sensors

MDPI

Article

Sensing Urban Transportation Events from Multi-Channel Social Signals with the Word2vec Fusion Model

Hao Lu [1,2,†], Kaize Shi [1,†], Yifan Zhu [1], Yisheng Lv [2,*] and Zhendong Niu [1,*]

[1] School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China; hao.lu@ia.ac.cn (H.L.); kzshi@bit.edu.cn (K.S.); zhuyifan@bit.edu.cn (Y.Z.)

[2] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academic of Science, Beijing 100190, China

* Correspondence: yisheng.lv@ia.ac.cn (Y.L.); zniu@bit.edu.cn (Z.N.)

† These authors contributed equally to this work.

Received: 20 September 2018; Accepted: 19 November 2018; Published: 22 November 2018

check for updates

Abstract: Social sensors perceive the real world through social media and online web services, which have the advantages of low cost and large coverage over traditional physical sensors. In intelligent transportation researches, sensing and analyzing such social signals provide a new path to monitor, control and optimize transportation systems. However, current research is largely focused on using single channel online social signals to extract and sense traffic information. Clearly, sensing and exploiting multi-channel social signals could effectively provide deeper understanding of traffic incidents. In this paper, we utilize cross-platform online data, i.e., Sina Weibo and News, as multi-channel social signals, then we propose a word2vec-based event fusion (WBEF) model for sensing, detecting, representing, linking and fusing urban traffic incidents. Thus, each traffic incident can be comprehensively described from multiple aspects, and finally the whole picture of unban traffic events can be obtained and visualized. The proposed WBEF architecture was trained by about 1.15 million multi-channel online data from Qingdao (a coastal city in China), and the experiments show our method surpasses the baseline model, achieving an 88.1% F_1 score in urban traffic incident detection. The model also demonstrates its effectiveness in the open scenario test.

Keywords: intelligent sensors; social transportation; multi-channel signals; event detection; word2vec-based event fusion

1. Introduction

Intelligent transportation systems (ITS) are highly involved in improving transportation efficiency and services [1,2]. Successful operation of ITS relies on multi-modal big data. To collect traffic data, physical sensors like inductive loops, radars, and cameras, are deployed in real world transportation systems [3,4]. However, such conventional physical sensors are expensive and provide limited coverage of transportation networks.

Today, social media and online web services provide a new path to access traffic information through the internet. Due to the widespread applications of smart devices and social networks, people can create and diffuse user generated contents (UGC) anywhere at any time. Therefore, everyone in such a network can be a social sensor to perceive the real world, which makes the collection of social signals in multiple domains possible [5,6]. Compared with traditional physical traffic sensors, social traffic sensors offer advantages of costing essentially nothing and large-scale coverage [7]. Social media data have been applied to detect traffic events, explain traffic status, analyze traffic sentiment, etc. [8–10].

Twitter, Facebook, Sina Weibo, News, BBS, and official transportation administration websites are the main data sources to generate social signals in transportation research [11–22]. However, current research focuses on using single channel social signals to extract traffic information. Clearly, exploiting multi-channel social signals from cross-platform online data could effectively help gain a deeper understanding of transportation events. Such multi-channel social signals provide significant opportunities, but are also accompanied by serious challenges.

In this paper, we utilize multi-channel social signals from cross-platform online data for more accurate detection and more comprehensive description of urban traffic. The main challenge is how to semantically represent and fuse the multi-channel social signals into traffic event descriptions. We firstly detect traffic events like traffic congestion, traffic incidents and traffic accidents from the News and Weibo platforms, respectively. Then we propose a word2vec-based event fusion (WBEF) model to semantically fuse the same topics detected from cross-platform data into overall traffic events. The proposed WBEF model is trained with two years' data collected from Weibo and News platforms. We apply the proposed methods to Qingdao, a city on China's eastern coast. Experiments demonstrate the effectiveness of the proposed methods in traffic event detection and fusion from cross-platform online data.

The main contributions of this paper are as follows: (1) We exploited cross-platform online data as multi-channel social streams to mine traffic events, which benefits from the strengths of News media's objectivity and social media's timeliness. (2) We proposed a word2vec-based event fusion (WBEF) model to fuse traffic events from cross-platform data semantically, and the WBEF model shows better recall and F_1 score compared to baseline models in the experiments. (3) We deployed the proposed methods in an open scenario, obtained a complete picture of the city traffic situation, and qualitatively investigated the application efforts of proposed model.

The rest of this paper is organized as follows: Section 2 reviews related work from three perspectives which are social transportation, topic modeling and cross-platform event detection. Section 3 presents the details of the WBEF model to detect and fuse urban traffic events from cross-platform data. Section 4 presents details of our experiments. Section 5 demonstrates a case study in Qingdao. Section 6 concludes this paper.

2. Related Work

Social media and online web services generate social signals via crowdsourcing technologies supported by smart mobile devices. Mining such social signals for traffic management and control has received increasing attention, and the number of publications in this field has grown quickly in recent years [23–25]. Typical techniques used in current social sensor-based transportation studies include natural language processing, social computing, machine learning, recommendation systems and expert systems. To our knowledge, this is the first attempt to apply the cross-platform event detection and fusion model to multi-channel social signals for urban traffic situation monitoring. In this section, we review social mediabased transportation research, and summarize topic modeling and cross-platform event fusion methods.

2.1. Social Media Based Transportation Research

Social mediabased transportation research mainly falls into five categories, i.e., traffic prediction, traffic event detection, traffic sentiment, traffic reasoning, and traffic behavior analysis. Traffic prediction has a long research history. As early as the 1970s, Ahmed and Cook applied the autoregressive integrated moving average (ARIMA) model to predict short-term freeway flow [26]. Recently, social media data were used to improve traffic prediction. He et al. [27] developed a linear regression model incorporating real traffic data and Twitter data for longer-term traffic flow prediction. Ni et al. [16] showed a moderate positive correlation between subway passenger flow and Twitter post rates and proposed a hybrid model to fuse linear regression and SARIMA models to forecast subway passenger flow. Chaniotakis et al. [19] used social media data to analyze and visualize Twitter signals at both temporal and spatial levels, which showed there are indicative group mobility patterns and behavioral characteristics in urban transportation. Zeng et al. [11] studied online transportation-related topic features on the national holiday of China, from the perspectives of topic evolution analysis, opinion analysis, and geographic analysis, which could potentially help administrative sectors for traffic management.

Traffic event sensing and detection is one of the major tasks in social mediabased transportation research and applications, and many scholars have proposed inspiring models. D'Andrea, et al. [13] treated traffic event detection from Twitter as a binary classification task, which assigns traffic/non-traffic class labels to each tweet. They compared seven different classification models, such as SVM, NB, C4.5, KNN, and employed the SVM model in their proposed system since SVM achieved the best accuracy value in their tests. Gu et al. [28] defined five traffic incident categories and extracted traffic incident information on highways and arterial roads from tweet texts and they firstly used the Semi-Naive-Bayes classifier to categorize traffic incident tweets and non-traffic incident tweets, and then they trained the Supervised Latent Dirichlet Allocation classifier to identify traffic incident categories. Fu et al. [29] proposed an association rule-based keyword generation scheme to iteratively extract real time transportation incidents. Meanwhile they implemented LexRank algorithms on the complete sentence graph to rank the most influencing node, and the ranked words are regarded as the summarization of traffic incidents. Gutierrez et al. [30] presented a computational framework to detect real-time traffic events in the UK from Twitter, including tweet filtering, event type classification, name entity recognition, geo-location extraction and event tracking. Nguyen et al. [15] built a system named TrafficWatch that leveraged twitter signals and integrated them with online clustering and classification algorithms for traffic monitoring and event detection. TrafficWatch demonstrated the potential to report traffic incidents earlier than other data sources when deployed in the traffic management center of Australia. Hao et al. [31], mined the correlation between adverse weather topic heat and traffic incidents in social media, and further proposed traffic situation awareness and alerting model assisted by adverse weather data to provide information on city-level traffic situations.

2.2. Topic Modeling

Topic modeling tries to uncover the hidden semantic structures of different types of documents. Topic modeling technologies can obtain traffic topics from online social media and news texts. In recent years, Latent Dirichlet Allocation (LDA) [32] and its extension models have become dominant for topic modeling. Zhai et al. [33] proposed an online topic model by extending LDA to draw topics from a Dirichlet process whose base distribution is over all possible words rather than from a finite Dirichlet distribution. They also develop an online variational inference method to heuristically expand the set of words in vocabulary. Paisley et al. [34] proposed a nested hierarchical Dirichlet process for hierarchical topic modeling, and developed a stochastic variational inference algorithm for the model. The proposed method was tested on 1.8 million documents from The New York Times and 2.7 million documents from Wikipedia.

LDA and related models are traditionally applied to long text documents like news articles, while there are increasing needs for modeling topics in short text documents like Twitter posts. Quan et al. [35] integrated topic modeling with automatic short text aggregation to alleviate the sparsity problem in short and sparse texts. Their experiments indicate the proposed scheme can extract more meaningful and interpretable topics than traditional topic models. Ramage et al. [36] developed a partially supervised LDA model which labels the content of twitter posts with four characteristics regarding substance, style, status and social relationship. The experiments indicated weighted combination of L-LDA model's latent topic features and TF-IDF feature achieve satisfactory results for topic ranking and recommendation task. Zhao et al. [37] assumed that each tweet has been associated with a Twitter topic, and each user has its own topic distribution. Based on such a hypothesis, they proposed a twitter LDA generation process, and used Gibbs sampling to perform model inference. Experiments demonstrated the Twitter-LDA can get more meaningful topic words than standard LDA models.

2.3. Cross-Platform Event Detection

Multi-channel social signals processing benefits from the characteristics of multi-modality and multi-domain, which integrate different kinds of information from different sources to obtain a more comprehensive view ("big picture") of objects compared to a single data stream. Many models with cross-platform data can produce better event prediction and detection results. Hou et al. [38] developed a cross-dependence temporal topic model to extract topics, and studied the mutual influence between news and user-generated content streams. The proposed methods were evaluated on five datasets from Sina, The New York Times and Twitter. Oghina et al. [39] used tweets from Twitter and comments from YouTube to predict IMDb movie ratings, and the best performance model could rate movies close to the observed values. Bao et al. [40] used a co-clustering model to detect emerging topics from The New York Times and Flickr which experimentally achieved effective evaluation results. Daichi et al. [41] applied a time series topic detection model to mix news and twitter streams during the London Olympic games, which detected 34 topics with a precision of 87.5%.

Some cross-platform event detection architectures and systems have been proposed. Qian et al. [42] proposed a generic framework for social event detection, tracking and evolution analysis. They developed specific models for each task. For example, they used a boosted multi-modal supervised LDA model for social event detection, and applied an incremental topic model learning algorithm for analyzing the evolutionary processes of social events. Li et al. [43] presented the Event Knowledge from News and Opinions in Twitter (EKNOT) system which could extract summaries combining an objective description from news and opinions from tweets. They used an entity graph to link entities for an event, also used an opinion graph to get a joint summarization of an event. Wang et al. [44] proposed an event-based multi-aspect reflection mining framework to discover, link and present major events. News and tweets about a major event can complement each other to describe the event.

3. Methodology

Social signals effectively connect the physical space and cyber space, which provides a new paradigm for traffic situation awareness. The main challenges that need to be faced are how to collect, process, analyze, and fuse several types of signals in social transportation. As mentioned in the related works, social transportation has attracted increasing numbers of researchers, and some frameworks/architectures were proposed for specific social transportation tasks, such as TrafficWatch [15] for traffic incident extraction, STAR-CITY [2] for traffic flow analysis and urban planning, Steds [29] for traffic event summarization, Hao's framework [31] for weather-related traffic incidents perception. However, these architectures only leverage single social signals either from news or social media.

Meanwhile, the frameworks for multi-modal data sensing and fusion are also getting more and more attention, with representative event detection frameworks like EKNOT [43], and the work of Wang et al. [44] and Qian et al. [42], all of which combine objective descriptions from news and opinions from tweets together, linking the event descriptions and reflection from a cross-platform with the entity graphs, finally fused into a joint summary of events. Although these frameworks recognized that fused multi-channel social signals will improve the accuracy of event detection and the diversity of event description, the entities (words representing person, location, organization, etc.) network-based event fusion methods are unable to deeply fuse the event description in a semantic way. Moreover, it is very necessary to integrate the traffic domain knowledge when applying the frameworks above to intelligent transportation systems.

To our knowledge, this is the first attempt to transfer the traffic event detection task from single channel social signals to multi-channel social signals, the main challenge being how to sense, process, analyze, link and fuse several types of signals in social transportation. To address these problems, it is necessary to integrate natural language processing, information retrieval and machine learning methods together with transportation domain knowledge to utilize the architecture.

In this paper, by sufficiently utilizing the objectivity of news media and immediacy of social media, a word2vec-based event fusion (WBEF) model is proposed for the urban transportation event detection, which extracts topics from multi-channel social signals, semantically coupled and fused topics into cross-platform urban traffic events description. Furthermore, we develop a cross-platform traffic event detection system integrating the above methods for real world applications.

The system architecture is shown in Figure 1. We choose news articles and Weibo posts as our cross-platform data sources. The multi-channel social signals from cross-platform online data are sensed through a keyword-based social sensor network configured by domain experts. The sensed webpages are filtered and decomposed into structured data, then aggregated into data blocks assigned to city roads. Traffic event topics from news articles are extracted with a news LDA model, and traffic event topics from Weibo posts are extracted with a Weibo specific model named w-LDA model. Furthermore, topic words are transformed into semantic representation with word vectors, and then the cross-platform traffic events can be fused based on the topic distance matrix semantically. The expressions and descriptions of variables in the following algorithms or sub-models are summarized in Appendix A.

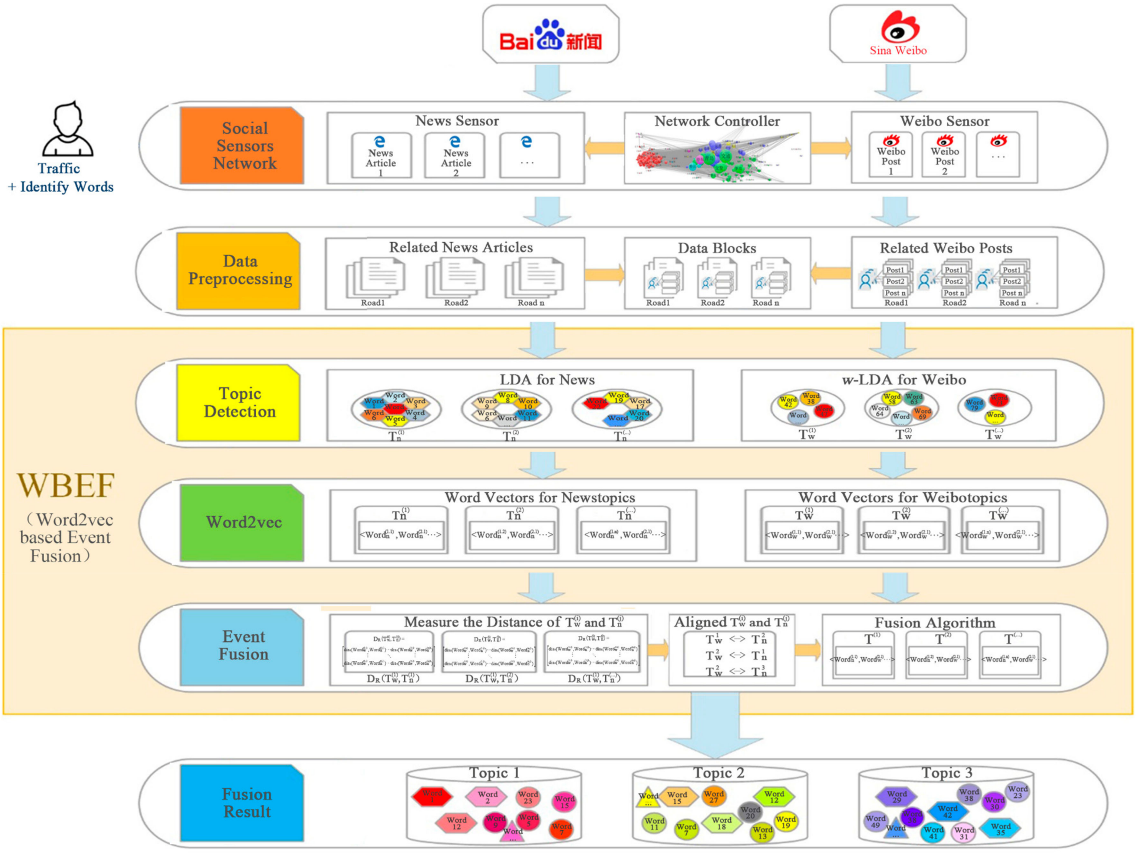
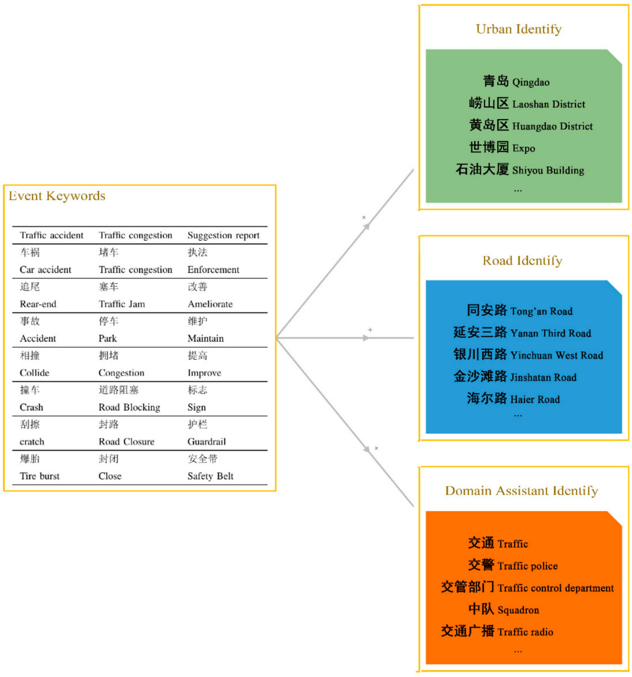


Figure 1. Architecture of cross-platform traffic event detection system.

3.1. Social Sensors Network

The social signals can be typically sensed by two approaches. The first approach is using API services provided by the platform to parse the XML or JSON file, such as Twitter and Weibo. The second approach is by deploying a web crawler that periodically monitors keywords, accounts and URL lists. In our system, we take both approaches according to the availability and limits of service providers. By consulting experts in traffic administrative agencies, the traffic keywords are grouped into four categories which are traffic event keywords, urban identity keywords, road identity keywords and domain assistant identity keywords, respectively. Specifically, traffic event keywords are mainly exploited to describe three types of traffic events, which are traffic accidents, traffic jams and traffic suggestions. The identity keywords (Figure 2) describe corresponding cities, roads, and domain assistants. All above traffic keywords are treated as search seeds that are feeding into the social sensor network.



Traffic accident	Traffic congestion	Suggestion report
车祸 Car accident	堵车 Traffic congestion	执法 Enforcement
追尾 Rear-end	塞车 Traffic Jam	改善 Ameliorate
事故 Accident	停车 Park	维护 Maintain
相撞 Collide	拥堵 Congestion	提高 Improve
撞车 Crash	道路阻塞 Road Blocking	标志 Sign
刮擦 cratch	封路 Road Closure	护栏 Guardrail
爆胎 Tire burst	封闭 Close	安全带 Safety Belt

Figure 2. Traffic keywords for social sensors.

The News sensor perceived the latest news data according one specific keyword from the News search engine. Correspondingly, the Weibo sensor also perceived latest Weibo data according one specific keyword from the Weibo search engine. The News and Weibo sensors were deployed to continuously monitor the data according to a keywords list, and the links of different sensors were built if both sensors acquired the same News or Weibo article. Then, all the sensors and corresponding links eventually formed a social sensors network, which dynamically adapted to perceiving multi-channel social signals.

The keywords-based social sensors network consists of a crawler, page parser, duplicated URLs filter, etc. [45]. Particularly, to efficiently utilize the network bandwidth and computation resources, the social sensors network has a keywords priority adaptor, which dynamically sorts the keyword query priorities according to the value of importance ranking for each node in the keyword network.

The keywords-based social sensors network is visualized in Figure 3. The nodes denote traffic keywords. The size of each nodes represents the number of webpages that are collected regarding the node's keywords. The edges are constructed by calculating the co-occurrence of keyword pairs in the same document. All the search keywords are aggregated into network clusters, the main clusters include the roads cluster, the traffic incidents cluster and traffic suggestion keywords. Since the experiment is based on a Chinese corpus, here we only show the social sensors network with Chinese nodes and annotate the important information in English. The social sensors network can also be analyzed by a social network algorithm, such as ranking the betweenness, closeness and degree.

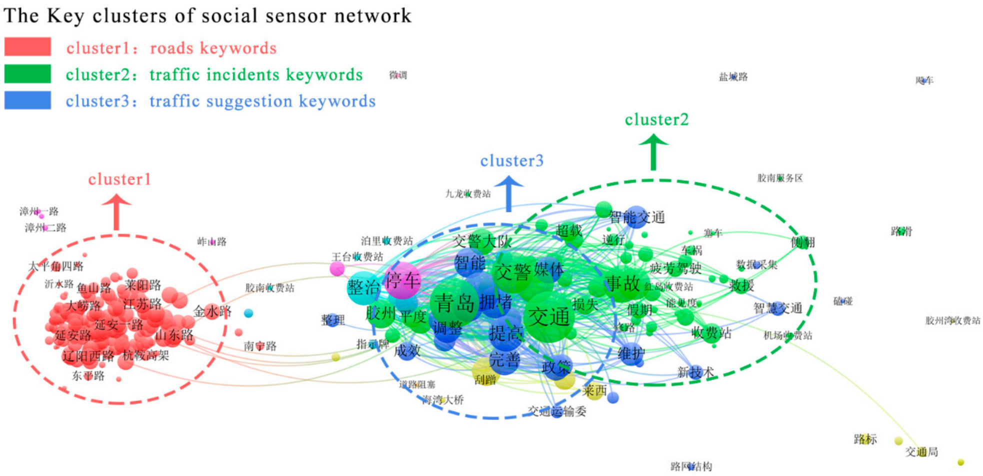


Figure 3. The keywords-based social sensors network.

3.2. Data Preprocessing

Data preprocessing is a fundamental step for traffic event sensing from cross-platform media because raw data are unstructured and full of noise. The major preprocessing steps include meta-data extraction, noise filtering, word segmentation and data blocks aggregation:

Noise filtering: The URL links, paragraph marks, emoji, etc. in the texts are regarded as noise information, which negatively influence the accuracy and efficiency of word segmentation, information processing and model training. To solve the problems, we define regular expressions to represent various noise patterns, and then use the regex to search and delete noises in the text. Meanwhile, either too short or too long texts are removed.

Meta-data extraction: we deployed Dom wrapper and XPath parsers to extract the title, post time and other meta-data like authors, review number and repost number, etc. from the sensed web page cache. Then the meta-data information is stored in databases.

Word segmentation: The space is a natural word delimiter in English texts, however, there is no such equivalent in Chinese, so word segmentation is needed for Chinese NLP tasks. The Language Technology Platform (LTP) [46] was deploy to segment words, remove punctuation and stop words, and tagging the road entity with customized dictionary. The customized dictionary for LTP includes all the keywords we used for the social sensors.

Data blocks aggregation: The sensed multi-channel social signals are aggregated into data blocks, which are defined as a dataset containing cross-platform online data related to every urban road. With the LTP's entity recognition tools, the road entity of every News article or Weibo post is extracted, then the cross-platform data are aggregated into data blocks corresponding to the road entity. Moreover, when news articles or Weibo posts contain multiple road entities they will be assigned to each corresponding data block separately. Each data block was fed into the event detection and fusion models iteratively.

3.3. Word2vec Based Event Fusion Model

News articles usually describe a traffic event with relatively standard language, while Twitter and Weibo may have posts expressing opinions/comments/discussions on the same traffic event [47,48]. Matching news articles and Twitter/Weibo posts on the same transportation topic can give a more comprehensive description of a traffic event. Consequently, we propose to use LDA-based models to extract topics in each respective source channels and link them together with the WBEF model.

3.3.1. Transportation Events Detection from News and Weibo

As described in the literature review section, the LDA and its variants have been widely applied for event detection. To group transportation topics and find topic words from News, we use the standard LDA model for event detection. While for Weibo, we detect traffic events with a *w*-LDA model which is described in details as follows:

w-LDA model: The *w*-LDA model is based on the USER scheme which achieves good performances in Twitter classification [49]. The process is described as follows:

a) Combine all training messages generated by the same user into user profiles;
b) Train the *w*-LDA model with training user profiles;
c) Aggregate all testing messages generated by the same user into testing user profiles;
d) Use the trained *w*-LDA model to infer a topic mixture.

The aggregated user profiles can be viewed as a random mixture distribution over latent topics, where each topic is characterized by a distribution over words. Both distributions are assumed to have a sparse Dirichlet prior. Suppose the corpus consists of T Weibo posts and U users that are aggregated into P user profiles, each user profile p contains N_p words. The total number of topics denoted as K, the unique words in vocabulary are denoted as V.

There are five latent variables and one observable variable, where latent variable α is a K-dimensional vector giving uniform prior weight for all topics in a user profile p, latent variable β is a V-dimensional vector with uniform prior weights for all words in a topic k, latent variable z_i is the topic for i-th word in user profile p and observable variable w_i is the specific word, latent variable φ_z is a V-dimensional vector representing the Dirichlet topic distribution for user profiles, latent variable ϑ_p is a K-dimensional vector representing the Dirichlet word distribution for topics. The variable z_i and w_i are drawn from multinomial distributions. The generative process can be seen in Algorithm 1.

Algorithm 1. *w*-LDA Generation Process

```
Input:$K$, $U$, $T$, $\alpha$, $\beta$
Output: $\varphi_z$, $\vartheta_p$, uProfileSet
//Step 1: User profiles data generation
For each author $u$ = 1 ... $U$
    Traverse all the Weibo posts $t$ = 1 ... $T$;
Aggregate the posts generated by author $u$ into users' profile s uProfileSet(p)
Return uProfileSet
//Step 2: Topics generation
For each topic $z$ = 1 ... $K$
    Draw $\varphi_z$ ~ Dirichlet ($\beta$)// sample mix components for topic-word
//Step 3: Topic words generation
For each user profile $p$ = 1 ... $P$
    Draw $\vartheta_p$ ~ Dirichlet ($\alpha$)// sample mix components for user profile-topic
    For each word in generated user profile $p$, $i$ = 1 ... $N_p$
        Draw $z_i$ ~ Multinomial ($\vartheta_p$) // Sample topic $z_i$ for user profile $p$
        Draw $w_i$~ Multinomial ($\varphi_{z_i}$) //Sample word $w_i$ for topic $z_i$
Return $\varphi_z$ and $\vartheta_p$
```

The *w*-LDA model focuses on finding out topics for each user profile, and we use collapsed Gibbs sampling to inference the final goal that is to approximate the distribution of P $(z_i = j|\, Z_{-i}, w_i, p_i)$, which is:

$$P(z_i = j|\, \mathbf{Z}_{-i}, w_i, p_i) \propto \frac{n_{-i,j}^{(w_i)} + \beta}{\sum_{i=1}^{v} n_{i,j}^{(w_i)} + \mathrm{V}\beta} \times \frac{n_{-i,j}^{(p_i)} + \alpha}{\sum_{j=1}^{K} n_{i,j}^{(p_i)} + \mathrm{K}\alpha} \tag{1}$$

where P $(z_i = j|\, \mathbf{Z}_{-i}, w_i, p_i)$ denotes as the probability that word w_i is assigned to topic j, $\mathbf{Z}_{-i}$ represents topic j assigned to all other words, w_i represents the i-th word in the vocabulary, p_i represents the user profile containing the word w_i. Then $n_{-i,j}^{(w_i)}$ is the number of times all word $\mathbf{W}_{-i}$ assigned to the topic j excluding the current word w_i, $\sum_{i=1}^{v} n_{i,j}^{(w_i)}$ calculates the total number of all words $\boldsymbol{W}$ assigned to the topic j. $n_{-i,j}^{(p_i)}$ is the number of times topic j is assigned to words in the user profile $\boldsymbol{P}_{-\boldsymbol{i}}$ excluding the current user profile p_i, $\sum_{j=1}^{K} n_{i,j}^{(p_i)}$ calculates the total number of all words in user profile p_i.

$\boldsymbol{\varphi_z}$ represents the predictive distributions of words in topic z, $\boldsymbol{\vartheta_p}$ represents the predictive distributions of topics in user profile p. For any obtained sample we can estimate $\varphi_z^{(j)}$ and $\vartheta_p^{(j)}$ by:

$$\varphi_z^{(j)} = \frac{n_{i,j}^{(w_i)} + \beta}{\sum_{i=1}^{v} n_{i,j}^{(w_i)} + V\beta}, \ \vartheta_p^{(j)} = \frac{n_{i,j}^{(p_i)} + \alpha}{\sum_{j=1}^{K} n_{i,j}^{(p_i)} + K\alpha} \tag{2}$$

where $n_{i,j}^{(w_i)}$ is number of times that word w_i has been assigned to topic j, $n_{i,j}^{(p_i)}$ is the number of times topic j has been assigned to words in user profile p_i.

3.3.2. Transportation Events Representation

After detecting topics from News articles and Weibo posts separately, we obtain the bags of words as event descriptions. The widely used one hot vector presentation for each topic word is unable to calculate the semantic similarity of topic words from different platforms efficiently, therefore, we trained the transportation word embeddings to represent topic words and calculated the semantic similarity between topic pairs. Finally, we linked and fused the topic pairs into event descriptions. Traditionally, in natural language processing each word is represented as a one-hot vector which is 1 at the position associated with the word an 0 at other positions. Clearly, the one-hot representation cannot capture any information about the semantic similarity between words. Moreover, the one-hot vector is high-dimensional and sparse. Recently, word embedding is proposed to represent each word in a continuous vector space and encode many semantic patterns [50,51]. Word embedding was firstly presented by Bengio et al. in [52], and implemented by Mikolov et al. in word2vec [53]. Since then word2vec has gained popularity for natural language processing [54], question and answer systems [55], information retrieval [56], recommending systems [57,58], sentiment analysis [59] etc.

The basic idea of word2vec is to combine a word and its contextual information together, and encode them into a low-dimensional vector. Words with similar contexts in the corpus are located in close proximity to each other in the representation space. Word embedding can be trained either by the Continuous Bag of Words (CBOW) model or the Skip-gram model. Both of them are neural networks which map word(s) to the target variable which is also a word(s), and the learned weights are word embedding representations. Specifically, the CBOW model is learning to predict the word by the context words, in contrast the skip-gram model is learning to predict the context words from the current word. The simplified structure of the two models is shown in Figure 4. Herein, we choose the CBOW model for training transportation word embedding.

However, computing probabilities in softmax layer is the most resource consuming phase when training CBOW models, since it requires summing over all words in the large vocabulary. Therefore, we use the Negative-Sampling [60] method to approximate the softmax layer in the CBOW model, which moves the embedding toward the neighbor words and away from the noise words. In this paper, the noise words are sampled from vocabulary according to their weighted the 3/4 power of unigram probability.

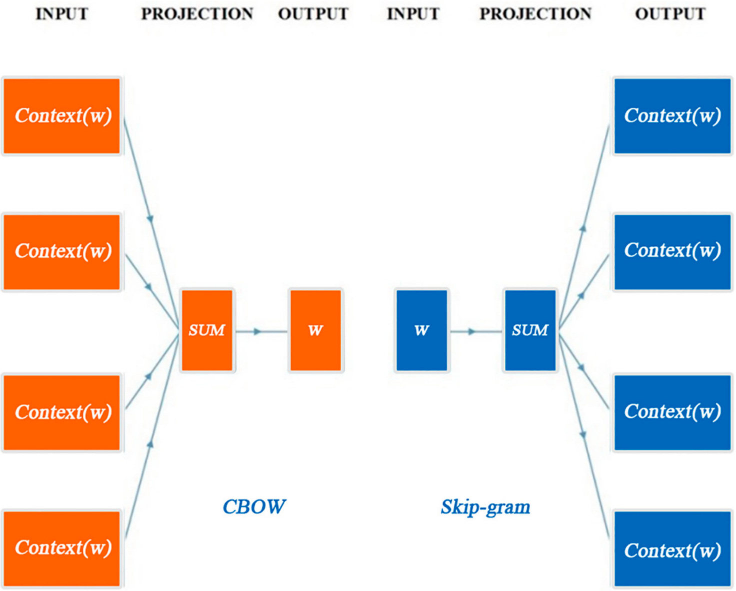


Figure 4. Model structure of CBOW and Skip-gram.

3.3.3. Transportation Events Fusion

Transportation topic clusters and corresponding topic words associated with each cluster can be obtained through the LDA model for news articles and w-LDA model for Weibo. The topic word clusters for Weibo and News are denoted as $\boldsymbol{T}_{\boldsymbol{w}}^{(\boldsymbol{i}=\boldsymbol{1}\ldots\boldsymbol{K})}$ and $\boldsymbol{T}_{\boldsymbol{n}}^{(\boldsymbol{j}=\boldsymbol{1}\ldots\boldsymbol{K})}$, respectively, where K is the total number of topic clusters. Each topic has a set of words, and we denote words in the given topic cluster i and j, as $\boldsymbol{WE}_{\boldsymbol{w}}^{(\boldsymbol{m}=1\ldots\mathbf{M},\boldsymbol{i})}$ and $\boldsymbol{WE}_{\boldsymbol{n}}^{(\boldsymbol{n}=1\ldots\boldsymbol{N},\boldsymbol{j})}$ separately, where M, N are the total number of words in the given topic clusters.

- *Events similarity measure:*

Meanwhile, each word is represented as a 200-dimensional continuous vector. Based on the semantic representation, we define words distance $dis_R\left(\boldsymbol{WE}_w^{(m,i)}, \boldsymbol{WE}_n^{(n,j)}\right)$ to measure the similarity between words in the given topic cluster $\boldsymbol{T}_{\boldsymbol{w}}^{(\boldsymbol{i})}$ and $\boldsymbol{T}_{\boldsymbol{n}}^{(\boldsymbol{j})}$, where R = {Cosine distance, Euclidean distance, Manhattan distance, Chebyshev distance, Weighted Euclidean distance}.

The formulas of the above distances are listed in Table 1. Furthermore, a topic distance matrix $\boldsymbol{D}_{\boldsymbol{R}}\left(\boldsymbol{T}_{\boldsymbol{w}}^{(\boldsymbol{i})}, \boldsymbol{T}_{\boldsymbol{n}}^{(\boldsymbol{j})}\right)$ can be defined as follows to measure the distance between $\boldsymbol{T}_{\boldsymbol{w}}^{(\boldsymbol{i})}$ and $\boldsymbol{T}_{\boldsymbol{n}}^{(\boldsymbol{j})}$:

$$\boldsymbol{D}_{\boldsymbol{R}}\left(\boldsymbol{T}_{\boldsymbol{w}}^{(\boldsymbol{i})}, \boldsymbol{T}_{\boldsymbol{n}}^{(\boldsymbol{j})}\right) = \begin{bmatrix} dis_R\left(\boldsymbol{WE}_{\boldsymbol{w}}^{(\boldsymbol{1},\boldsymbol{i})}, \boldsymbol{WE}_{\boldsymbol{n}}^{(\boldsymbol{1},\boldsymbol{j})}\right) & \cdots & dis_R\left(\boldsymbol{WE}_{\boldsymbol{w}}^{(\boldsymbol{1},\boldsymbol{i})}, \boldsymbol{WE}_{\boldsymbol{n}}^{(\boldsymbol{N},\boldsymbol{j})}\right) \\ \vdots & \ddots & \vdots \\ dis_R\left(\boldsymbol{WE}_{\boldsymbol{w}}^{(\boldsymbol{M},\boldsymbol{i})}, \boldsymbol{WE}_{\boldsymbol{n}}^{(\boldsymbol{1},\boldsymbol{j})}\right) & \cdots & dis_R\left(\boldsymbol{WE}_{\boldsymbol{w}}^{(\boldsymbol{M},\boldsymbol{i})}, \boldsymbol{WE}_{\boldsymbol{n}}^{(\boldsymbol{N},\boldsymbol{j})}\right) \end{bmatrix} \tag{3}$$

Table 1. Formulas of the words vector Distance.

R=	Formula
Euclidean Distance	$\sqrt{\sum_{l=1}^{200}\left(WE_{w}^{(m,i)}[l]\ +\ WE_{n}^{(n,j)}[l]\right)^{2}}$
Cosine Distance	$\frac{\sum_{l=1}^{200} WE_{w}^{(m,i)}[l]\ \times\ WE_{n}^{(n,j)}[l]}{\sqrt{\sum_{l=1}^{200} WE_{w}^{(m,i)}[l]^{2}}\ \times\ \sqrt{\sum_{l=1}^{200} WE_{n}^{(n,j)}[l]^{2}}}$
Manhattan Distance	$\sum_{l=1}^{200}\left\lvert WE_{w}^{(m,i)}[l]\ -\ WE_{n}^{(n,j)}[l]\right\rvert$
Chebyshev Distance	$\lim\limits_{\mathrm{k}\to\infty}\left(\sum_{l=1}^{200}\left\lvert WE_{w}^{(m,i)}[l]\ -\ WE_{n}^{(n,j)}[l]\right\rvert\right)^{\frac{1}{k}}$
Standardized Euclidean Distance	$\sqrt{\sum_{l=1}^{200}\left(\frac{WE_{w}^{(m,i)}[l]+WE_{n}^{(n,j)}[l]}{s[l]}\right)^{2}}$ where $s[l]$ is standard deviation

- *Events alignment:*

Topic alignment is to align topics detected from different platforms. The topics from different platforms are latent variables, thus we do not know each topic labels like traffic jam and traffic accidents. Hence, we need to align the topic clusters detected from the News platform and Weibo platform into topic pairs, which will support the multi-view descriptions in the fusion step.

From bottom up, we choose a $\boldsymbol{WE_{w}^{(m,i)}}$ in $\boldsymbol{T_{w}^{(i)}}$ to calculate the similarity with each word in $\boldsymbol{T_{n}^{(j)}}$, then find the most similar or closest $\boldsymbol{WE_{w}^{(n,j)}}$ in $\boldsymbol{T_{n}^{(j)}}$. Based the closest word pairs, the shortest distance between $\boldsymbol{T_{w}^{(i)}}$ and $\boldsymbol{T_{n}^{(j)}}$ denoted as $\min\limits_{\boldsymbol{m,n=1\ldots N}} dis_{R}\left(\boldsymbol{WE_{w}^{(m,i)}}, \boldsymbol{WE_{n}^{(n,j)}}\right)$, the average shortest distance between $\boldsymbol{T_{w}^{(i)}}$ and $\boldsymbol{T_{n}^{(j)}}$ denoted as $\mu_{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)$, the normalized average shortest distance $\mu_{R}^{*}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)$, and the standard deviation $\sigma_{R}^{*}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)$ can be derived:

$$\mu_{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right) = \frac{\sum_{m=1}^{\mathrm{M}} \min\limits_{n=1\ldots N, n\neq m} dis_{R}\left(\boldsymbol{WE_{w}^{(m,i)}}, \boldsymbol{WE_{n}^{(n,j)}}\right)}{M} \tag{4}$$

Next, we select a topic $\boldsymbol{T_{w}^{(i)}}$, calculate the normalized average distance $\mu_{*}^{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)$ between topic $\boldsymbol{T_{w}^{(i)}}$ and topics in $\boldsymbol{T_{n}^{(j)}}$,

$$\mu_{R}^{*}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right) = \frac{\mu_{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right) - \min\limits_{i=1\ldots k, j=1\ldots k} \mu_{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)}{\max\limits_{i=1\ldots k, j=1\ldots k} \mu_{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right) - \min\limits_{i=1\ldots k, j=1\ldots k} \mu_{R}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)} \tag{5}$$

Then, the topic $\boldsymbol{T_{w}^{(i)}}$ was aligned with the most similar topic $\boldsymbol{T_{n}^{(j)}}$ through calculating $\underset{j=1\ldots k}{argmin}\ \mu_{R}^{*}\left(\boldsymbol{T_{w}^{(i)}}, \boldsymbol{T_{n}^{(j)}}\right)$. Finally, the index pair $(i,\ j)$ of aligned topics was returned.

- *Events fusion:*

After aligning the cross-platform traffic topics for each urban road, we can fuse the aligned topic pairs into a unified event description.

First, the anomalous words were removed. We scan the words list $\boldsymbol{WE_{w}^{(*,i)}}$ in $\boldsymbol{T_{w}^{(i)}}$, and calculate the shortest distance $\min\limits_{n=1\ldots N, n\neq m} dis_{R}\left(\boldsymbol{WE_{w}^{(m,i)}}, \boldsymbol{WE_{n}^{(n,j)}}\right)$ from current word $\boldsymbol{WE_{w}^{(m,i)}}$ to the aligned topic words $\boldsymbol{WE}_{\boldsymbol{n}}^{(*,j)}$. If the shortest distance falls outside the region of three standard deviations $\sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$, $\boldsymbol{WE}_{\boldsymbol{w}}^{(m,i)}$ will be regard as an anomalous word:

$$\sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) = \sqrt{\frac{\sum_{i=1}^{K}\left(\mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) - \frac{1}{K}\sum_{j=1}^{K}\mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)\right)^2}{K}} \tag{6}$$

Second, the topic words from cross-platform were fused. If the shortest distance falls inside the region of one standard deviation $\sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$, the current word will be replaced with candidate $\boldsymbol{WE}_{\boldsymbol{n}}^{(n,j)}$ for a more objective and formal event description. The detailed algorithm for event fusion is shown in Algorithm 2.

Algorithm 2. Transportation Events Alignment and Fusion

```
Input: $\boldsymbol{T_w}$, $\boldsymbol{T_n}$, $\boldsymbol{WE_{weibo}}$, $\boldsymbol{WE_{news}}$
Output: AlignedTopicsList, FusionEventMatrix
//Step 1: aligning the topic clusters in $\boldsymbol{T_w}$ and $\boldsymbol{T_n}$, return AlignedTopicsList
  For each $\boldsymbol{T}_{\boldsymbol{w}}^{(i)}$ in $\boldsymbol{T_w}$
    WeiboIndex = $i$
    For each $\boldsymbol{T}_{\boldsymbol{n}}^{(j)}$ in $\boldsymbol{T_n}$
      If $\mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) <$ minTopicDis
      minTopicDis = $\mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$
      NewsAlignedIndex = $j$
  End for // Find the closest topic cluster
  Append (WeiboIndex, NewsAlignedIndex) to AlignedTopicsList
  End for // couple closet the topic clusters into pairs
//Step 2: fusing the words in the cross-paired topics, return FusionEventMatrix
  For each index pair in AlignedTopicsList
    $i$ = WeiboIndex;
    $j$ = NewsAlignedIndex
    $\mathrm{LB_n} = \mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) - \sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$
    $\mathrm{UB_w} = \mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) + \sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$ // calculate the low boundary and up boundary for news words
    $\mathrm{LB_w} = \mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) - 3\sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$
    $\mathrm{UB_w} = \mu_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right) + 3\sigma_R^*\left(\boldsymbol{T}_{\boldsymbol{w}}^{(i)}, \boldsymbol{T}_{\boldsymbol{n}}^{(j)}\right)$ // calculate the low boundary and up boundary for weibo words
    For each $\boldsymbol{WE}_{\boldsymbol{w}}^{(m,i)}$ in $\boldsymbol{T}_{\boldsymbol{w}}^{(i)}$
      For each $\boldsymbol{WE}_{\boldsymbol{n}}^{(n,j)}$ in $\boldsymbol{T}_{\boldsymbol{n}}^{(j)}$
        If $dis_R\left(\boldsymbol{WE}_{\boldsymbol{w}}^{(m,i)}, \boldsymbol{WE}_{\boldsymbol{n}}^{(n,j)}\right) <$ minWordDis
        minWordDis = $dis_R\left(\boldsymbol{WE}_{\boldsymbol{w}}^{(m,i)}, \boldsymbol{WE}_{\boldsymbol{n}}^{(n,j)}\right)$
        minWordIndex = $n$
      End for // Find the shortest word embedding distance in coupled clusters
      If $\mathrm{LB_w}$ < minWordDis < $\mathrm{UB_w}$ // remove anomaly word in Weibo cluster
        If $\mathrm{LB_n}$ < minWordDis < $\mathrm{UB_n}$
          Append $\boldsymbol{WE}_{\boldsymbol{n}}^{(\boldsymbol{minWordIndex},j)}$ to FusionEventList // append words in News cluster
        Else
        Append $\boldsymbol{WE}_{\boldsymbol{w}}^{(m,i)}$ to FusionEventList // append words in Weibo cluster
    End for
    Append FusionEventList to FusionEventMatrix
  End for
```

4. Experiments

4.1. Data Description

The proposed methodology was applied to detect and fuse urban traffic events in Qingdao (a coastal city of China) from cross-platform social signals. Hence, first we used a social sensors network with 337 traffic keywords to collect data from News and Weibo which related to Qingdao transportation.

After obtaining the raw webpages, we removed the News articles which content length was greater or less than 90% of articles, and also removed the Weibo posts with a number of words of less than 5. Meanwhile, considering there are lots of social bots or online spammers in Weibo [61–64], we only retained the authors that published less than 10 articles in one day.

Next, we removed the punctuation, paragraph symbols, and noise patterns in the Weibo posts and news articles. Then we segmented texts into words, removed stop words, and tagged the words representing city roads with the LTP toolkit. After consulting the local transportation agency, we annotated 132 main roads in Qingdao, and aggregated the preprocessed texts into road data blocks with the criteria we mentioned in Section 3.

Finally, the multi-channel Qingdao transportation dataset from 1 August 2015 to 4 August 2017 was built. The dataset has about 1.15 million texts in total, including 301,684 News articles and 839,587 Weibo posts. The dataset was divided into training dataset, testing dataset and case study dataset, as shown in Table 2. In the following section, we separately used the training dataset to learning WBEF model parameters, the testing dataset to evaluate model's performance, and the case study dataset to discuss the practical application effect of proposed model in open scenario.

Table 2. The Dataset of cross-platform transportation in Qingdao.

	Training Dataset (about 2 Years)			**Testing Dataset (7 Days)**			**Case Study Dataset (1 Day)**		
	Text	**Authors**	**Roads**	**Text**	**Authors**	**Roads**	**Text**	**Authors**	**Roads**
News Articles	301,684	-	132	2369	-	37	334	-	12
Weibo Posts	839,587	271,260	147	4072	2600	49	581	428	25

4.2. WBEF Model Verification

In this section, we first describe the implementation details of WBEF model, show the performance of each sub-model, evaluate the entire performance of the overall WBEF model, and compare the WBEF model with the baseline model.

4.2.1. Transportation Word Embedding

We utilized the CBOW algorithm to train transportation word embedding (200 dimensions). The window size is 5, which indicates the maximum distance between the current and predicted word within a sentence. The number of training epochs over the dataset is 10. The negative sampling algorithm is used to approximate the parameters' gradient, and the number of noise words drawn for current word is 5.

To evaluate transportation word embedding, we selected the most representative words in transportation events (i.e., traffic congestion, traffic accident), while searching for the most similar word embedding semantically. Table 3 gives top five most similar words for traffic congestion and traffic accident. The results show that our transportation word vectors contain essential semantic information.

Table 3. Most similar words for traffic congestion and traffic accident ranked by transportation word2vec.

Traffic Words	Similar Words	Similarity	Traffic Words	Similar Words	Similarity
交通拥堵 Traffic congestion	交通拥挤 Heavy traffic	83.47%	交通事故 Traffic accident	交通事件 Traffic incident	82.19%
	交通阻塞 Traffic jam	83.46%		车祸 Car accident	69.77%
	堵车 Caught in traffic	75.28%		伤亡事故 Fatality accident	62.43%
	塞车 Stuck in traffic	60.27%		碰撞 Traffic crash	60.86%
	滞留 Traffic blocking	59.45%		重大事故 Major accident	59.33%

4.2.2. Transportation Event Detection

During the experiments, we tried different K values (K = 1–5) for aggregating topic clusters, and when K = 3 we get more meaningful word clusters for every topic. We used the perplexity value to evaluate the LDA and *w*-LDA models, respectively. Since there were 132 data blocks that represent multi-channel social signal sensed from the main roads in Qingdao, so the LDA and *w*-LDA model were deployed on every data block for topic generation. Hence, the LDA and *w*-LDA model parameters were determined by the average perplexity of topic model on multiple data blocks as shown in Figure 5, the average perplexity of *w*-LDA converge to the value of 167 after 347 training iterations, the average perplexity of LDA converge to the value of 279 after 182 training iterations. Considering the model generalization on different roads in Qingdao, we choose the topic model which perplexity value was closest to the average perplexity.

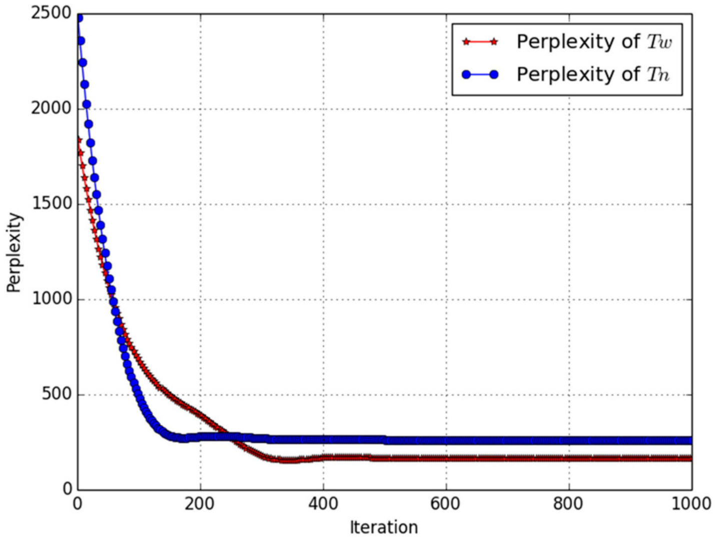


Figure 5. Average Perplexity of LDA for News and *w*-LDA for Weibo posts.

In order to validate the LDA and *w*-LDA topic models, we selected the three largest data blocks in the testing dataset to extract topics for the corresponding road. Herein, to intuitively show the results, we list the words in the largest cluster of $\boldsymbol{T_w}$ and the words in the aligned cluster of $\boldsymbol{T_n}$ for each road (Table 4).

Table 4. The topic detection results from Weibo and News.

Road	T_n	T_w	Type
抚顺路 Fushun Road	抚顺路 施工现场 堵车 交叉路口 缓慢 . . . Fushun Road, construction site, traffic jams, intersection, slow moving . . .	交通广播 抚顺路 一动不动 堵死 附近 高峰时段 疏通 水泄不通 . . . Traffic broadcast, Fushun Road, blocked, slow moving, nearby, peak hours, dispersion, crowd and jam . . .	拥堵 Traffic Jam
银川西路 Yinchuan WestRoad	银川西路 会场 一段路 规划 答复 拥堵 . . . Yinchuan WestRoad, theater, road section, city planning, response, traffic jams . . .	信号灯 失灵十字路口 怎么回事 东向西 吐槽 为何 . . . Traffic signals, not work, crossroads, what happened, east to west, complain, reason . . .	投诉 Traffic Complaint
人民路 RenMin Road	人民路 交警 民警 男子 驾驶 撞墙 双腿 受伤 . . . RenMin Road, traffic police, police officer, man, car driving, hit the wall, legs, injured . . .	人民路 救援 现场 不慎 情况危急 求助 车底 . . . RenMin Road, rescue, accident scene, critical situation, help, car bottom . . .	事故 Traffic Accident

4.2.3. Distance Metrics for Transportation Events Fusion

We tested different distance metrics for fusing traffic topics, and chose the proper distance metric for semantic event fusion. The top 10 largest road data blocks in the testing dataset were selected, and the topics were extracted from each data block with the LDA and w-LDA models, respectively. The words in the largest topic cluster $\boldsymbol{T}_w^{(i)}$ and the corresponding aligned cluster $\boldsymbol{T}_n^{(j)}$ were considered as the fusion corpus for distance comparison experiments.

For each distance measure methods we computed the average and standard deviation of topic distances $D_R\left(\boldsymbol{T}_w^{(i)},\ \boldsymbol{T}_n^{(j)}\right)$ on different data blocks, expressed as $\mu_R\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$ and $\sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$. For the model robustness, we chose the distance has the minimum standard deviation on multiple roads, which can be obtained through $\underset{R}{\operatorname{argmin}} \sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$. The results in Table 5 show that the most stable distance measure is the cosine distance for word embedding presentation.

Table 5. Average and standard deviation of topic distances for different roads.

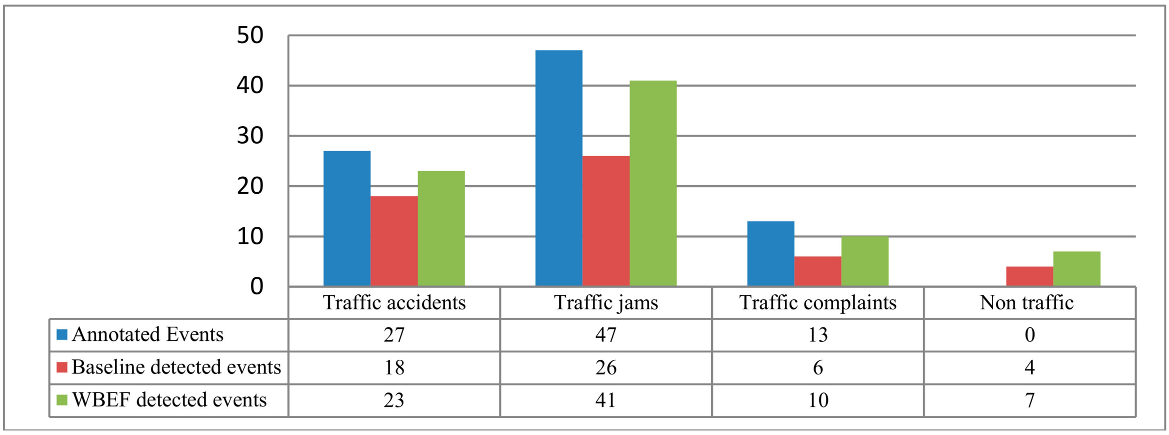
Distance Measure (R=)	**Average and Standard Deviation**	**人民路 RenMin Road**	**伊春路 YiChun Road**	**山东路 ShanDong Road**	**延吉路 JiYan Road**	**延安三路Yan' An San Road**	**抚顺路 FuShun Road**	**敦化路 AnHua Road**	**登州路 DengZhou Road**	**银川西路 YinChuan Xi Road**	**鞍山路 AnShan Road**
Euclidean	$\mu_R\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.896	1.088	1.03	1.09	0.973	0.986	1.101	0.96	1.083	0.916
	$\sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.07278									
Cosine	$\mu_R\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.0031	0.0041	0.0033	0.0036	0.0029	0.0032	0.0071	0.0034	0.0036	0.0029
	$\sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.00117									
Manhattan	$\mu_R\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	10.15	12.33	11.31	12.28	10.97	11.12	12.36	10.81	12.21	10.32
	$\sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.8176									
Chebyshev	$\mu_R\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.169	0.206	0.191	0.207	0.185	0.189	0.207	0.181	0.208	0.174
	$\sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.01386									
Weighted Euclidean	$\mu_R\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	19.78	19.99	19.68	19.54	19.91	19.6	19.54	20	19.91	20
	$\sigma_R^*\left(\boldsymbol{T}_w^{(i)}, \boldsymbol{T}_n^{(j)}\right)$	0.17455									

4.3. Overall Performance

To evaluate the overall performance of the proposed model, we selected 10 roads in Qingdao with the most frequently occurring traffic events, checked the content of News and Weibo articles in every road data block, and manually annotated the traffic events including traffic accidents, traffic jams and traffic complements in the testing dataset. By consulting a traffic domain expert from the Qingdao

Transportation Committee, the traffic events annotation results which consisted of 27 traffic accidents, 47 traffic jams, and 13 traffic complaints in testing dataset were finally confirmed.

Moreover, the dataset was built with keywords-based social sensors, thus there will be non-traffic News and Weibo articles which also contain traffic keywords, such as traffic products release, traffic safety lectures, and civilized traffic initiative, etc. Hence, if the events were widespread in social media or news, the model may also falsely flag these non-traffic events. In that case, the precision, recall and F_1 Score were adopted to evaluate the model performance. The proposed model was applied to detect and fuse traffic events on the 10 roads during one week. Meanwhile, we selected the "standard LDA + keywords matching" approach as the baseline model. Experimental results are given in Figure 6 and Table 6. Although the precision value of our proposed model (91.4%) is less than baseline model (92.6%), the recall value (85.1%) and F_1 score (88.1%) is much better than baseline. Overall the WBEF model surpassed the baseline model.



	Traffic accidents	Traffic jams	Traffic complaints	Non traffic
🟦 Annotated Events	27	47	13	0
🟥 Baseline detected events	18	26	6	4
🟩 WBEF detected events	23	41	10	7

Figure 6. Comparison of detected and annotated number of traffic events.

Table 6. Performance comparison of Baseline and the proposed model.

	Hits	Miss	False Alarm	Precision	Recall	F_1 Score
Baseline Model	50	37	4	92.6%	57.4%	70.9%
The Proposed Model	74	13	7	91.4%	85.1%	88.1%

The experimental results shows that the baseline model "standard LDA + keywords matching" is unable to effectively process short content in Weibo, and also lacks semantic meaning when the different word styles from cross-platform media are fused, so fewer traffic events have been sensed. However, the event words both exactly occurred in News articles and Weibo posts imply the traffic events have been confirmed by both officials and the masses, which leads to higher accuracy of the baseline model. Compared to the baseline model, the WBEF model grouped the short messages into user profiles, then processed and clustered user-central context through the w-LDA scheme, hence traffic topics in Weibo can be detected more effectively. Meanwhile, with the WBEF model, the detected cross-platform topics can be embedded with semantics, which guarantees topic similarity calculation and event fusion, so more traffic events can be detected, resulting in much higher recall value, which effectively solves the problem of missing detection in the baseline model. However, the noise information in the dataset that is collected by the keywords-based social network sensors, also caused bias in the processing and fusion step, which leads to slightly lower precision.

Quantitatively evaluation, the F_1 score of WBEF model exceeded the baseline model by nearly 17 percentage points, which means the proposed model is much superior to the baseline model in sensing and detecting traffic events from multi-channel social signals. In the next section, we will choose one case, qualitatively study and discuss the traffic event detection effects of WBEF approaches in open scenario.

5. Case Study in Application Scenario

In this section, we chose the traffic situation in Qingdao on 4 August 2017 as one case, sensed multi-channel social transportation signals and deployed the WBEF model in the open scenario. As shown in Table 2, the case study dataset contains about one thousand articles. The WBEF model successfully identified 11 traffic events, missed two traffic events and false alarmed one traffic event, finally achieving 91.6% precision, 84.6% recall and an 87.9% F_1 score.

All the traffic events were mapped to the city roads, as shown in Figure 7, and the overall urban transportation situation can be visualized, which will intuitively support traffic management and traffic plan decisions. Meanwhile, in order to qualitatively investigate the practical usability and effectiveness of the proposed model, we chose the Top 3 hottest events (largest clusters) that occurred on different roads, and imported the corresponding road block data into the baseline and WBEF models. Although both models successfully detected Top 3 hottest events, the baseline model generated fewer words with less semantic. In contrast, the WBEF model presented more comprehensible and understandable words description of transportation events.

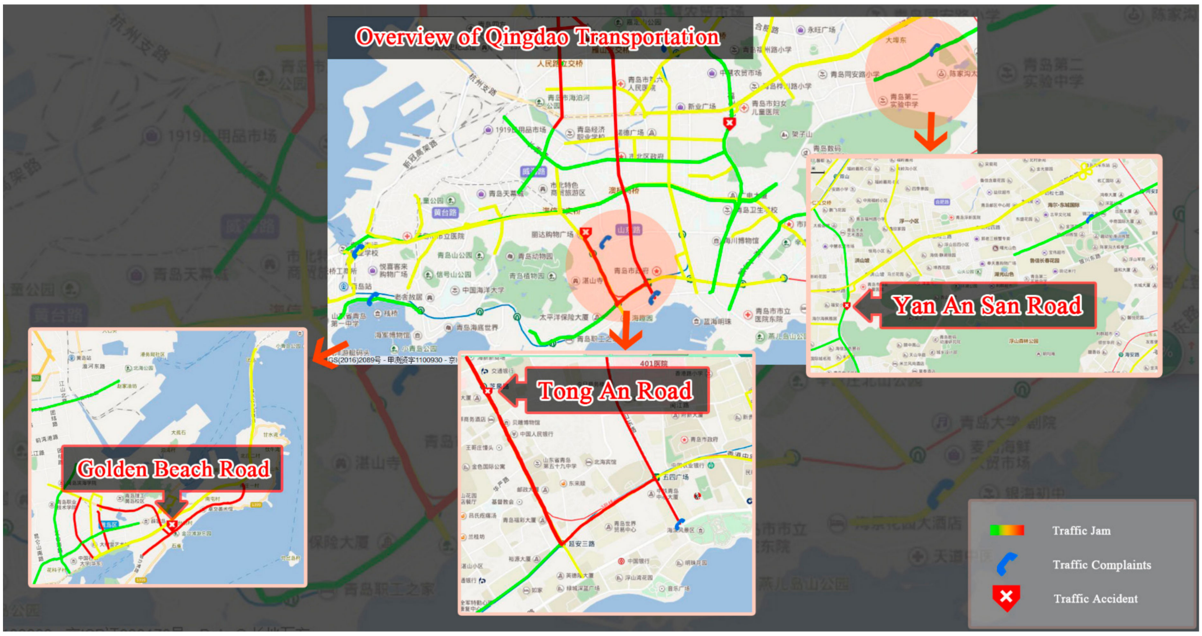


Figure 7. Overview of Qingdao transportation on 4 August 2017.

Furthermore, we investigated event detail and corresponding causal factors. To clearly demonstrate event words, we manually categorized and tagged event words into target road words, relevant location words, traffic words, reason words, and other words (shown in Table 7). We also double-checked News and Weibo data related to the event and got insights into the event details. Three target roads with detected traffic events are detailed in the following.

Table 7. Top 3 hottest fused traffic events in Qingdao on 4 August 2017.

Target Road	Relevant Locations	Traffic Words	Traffic Causal Words	Other Words
金沙滩路 Gold Beach Road	黄岛区 啤酒城 隧道 Huangdao District, Beer Square, Cross-sea Tunnel.	缓慢 交通拥堵 高峰 客流 车流量 停车 交警 Slow, traffic jam, rush hour, passenger flow, traffic flow, parking, traffic polices	啤酒节 开幕式 开园 Beer festival, opening ceremony, park opening	黄晓明 明星 直播 Xiaoming Huang, Superstar, live show
同安路 Tong An Road	国信体育场 银川东路 劲松七路 海尔路 Guoxin Stadium, Yin Chuan Xi Road, Jin Song Qi Road, Hai Er Road	绕行 高峰 慢行 管制 拥堵 停车 调流 交通广播 预案 Passing round, rush hour, slow, traffic control, traffic jam, parking, traffic flow regulation, traffic broadcast, Traffic plan.	五月天 演唱会 Mayday singer group, concert	门票 五迷 提示 Tickets, fans, notice
延安三路 Yan An San Road	石油大厦 Petroleum Building	交通拥堵 交警 交通广播 道路封锁 车辆疏导 疏散 Traffic jam, traffic policies, traffic broadcast, road closed, traffic dispersion, evacuation	火情 火警 起火 扑灭 Fire situation, fire alarm, catch fire, Fire extinguished	伤亡 电梯井 Casualties, elevator hoist way

Gold Beach Road (金沙滩路): on this road, event fusion words included traffic words "parking place (停车场)", "traffic jam (交通拥堵)", "slow down (慢行)", "traffic broadcast (交通广播)", social activity words like "beer festival (啤酒节)", "opening ceremony (开幕式)", and famous star words like "Xiaoming Huang (黄晓明)". After referring to Weibo and News data associated to this event, we found that the annual Qingdao International Beer Festival was opening on the square of Qingdao Beer City, where the film star Xiaoming Huang and other famous stars presented a show for the celebration. In addition, the ceremony was opening at the evening rush hour (7 p.m.) which caused a heavy traffic jam on Gold Beach Road.

Tong An Road (同安路): with the proposed methods, we observed the location words "GuoXin stadium (国信体育场)" and traffic irrelevant words "Mayday (五月天, a singer group)", "concert (演唱会)" in the event description. However, there was no words about traffic jams or accidents, but only traffic words like "detour (绕行)", "control (管制)", "dispatch (调流)", "plan (预案)" etc. We further checked the relevant Weibo and News data, then found that a concert would be held on the next day, so the traffic agency was providing early alerts about the traffic situation on the road, releasing the traffic control notice through social media and news. Massive fans and audiences forwarded and disseminated these posts.

Yan An San Road (延安三路): The event words include traffic words like "traffic jam (交通拥堵)", "evacuation (疏散), and also fire emergency words like "fires (火情)", "fireman (火警)". Associated with the location word "Petroleum Building (石油大厦)", we inferred there was a traffic jam on "Yan'An San Road (延安三路)" which was caused by a fire emergency. After further checked, the real fact was consistent with the detected results from cross-platform media. The elevator in the Petroleum Building caught fire at 7 p.m. Lots of citizens posted the live fire and traffic situation, gave notices and conveyed safety messages to their family. The official announcement also reported the event through news and social media.

6. Conclusions and Future Work

In Intelligent transportation systems, the method of analyzing social signals (social transportation) offers advantages of low cost and large coverage over traditional methods which depend on physical sensors. In this paper, we addressed the challenges of cross-platform traffic event detection when shifting the social signals from a single channel to multiple channels. The WBEF model for urban transportation event detection has been proposed, which benefits from comprehensive social signals, domain knowledge and semantic representation. The model was trained with about 1.15 million News and Weibo data from the past 2 years, and deployed to assess the traffic situation in Qingdao. Experiments show that the sub-models of WBEF achieved the expected performance, and the overall performance of WBEF is much superior to the baseline model. Moreover, from the case study in the open scenario, the accuracy and robustness of WBEF have been further demonstrated.

Further investigation can be conducted: firstly, more social signal sources can be involved into the WBEF model, such as Instagram, Facebook, Quora, etc., which will make traffic incident detection more accurate and comprehensive. Secondly, powerful deep learning methods have highly potentials to improve accuracy and robustness for cross-platform event detection. Thirdly, the WBEF model can be extended to heterogeneous recommending system [65,66], which will achieve more personalized and accurate information services in transportation domain. Furthermore, social sensors combine with physical sensors (Cyber-Physical-Social-System, CPSS) will lead a novel way to monitor, control, and optimize intelligent transportation systems.

Author Contributions: Conceptualization, H.L.; methodology, H.L., K.S.; software, K.S., Y.Z.; writing—original draft preparation, K.S.; writing—review and editing, H.L.; supervision, Y.L., Z.N.; funding acquisition, Y.L., Z.N.

Funding: This work was supported by the National Natural Science Foundation of China under grants 61233001, 61773381 and 61370137, Ministry of Education-China Mobile Research Foundation Project No.2016/2-7.

Acknowledgments: The authors would like to thank the Qingdao Transportation Committee for providing professional advice on improving our model. The authors also would like to thank NSFC for supporting this research. Finally, the authors would like to thank the editor and anonymous reviewers for their suggestions that improved the work.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Table of Variables

Symbol	Description	Data Structure	Supporting Process
T	Number of Weibo posts	Int T	w-LDA
U	Number of Weibo users	Int U	
K	Number of topics	Int K	
V	Number of words in the vocabulary	Int V	
V	Number of uers profiles	Int V	
N_p	Number of words in p-th user profile	Int N[P]	
$\boldsymbol{\alpha}$	K-dimensional prior weight vectors of topics in a document,	Float a[K]	
$\boldsymbol{\beta}$	V-dimensional vector prior weight of words in a topic	Float b[V]	
$\boldsymbol{\varphi_z}$	V-dimensional vector of probabilities, represents distribution of words in topic z	Double phi [Z][V]	
$\boldsymbol{\vartheta_p}$	K-dimensional vector of probabilities, represents distribution of topics in user profile p	Double theta [P][K]	
z_i	Identity of current topic of word w_i in user profile p_i	Int *1 … K*	
w_i	Identity of current word in user profile p_i	Int *1 … V*	
p_i	Identity of current user profiles	Int *1 … P*	
$n_{i,j}^{(p_i)}$	Document-Topic matrix, the number of times topic j has been assigned to words in user profile p_i.	int npt [P][K]	
$n_{i,j}^{(w_i)}$	Topic-Word matrix, Number of times that word w_i has been assigned to topic j	int ntw [K][V]	
$\boldsymbol{W_i}$	Identity of current word vector (200 dimensions) trained by traffic word2vec	Double [200]	*Similarity measure*
$\boldsymbol{T_w^{(i=1\ldots K)}}$	Identity of current topic word cluster detected from Weibo	Double tw[K]	
$\boldsymbol{T_n^{(j=1\ldots K)}}$	Identity of current topic word cluster detected from News	Double tn[K]	
$\boldsymbol{WE}_w^{(m,i)}$	Word embedding- cluster tensor, identity of the current word embedding in i-th cluster detected from Weibo	Double cew[K] [$\boldsymbol{W_i}$]	
$\boldsymbol{WE}_n^{(n,j)}$	Word embedding- cluster tensor, identity of the current word embedding in i-th cluster detected from News	Double cen [K] [$\boldsymbol{W_i}$]	
$dis_R\left(\boldsymbol{WE}_w^{(m,i)}, \boldsymbol{WE}_n^{(n,j)}\right)$	Words similarity, measure the similarity between word embedding $\boldsymbol{WE}_w^{(m,i)}$ and $\boldsymbol{WE}_n^{(n,j)}$	Double wd	
$\boldsymbol{D_R}\left(\boldsymbol{T_w^{(i)}}, \boldsymbol{T_n^{(j)}}\right)$	Topic similarity matrix, measure the distances between each words in the given topic cluster $\boldsymbol{T_w^{(i)}}$ and $\boldsymbol{T_n^{(j)}}$	Double td[K][K]	*Event fusion*
$\mu_R\left(\boldsymbol{T_w^{(i)}}, \boldsymbol{T_n^{(j)}}\right)$	Average shortest distance between $\boldsymbol{T_w^{(i)}}$ and $\boldsymbol{T_n^{(j)}}$	Double atd	
$\mu_R^*\left(\boldsymbol{T_w^{(i)}}, T_{\boldsymbol{n}}^{(j)}\right)$	Normalized average shortest distance between $\boldsymbol{T_w^{(i)}}$ and $\boldsymbol{T_n^{(j)}}$	Double natd	
$\sigma_R^*\left(\boldsymbol{T_w^{(i)}}, \boldsymbol{T_n^{(j)}}\right)$	The standard deviation of normalized topic distances	Double sd	

References

1. Cheng, S.-T.; Li, J.-P.; Horng, G.-J.; Wang, K.-C. The Adaptive Road Routing Recommendation for Traffic Congestion Avoidance in Smart City. *Wirel. Pers. Commun.* **2014**, *77*, 225–246. [CrossRef]

2. Lecue, F.; Tallevi-Diotallevi, S.; Hayes, J.; Tucker, R.; Bicer, V.; Sbodio, M.; Tommasi, P. Smart traffic analytics in the semantic web with STAR-CITY: Scenarios, system and lessons learned in Dublin City. *J. Web Semant.* **2014**, *27–28*, 26–33. [CrossRef]
3. Papageorgiou, M.; Diakaki, C.; Dinopoulou, V.; Kotsialos, A.; Wang, Y.B. Review of road traffic control strategies. *Proc. IEEE* **2003**, *91*, 2043–2067. [CrossRef]
4. El Faouzi, N.E.; Leung, H.; Kurian, A. Data fusion in intelligent transportation systems: Progress and challenges—A survey. *Inf. Fusion* **2011**, *12*, 4–10. [CrossRef]
5. Morgul, E.F.; Yang, H.; Kurkcu, A.; Ozbay, K.; Bartin, B.; Kamga, C.; Salloum, R. Virtual Sensors Web-Based Real-Time Data Collection Methodology for Transportation Operation Performance Analysis. *Transp. Res. Rec.* **2014**, *2442*, 106–116. [CrossRef]
6. Zheng, X.H.; Chen, W.; Wang, P.; Shen, D.Y.; Chen, S.H.; Wang, X.; Zhang, Q.; Yang, L. Big Data for Social Transportation. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 620–630. [CrossRef]
7. Wang, X.; Zheng, X.H.; Zhang, Q.P.; Wang, T.; Shen, D.Y. Crowdsourcing in ITS: The State of the Work and the Networking. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 1596–1605. [CrossRef]
8. Xiong, G.; Zhu, F.; Liu, X.; Dong, X.; Huang, W.; Chen, S.; Zhao, K. Cyber-physical-social system in intelligent transportation. *IEEE J. Autom. Sin.* **2015**, *2*, 320–333.
9. Wang, F.Y. The Emergence of Intelligent Enterprises: From CPS to CPSS. *IEEE Intell. Syst.* **2010**, *25*, 85–88. [CrossRef]
10. Wang, F.Y. Parallel Control and Management for Intelligent Transportation Systems: Concepts, Architectures, and Applications. *IEEE Trans. Intell. Transp. Syst.* **2010**, *11*, 630–638. [CrossRef]
11. Zeng, K.; Liu, W.L.; Wang, X.; Chen, S.H. Traffic Congestion and Social Media in China. *IEEE Intell. Syst.* **2013**, *28*, 72–77. [CrossRef]
12. Mo, H.; Hao, X.X.; Zheng, H.B.; Liu, Z.Z.; Wen, D. Linguistic Dynamic Analysis of Traffic Flow Based on Social Media-A Case Study. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 2668–2676. [CrossRef]
13. D'Andrea, E.; Ducange, P.; Lazzerini, B.; Marcelloni, F. Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 2269–2283. [CrossRef]
14. Cao, J.; Zeng, K.; Wang, H.; Cheng, J.; Qiao, F.; Wen, D.; Gao, Y. Web-Based Traffic Sentiment Analysis: Methods and Applications. *IEEE Trans. Intell. Transp. Syst.* **2014**, *15*, 844–853.
15. Nguyen, H.; Liu, W.; Rivera, P.; Chen, F. TrafficWatch: Real-Time Traffic Incident Detection and Monitoring Using Social Media. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Auckland, New Zealand, 19–22 April 2016.
16. Ni, M.; He, Q.; Gao, J. Forecasting the Subway Passenger Flow Under Event Occurrences with Social Media. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1623–1632. [CrossRef]
17. Maghrebi, M.; Abbasi, A.; Waller, S.T. Transportation Application of Social Media: Travel Mode Extraction. In Proceedings of the 19th International Conference on Intelligent Transportation Systems, Rio de Janeiro, Brazil, 1–4 November 2016.
18. Abidin, A.F.; Kolberg, M.; Hussain, A. Integrating Twitter Traffic Information with Kalman Filter Models for Public Transportation Vehicle Arrival Time Prediction. *Big-Data Anal. Cloud Comput.* **2015**. [CrossRef]
19. Chaniotakis, E.; Antoniou, C. Use of Geotagged Social Media in Urban Settings: Empirical Evidence on Its Potential from Twitter. In Proceedings of the 2015 IEEE 18th International Conference on Intelligent Transportation Systems, Las Palmas, Spain, 15–18 September 2015.
20. Xiong, G.; Zhu, F.; Dong, X.; Fan, H.; Hu, B.; Kong, Q.; Kang, W.; Teng, T. A Kind of Novel ITS Based on Space-Air-Ground Big-Data. *IEEE Intell. Transp. Syst. Mag.* **2016**, *8*, 10–22. [CrossRef]
21. Zhou, T.; Gao, L.; Ni, D. Road traffic prediction by incorporating online information. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014.
22. Zhu, F.; Li, Z.; Chen, S.; Xiong, G. Parallel Transportation Management and Control System and Its Applications in Building Smart Cities. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 1576–1585. [CrossRef]
23. Zhang, J.; Wang, F.Y.; Wang, K.; Lin, W.H.; Xu, X.; Chen, C. Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 1624–1639. [CrossRef]
24. Lv, Y.; Chen, Y.; Zhang, X.; Duan, Y.; Li, N.L. Social media based transportation research: The state of the work and the networking. *IEEE J. Autom. Sin.* **2017**, *4*, 19–26. [CrossRef]
25. Chaniotakis, E.; Antoniou, C.; Pereira, F. Mapping Social Media for Transportation Studies. *IEEE Intell. Syst.* **2016**, *31*, 64–70. [CrossRef]

26. Ahmed, M.S.; Cook, A.R. Analysis of Freeway Traffic Time-series Data by Using Box-jenkins Techniques. *Transp. Res. Board* **1979**, *773*, 1–9.
27. He, J.; Shen, W.; Divakaruni, P.; Wynter, L.; Lawrence, R.D. Improving traffic prediction with tweet semantics. In Proceedings of the Proceedings of the 23th International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013.
28. Gu, Y.; Qian, Z.; Chen, F. From Twitter to detector: Real-time traffic incident detection using social media data. *Transp. Res. Part C Emerg. Technol.* **2016**, *67*, 321–342. [CrossRef]
29. Fu, K.; Lu, C.T.; Nune, R.; Tao, J.X. Steds: Social Media Based Transportation Event Detection with Text Summarization. Available online: http://europa.nvc.cs.vt.edu/~{}ctlu/Publication/2015/IEEE-ITSC-Proceedings-STEDS-2015.pdf (accessed on 20 November 2018).
30. Gutiérrez, C.; Figuerias, P.; Oliveira, P.; Costa, R.; Jardim-Goncalves, R. Twitter mining for traffic events detection. In Proceedings of the 2015 Science and Information Conference (SAI), London, UK, 28–30 July 2015.
31. Lu, H.; Zhu, Y.; Shi, K.; Lv, Y.; Shi, P.; Niu, Z. Using Adverse Weather Data in Social Media to Assist with City-Level Traffic Situation Awareness and Alerting. *Appl. Sci.* **2018**, *8*, 1193. [CrossRef]
32. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
33. Zhai, K.; Boyd-Graber, J. Online latent Dirichlet Allocation with Infinite Vocabulary. Available online: http://proceedings.mlr.press/v28/zhai13.pdf (accessed on 20 November 2018).
34. Paisley, J.; Wang, C.; Blei, D.M.; Jordan, M.I. Nested Hierarchical Dirichlet Processes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 256–270. [CrossRef] [PubMed]
35. Quan, X.; Kit, C.; Ge, Y.; Pan, S.J. Short and sparse text topic modeling via self-aggregation. In Proceedings of the 24th International Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015.
36. Ramage, D.; Dumais, S.T.; Liebling, D.J. Characterizing Microblogs with Topic Models. In Proceedings of the Fourth International Conference on Weblogs and Social Media, Washington, DC, USA, 23–26 May 2010.
37. Zhao, W.X.; Jiang, J.; Weng, J.; He, J.; Lim, E.-P.; Yan, H.; Li, X. Comparing twitter and traditional media using topic models. In Proceedings of the 33rd European conference on Advances in information retrieval, Dublin, Ireland, 18–21 April 2011.
38. Hou, L.; Li, J.; Li, X.-L.; Su, Y. Measuring the Influence from User-Generated Content to News via Cross-dependence Topic Modeling. In Proceedings of the International Conference on Database Systems for Advanced Applications, Hanoi, Vietnam, 20–23 April 2015.
39. Oghina, A.; Breuss, M.; Tsagkias, M.; De Rijke, M. Predicting IMDB Movie Ratings Using Social Media. In Proceedings of the European Conference on Information Retrieval, Barcelona, Spain, 1–5 April 2012.
40. Bao, B.-K.; Xu, C.; Min, W.; Hossain, M.S. Cross-Platform Emerging Topic Detection and Elaboration from Multimedia Streams. *ACM Trans. Multimed. Comput. Commun. Appl.* **2015**, *11*, 1–21. [CrossRef]
41. Koike, D.; Takahashi, Y.; Utsuro, T.; Yoshioka, M.; Kando, N. Time Series Topic Modeling and Bursty Topic Detection of Correlated News and Twitter. Available online: http://www.aclweb.org/anthology/I13-1118 (accessed on 20 November 2018).
42. Qian, S.; Zhang, T.; Xu, C. A Generic Framework for Social Event Analysis. In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, Bucharest, Romania, 6–9 June 2017.
43. Li, M.; Wang, J.; Tong, W.; Yu, H.; Ma, X.; Chen, Y.; Cai, H.; Han, J. EKNOT: Event knowledge from news and opinions in Twitter. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
44. Wang, J.; Tong, W.; Yu, H.; Li, M.; Ma, X.; Cai, H.; Hanratty, T.; Han, J. Mining Multi-aspect Reflection of News Events in Twitter: Discovery, Linking and Presentation. In Proceedings of the 2015 IEEE International Conference on Data Mining, Atlantic City, NJ, USA, 14–17 November 2015.
45. Wang, F.Y. A framework for social signal processing and analysis: from social sensing networks to computational dialectical analytics. *Sci. Sin. Inf.* **2013**, *43*, 1598–1611.
46. Liu, T.; Che, W.; Zhenghua, L.I. Language Technology Platform. *J. Chin. Inf. Process.* **2011**, *2*, 13–16.
47. Kim, E.H.J.; Jeong, Y.K.; Kim, Y.Y.; Kang, K.Y.; Song, M. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *J. Inf. Sci.* **2016**, *42*, 763–781. [CrossRef]
48. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, Raleigh, CA, USA, 26–30 April 2010.

49. Hong, L.; Davison, B.D. Empirical study of topic modeling in Twitter. In Proceedings of the SIGKDD Workshop on Social Media Analytics, Washington, DC, USA, 25 June 2010.
50. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global Vectors for Word Representation. Available online: http://www.aclweb.org/anthology/D14-1162 (accessed on 20 November 2018).
51. Lai, S.; Liu, K.; He, S.; Zhao, J. How to Generate a Good Word Embedding. *IEEE Intell. Syst.* **2016**, *31*, 5–14. [CrossRef]
52. Bengio, Y.; Ducharme, R.; Vincent, P.; Jauvin, C. A neural probabilistic language model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
53. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv*, 2013; arXiv:1301.3781. Available online: https://arxiv.org/abs/1301.3781(accessed on 20 November 2018).
54. Socher, R.; Bauer, J.; Manning, C.D.; Y, N.A. Parsing with Compositional Vector Grammars. Available online: http://www.aclweb.org/anthology/P13-1045 (accessed on 20 November 2018).
55. Zhou, G.; He, T.; Zhao, J.; Hu, P. Learning Continuous Word Embedding with Metadata for Question Retrieval in Community Question Answering. Available online: http://www.aclweb.org/anthology/P15-1025 (accessed on 20 November 2018).
56. Rekabsaz, N. Enhancing Information Retrieval with Adapted Word Embedding. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016.
57. Musto, C.; Semeraro, G.; De Gemmis, M.; Lops, P. Learning Word Embeddings from Wikipedia for Content-Based Recommender Systems. In Proceedings of the European Conference on Information Retrieval, Padua, Italy, 20–23 March 2016.
58. Lin, Q.; Niu, Y.; Zhu, Y.; Lu, H.; Shi, K.; Keith, Z.; Niu, Z. Heterogeneous Knowledge-Based Attentive Neural Networks for Short-Term Music Recommendations. *IEEE Access.* **2018**, *6*, 58990–59000. [CrossRef]
59. Giatsoglou, M.; Vozalis, M.G.; Diamantaras, K.; Vakali, A.; Sarigiannidis, G.; Chatzisavvas, K.C. Sentiment analysis leveraging emotions and word embeddings. *Expert Syst. Appl.* **2017**, *69*, 214–224. [CrossRef]
60. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013.
61. Ghosh, S.; Korlam, G.; Ganguly, N. Spammers' networks within online social networks: A case-study on Twitter. In Proceedings of the 20th International Conference Companion on World Wide Web, Hyderabad, India, 28 March–1 April 2011.
62. Yang, C.; Harkreader, R.; Gu, G. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Trans. Inf. Forensics Secur.* **2013**, *8*, 1280–1293. [CrossRef]
63. Hayati, P.; Chai, K.; Potdar, V.; Talevski, A. HoneySpam 2.0: Profiling Web Spambot Behaviour. In Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems, Nagoya, Japan, 14–16 December 2009.
64. Chen, C.; Wu, K.; Srinivasan, V.; Zhang, X. Battling the internet water army: detection of hidden paid posters. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, ON, Canada, 25–28 August 2013.
65. Tarus, J.K.; Niu, Z.; Yousif, A. A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining. *Future Gener. Comput. Syst.* **2017**, *72*, 37–48. [CrossRef]
66. Tarus, J.K.; Niu, Z.; Mustafa, G. Knowledge-based recommendation: A review of ontology-based recommender systems for e-learning. *Artif. Intell. Rev.* **2018**, *50*, 21–48. [CrossRef]

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).