pretrained from Chinese Sina Weibo data. The dimensionality of the word vector was set to 300. The word embeddings would first be used to map each word in Chinese Weibo to a dense representation. Then the CNN module was used to capture the contextual information around each word. The LSTM layer encoded the long-term dependencies of the word sequence. The output from the LSTM layer was then fed to a two-layer feedforward neural network for classification. The number of hidden units in the feedforward neural network was 100 and 3, respectively. To prevent overfitting, we further added dropout layers with a dropout rate of 0.5 between the convolutional and LSTM layers, between the LSTM layers, and the fully-connected feedforward layers (Srivastava et al., 2014).

When training the detection module, we first split the manually labeled Weibo dataset developed in Subsection 3.3.2 into training data, validation data, and test data, which account for 60%, 20%, and 20% of all the labeled Weibos. Then we tried different combinations of hyperparameter settings (presented in Table B1). For each model with a specific hyperparameter setting, we fed mini-batch data to the model. In backpropagation, we used the Adam Optimizer (Kingma and Ba, 2015) to update the parameters in the detection model. The number of epochs we set to train a detection model is 50. We would early stop the model if Categorical Cross-Entropy Loss did not decrease in 5 epochs based on the validation data (De Boer et al., 2005).

**Table B1**
Grid Search Hyperparameters for CNN-LSTM Model.

| Hyperparameter Settings | Considered Values |
|---|---|
| CNN Kernel Size | (2, 300), (3, 300) |
| Learning Rate in Adam Optimizer | [0.0001, 0.0002, 0.001] |
| Mini Batch Size | [16, 32, 64] |

## Appendix C:. Robustness checks of KDE modules

The results of the robustness check are shown in Fig. C1. The black dots show the *TP* and *FP* using threshold value defined in the Eq. (2). We find that the KDE module with the 300-meter unit size and considering the sentiment is slightly more robust than other modules in separating the spatial units containing actual accident records from others. Nevertheless, there is no significant difference between KDE modules with different parameter settings in distinguishing the spatial units containing actual congestion records from other units.

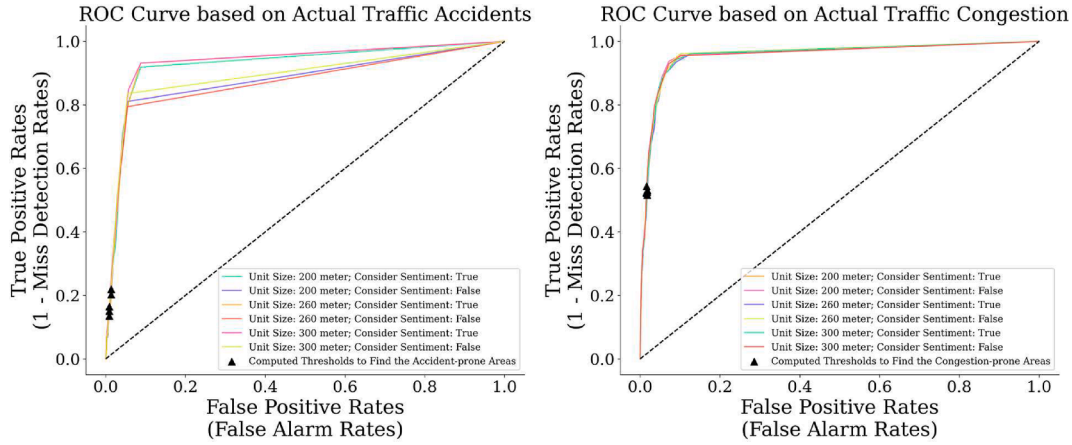


**Fig. C1.** Robustness Checks for KDE Modules Regarding Actual Accident and Congestion Records.

## References


Al-Aamri, A.K., Hornby, G., Zhang, L.-C., Al-Maniri, A.A., Padmadas, S.S., 2021. Mapping road traffic crash hotspots using GIS-based methods: A case study of Muscat Governorate in the Sultanate of Oman. Spat. Stat. 42, 100458. https://doi.org/10.1016/j.spasta.2020.100458.

Ali, F., Ali, A., Imran, M., Naqvi, R.A., Siddiqi, M.H., Kwak, K.-S., 2021. Traffic accident detection and condition analysis based on social networking data. Accid. Anal. Prev. 151, 105973. https://doi.org/10.1016/j.aap.2021.105973.

Ali, F., El-Sappagh, S., Kwak, D., 2019a. Fuzzy ontology and LSTM-based text mining: A transportation network monitoring system for assisting travel. Sensors (Switzerland) 19, 2. https://doi.org/10.3390/s19020234.

Ali, F., Kwak, D., Khan, P., El-Sappagh, S., Ali, A., Ullah, S., Kim, K.H., Kwak, K.S., 2019b. Transportation sentiment analysis using word embedding and ontology-based topic modeling. Knowledge-Based Syst. 174, 27–42. https://doi.org/10.1016/j.knosys.2019.02.033.

Ali, F., Kwak, D., Khan, P., Islam, S.M.R., Kim, K.H., Kwak, K.S., 2017. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling. Transp. Res. Part C Emerg. Technol. 77, 33–48. https://doi.org/10.1016/j.trc.2017.01.014.

Angel, S., Blei, A.M., Civco, D.L., Parent, J., 2012. Atlas of urban expansion. Lincoln Institute of Land Policy Cambridge, MA.

Bíl, M., Andrášik, R., Sedoník, J., 2019. A detailed spatiotemporal analysis of traffic crash hotspots. Appl. Geogr. 107 April, 82–90. https://doi.org/10.1016/j.apgeog.2019.04.008.

Bíl, M., Kubeček, J., Sedoník, J., Andrášik, R., 2017. Srazenazver.cz: A system for evidence of animal-vehicle collisions along transportation networks. Biol. Conserv. 213 February, 167–174. https://doi.org/10.1016/j.biocon.2017.07.012.

Cao, D., Wang, S., Lin, D., 2018. Chinese microblog users' sentiment-based traffic condition analysis. Soft Comput. 22 (21), 7005–7014. https://doi.org/10.1007/s00500-018-3293-8.

Cao, J., Zeng, K., Wang, H., Cheng, J., Qiao, F., Wen, D., Gao, Y., 2014. Web-based traffic sentiment analysis: Methods and applications. IEEE Trans. Intell. Transp. Syst. 15 (2), 844–853. https://doi.org/10.1109/TITS.2013.2291241.

Casas, I., Delmelle, E.C., 2017. Tweeting about public transit — Gleaning public perceptions from a social media microblog. Case Stud. Transp. Policy 5 (4), 634–642. https://doi.org/10.1016/j.cstp.2017.08.004.
Chainey, S., Tompson, L., Uhlig, S., 2008. The utility of hotspot mapping for predicting spatial patterns of crime. Secur. J. 21 (1–2), 4–28. https://doi.org/10.1057/palgrave.sj.8350066.
Chang, Z., Murakami, J., 2019. Transferring land-use rights with transportation infrastructure extensions: Evidence on spatiotemporal price formation in Shanghai. J. Transp. Land Use 12 (1), 1–19. https://doi.org/10.5198/jtlu.2019.1357.
Chen, Y., Lv, Y., Wang, X., Li, L., Wang, F.-Y., 2019. Detecting Traffic Information From Social Media Texts With Deep Learning Approaches. IEEE Trans. Intell. Transp. Syst. 20 (8), 3049–3058. https://doi.org/10.1109/TITS.2018.2871269.
Cortes, C., Vapnik, V., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297.
Cottrill, C., Gault, P., Yeboah, G., Nelson, J.D., Anable, J., Budd, T., 2017. Tweeting Transit: An examination of social media strategies for transport information management during a large event. Transp. Res. Part C Emerg. Technol. 77, 421–432. https://doi.org/10.1016/j.trc.2017.02.008.
D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F., 2015. Real-Time Detection of Traffic from Twitter Stream Analysis. IEEE Trans. Intell. Transp. Syst. 16 (4), 2269–2283. https://doi.org/10.1109/TITS.2015.2404431.
Dabiri, S., Heaslip, K., 2019. Developing a Twitter-based traffic event detection model using deep learning architectures. Expert Syst. Appl. 118, 425–439. https://doi.org/10.1016/j.eswa.2018.10.017.
de Boer, P.-T., Kroese, D.P., Mannor, S., Rubinstein, R.Y., 2005. A tutorial on the cross-entropy method. Ann. Oper. Res. 134 (1), 19–67.
Durahim, A.O., Coşkun, M., 2015. #iamhappybecause: Gross National Happiness through Twitter analysis and big data. Technol. Forecast. Soc. Change 99, 92–105. https://doi.org/10.1016/j.techfore.2015.06.035.
ESRI, 2017. ArcGIS 10.4. 1 software.
GaryBikini, 2020. GaryBikini/ChinaAdminDivisonSHP: ChinaAdminDivisonSHP v1.1. doi:10.5281/ZENODO.4167299.
Haghighi, N.N., Liu, X.C., Wei, R., Li, W., Shao, H., 2018. Using Twitter data for transit performance assessment: a framework for evaluating transit riders' opinions about quality of service. Public Transp. 10 (2), 363–377. https://doi.org/10.1007/s12469-018-0184-4.
Haklay, M., Weber, P., 2008. Openstreetmap: User-generated street maps. IEEE Pervasive Comput. 7 (4), 12–18.
Han, F., Xie, R., Lai, M., 2018. Traffic density, congestion externalities, and urbanization in China. Spat. Econ. Anal. 13 (4), 400–421. https://doi.org/10.1080/17421772.2018.1459045.
Harirforoush, H., Bellalite, L., 2019. A new integrated GIS-based analysis to detect hotspots: A case study of the city of Sherbrooke. Accid. Anal. Prev. 130, 62–74. https://doi.org/10.1016/j.aap.2016.08.015.
Holmgren, J., Knapen, L., Olsson, V., Masud, A.P., 2020. On the use of clustering analysis for identification of unsafe places in an urban traffic network. Procedia Comput. Sci. 170 (2019), 187–194. https://doi.org/10.1016/j.procs.2020.03.024.
Kim, Y., 2014. Convolutional Neural Networks for Sentence Classification, in. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. https://doi.org/10.3115/v1/D14-1181.
Kingma, Diederik P and Ba, J., 2015. Adam: A method for stochastic optimization, in: 3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.
Lansley, G., Longley, P.A., 2016. The geography of Twitter topics in London. Comput. Environ. Urban Syst. 58, 85–96. https://doi.org/10.1016/j.compenvurbsys.2016.04.002.
LeCun, Y., Bengio, Y., et al., 1995. Convolutional networks for images, speech, and time series. Handb. Brain Theory Neural Netw. 3361, 1995.
Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., Du, X., 2018. Analogical reasoning on Chinese morphological and semantic relations. In: ACL 2018–56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), pp. 138–143. https://doi.org/10.18653/v1/p18-2023.
Liu, Y., Zhou, Y., Liu, G., 2012. Chinese government use of social media: A case of Shanghai Weibo @Shanghaicity. IEEE Int. Conf. Digit. Ecosyst. Technol. 1–5 https://doi.org/10.1109/DEST.2012.6227946.
Luo, S., He, S.Y., 2021. Using data mining to explore the spatial and temporal dynamics of perceptions of metro services in China: The case of Shenzhen. Environ. Plan. B Urban Anal. City Sci. 48 (3), 449–466. https://doi.org/10.1177/2399808320974693.
Mihalcea, R.d., Tarau, P., 2004. Textrank: Bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 1–9.
Millward, S., 2012. China's forgotten 3rd Twitter clone hits 260 million users [WWW Document]. URL http://www.techinasia.com/netease-weibo-260-million-users-numbers.
National Bureau of Statistics of China, 2019. China Statistical Yearbook.
Ni, M., He, Q., Gao, J., 2014. Using Social Media to Predict Traffic Flow under Special Event Conditions. Meet. Transp. Res. Board. 716, 23p.
Ouni, F., Belloumi, M., 2018. Spatio-temporal pattern of vulnerable road user's collisions hot spots and related risk factors for injury severity in Tunisia. Transp. Res. Part F: Traffic Psychol. Behav. 56, 477–495. https://doi.org/10.1016/j.trf.2018.05.003.
Powers, D.M.W., 2020. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. ArXiv abs/2010.1.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Serna, A., Gerrikagoitia, J.K., Bernabé, U., Ruiz, T., 2017. Sustainability analysis on Urban Mobility based on Social Media content. Transp. Res. Procedia 24, 1–8. https://doi.org/10.1016/j.trpro.2017.05.059.
Shanghai Municipal Bureau of Statistics, 2015. Shanghai Statistical Yearbook 2015 [WWW Document]. URL http://tjj.sh.gov.cn/tjnj/zgsh/tjnj2015en.html.
Shanghai Urban and Rural Construction and Traffic Committee, 2010. Shanghai Fourth Comprehensive Traffic Survey Report.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958.
Tao, L., Zhu, D., Yan, L., Zhang, P., 2015. The traffic accident hotspot prediction: Based on the logistic regression method. ICTIS 2015–3rd Int. Conf. Transp. Inf. Safety, Proc. 107–110 https://doi.org/10.1109/ICTIS.2015.7232194.
Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., Wang, H., Wu, F., 2020. In: SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis, in, pp. 4067–4076.
Wang, D., Al-Rubaie, A., Davies, J., Clarke, S.S., 2014. Real time road traffic monitoring alert based on incremental learning from tweets. IEEE SSCI 2014 - 2014 IEEE Symp. Ser. Comput. Intell. - EALS 2014 2014 IEEE Symp. Evol. Auton. Learn. Syst. Proc. 50–57. doi:10.1109/EALS.2014.7009503.
Wang, D., Krebs, E., Nickenig Vissoci, J.R., de Andrade, L., Rulisa, S., Staton, C.A., 2020. Built Environment Analysis for Road Traffic Crash Hotspots in Kigali, Rwanda. Front. Sustain. Cities 2 June , 1–13. doi:10.3389/frsc.2020.00017.
Wu, D., Cui, Y., 2018. Disaster early warning and damage assessment analysis using social media data and geo-location information. Decision Support Systems 111, 48–59.
Xia, Li, Chen, Liao, 2019. Identify and Delimitate Urban Hotspot Areas Using a Network-Based Spatiotemporal Field Clustering Method. ISPRS Int. J. Geo-Information 8 (8), 344. https://doi.org/10.3390/ijgi8080344.
Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of Traffic Crashes Involving Pedestrians Using Big Data: Investigation of Contributing Factors and Identification of Hotspots. Risk Anal. 37 (8), 1459–1476.
Xie, Z., Yan, J., 2008. Kernel Density Estimation of traffic accidents in a network space. Comput. Environ. Urban Syst. 32 5 , 396–406. doi:10.1016/j.compenvurbsys.2008.05.001.
Xu, Q., Tao, G., 2018. Traffic accident hotspots identification based on clustering ensemble model. Proc. - 5th IEEE Int. Conf. Cyber Secur. Cloud Comput. 4th IEEE Int. Conf. Edge Comput. Scalable Cloud, CSCloud/EdgeCom 2018, 1–4. https://doi.org/10.1109/CSCloud/EdgeCom.2018.00010.
Xu, S., Li, S., Wen, R., 2018. Sensing and detecting traffic events using geosocial media data: A review. Comput. Environ. Urban Syst. 72 June, 146–160. https://doi.org/10.1016/j.compenvurbsys.2018.06.006.
Yao, S., Wang, J., Fang, L., Wu, J., 2018. Identification of vehicle-pedestrian collision hotspots at the micro-level using network kernel density estimation and random forests: A case study in Shanghai, China. Sustainability 10, 12. https://doi.org/10.3390/su10124762.