Geo-Tagged Social Media Data-Based Analytical Approach for Perceiving Impacts of Social Events

Journal Article

Author(s):
Zhu, Ruoxin; Lin, Diao; Jendryke, Michael; Zuo, Chenyu; Ding, Linfang; Meng, Liqiu

Article

Geo-Tagged Social Media Data-Based Analytical Approach for Perceiving Impacts of Social Events

Ruoxin Zhu [1],* , Diao Lin [1] , Michael Jendryke [2,3] , Chenyu Zuo [1], Linfang Ding [1,4] and Liqiu Meng [1]

[1] Chair of Cartography, Technical University of Munich, 80333 Munich, Germany; diao.lin@tum.de (D.L.); chenyu.zuo@tum.de (C.Z.); linfang.ding@tum.de (L.D.); liqiu.meng@tum.de (L.M.)
[2] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China; mjendryke@whu.edu.cn
[3] School of Resources and Environmental Science, Wuhan University, Wuhan 430079, China
[4] KRDB Research Centre, Faculty of Computer Science, Free University of Bozen-Bolzano, 39100 Bolzano, Italy
* Correspondence: ruoxin.zhu@tum.de; Tel.: +49-152-5723-5586

check for updates

Abstract: Studying the impact of social events is important for the sustainable development of society. Given the growing popularity of social media applications, social sensing networks with users acting as smart social sensors provide a unique channel for understanding social events. Current research on social events through geo-tagged social media is mainly focused on the extraction of information about when, where, and what happened, i.e., event detection. There is a trend towards the machine learning of more complex events from even larger input data. This research work will undoubtedly lead to a better understanding of big geo-data. In this study, however, we start from known or detected events, raising further questions on how they happened, how they affect people's lives, and for how long. By combining machine learning, natural language processing, and visualization methods in a generic analytical framework, we attempt to interpret the impact of known social events from the dimensions of time, space, and semantics based on geo-tagged social media data. The whole analysis process consists of four parts: (1) preprocessing; (2) extraction of event-related information; (3) analysis of event impact; and (4) visualization. We conducted a case study on the "2014 Shanghai Stampede" event on the basis of Chinese Sina Weibo data. The results are visualized in various ways, thus ensuring the feasibility and effectiveness of our proposed framework. Both the methods and the case study can serve as decision references for situational awareness and city management.

Keywords: social sensing; machine learning; social opinion mining; topic discovery; visual analysis

1. Introduction

The world has witnessed a tremendous upsurge of social events, which began during the late nineteenth century and continued into the twenty-first century. A variety of social events, especially negative ones (e.g., terrorist attacks, violent incidents), have undeniable impacts on a variety of aspects regarding individuals and society at large, which are worthy of recognition as a distinct academic discipline [1]. Scholars have mainly used traditional sociological methods (e.g., interviews and sample surveys) to study the impact of social events, leading to a large number of insightful empirical results. Social sensing was spawned as a result of rapid technological development, especially in communication technology and mobile positioning technology, which triggered a data-driven channel for understanding social events. Social sensing is powerful for crowd-sourcing users, with each individual playing the role of a sensor [2]. Various types of big data (e.g., social media data, trajectory

data, and check-in data) generated by social sensors bring us new opportunities for understanding the natural and social environment in which we exist.

The knowledge people have of an event develops from a perceptual to a rational stage [3]. One of the main driving forces people have is to share their knowledge of real-world events, such as a government election or a traffic jam [4]. Acting as an effective means of social sensing, social media services (e.g., Facebook and Chinese Sina Weibo) are especially popular among people wanting to share their thoughts about nearby events. These data from social media may be extracted and processed for various purposes. In particular, the spatial–temporal and semantic information embedded in geo-tagged social media data provides valuable indicators for us in investigating and understanding social events, which may be the most extensive information container for dynamic geo-historical phenomena. Current research dealing with geo-tagged social media data focuses more on event detection and tracking rather than the study of how the impacts of social events are reflected in geo-tagged social media data.

Using crowd-sourcing, users have contributed to social sensing, which simultaneously comes with an inevitable challenge [5]. Unlike natural environment monitoring, the expected data can be delivered in real time through automated sensors. In social perception systems, people acting as social sensors are uncontrolled. We need to extract valuable information from crowd-sourcing data and mine the expected knowledge. In this study, by taking advantage of social sensing, we develop a framework for analyzing the impacts of social events from geo-tagged social media data. Three dimensions (time, space, and semantics) are considered in the combined handling of machine learning, natural language processing, and visualization methods. On the basis of a case study using Chinese Weibo data, we demonstrate our approach, which includes three contributions:

(1) Utilizing social sensing, the proposed framework enables a combination of machine learning, natural language processing (NLP), and visualization methods to discover evidence from geo-tagged social media data that indicate the impacts of social events.
(2) A topic discovery method based on Latent Dirichlet Allocation (LDA) is applied, in which a topic coherence model is used to determine the optimal number of topics hidden in the event-related information, thus obtaining a more reasonable topic clustering.
(3) A social sentiment analysis is conducted to explore the impact of social events in terms of public sentiment. This analysis is based on a comparative assessment of the daily sentimental value of event-related information, event-irrelevant information, and comprehensive information.

With regard to the sections hereinafter, in Section 2, we give an overview of related work. Section 3 explains the computational model of studying event impact based on geo-tagged social media data. A case study is provided in Section 4, and research outcomes are discussed in Section 5. Finally, we conclude this paper and describe future work in Section 6.

2. Related Work

This section introduces, on the one hand, the research progress made by sociologists in applying various sociological methods to study social event impact and, on the other hand, the merits of geo-tagged social media data as well as the current research progress made in using geo-tagged social media data to understand social events.

2.1. Traditional Research on the Impact of Social Events

The impacts of various social events have typically been studied using traditional sociological methods (e.g., interviews, questionnaires and surveys, and documents and records). Regarding sports events, for example, Ohmann et al. studied the impact of the 2006 Football World Cup on the local residents through face-to-face interviews with 132 people [6]. Based on a questionnaire approach, Fredline and Faulkner used logistic regression analysis to interpret the influence indicators on residents' different attitudes towards local motor-racing events [7,8]. Using a social conflict analysis database,

Moreno and Miguel analyzed the short-term causal impact of sports events on social unrest events in Africa [9]. Using case studies, Scholtens and Peenstra [10] and Barreda et al. [11] analyzed the impact of sports events on the stock market and hotel industry, respectively. Regarding political aspects, Healy et al. collected original survey data during basketball championships to study the impact of irrelevant events on voters' evaluations of government performance [12]. Bariviera et al. studied the effect of geopolitical events on the oil market based on the price dynamics of West Texas Intermediate (WTI) [13]. Regarding public safety, Smith et al. analyzed public responses to the 11 September attacks through a random telephone survey of 2126 U.S. residents [14]. Arvanitidis et al. studied the impact of terrorist attacks on citizens' risk-perceptions based on European social surveys [15]. Regarding the tourism aspect, Breitsohl and Garrod used an online survey to examine people's cognitive reactions to an incident of unethical behavior at a tourist destination [16]. By means of questionnaires, Li et al. studied the public preferences for a post-event visit to the host city [17]. In recent years, the research potential of multi-sourcing social data has gradually attracted the attention of scholars. Cell phone mobility data was used to explore the relationship between various types of social events and participants' origins [18]. Various types of early network communication structures were used to learn and detect political abuse in social media [19]. In regard to political opinion mining, an opinion diffusion model was constructed to analyze and predict public reactions based on crowd social media content [20].

Current sociological studies on the impacts of social events mainly rely on sample surveys and in-depth interviews. It is irrefutable that the sample survey is the core methodological resource of sociology [21]. However, the increasing intensity of in-depth cooperations among sociology, computer science, and linguistics empowers us with new possibilities made possible by new data and methods for recognizing the impact of social events [22]. On the basis of widely used social media platforms, we can gather social data that is richer and is able to be used as a new avenue for understanding social events.

2.2. Event-Related Research Based on Geo-Tagged Social Media Data

Several key features of geo-tagged social media data make it feasible and valuable in lives monitoring event information. First, social media platforms like Twitter and China's Sina Weibo enable users to post about what has happened and share their knowledge. Second, users acting as smart social sensors are able to quickly report nearby events to others, thus raising attention to events. Third, social media data with spatial–temporal components can provide events with an explicit time–space description [23]. Fourth, given the broad geographic distribution and a huge number of users, events are able to be perceived efficiently. These features make geo-tagged social media data valuable for event study [24].

Crooks et al. assessed the potential of people to act as sensors for event monitoring and concluded that social media data could be used to enhance our situational awareness [25]. For event detection and tracking, some scholars integrated the temporal burstiness of terms or data volume and location information to find local events [26,27]. Adjusted classical probabilistic models combined with spatial and temporal features represent another frequently investigated approach of event detection [28,29]. Furthermore, some researchers have made use of time partitioning and spatial changes to track geo-events from geo-tagged social media data [30]. Regarding the extraction of event-related information, Murzintcev and Cheng proposed an automated process to collect event-related messages based on Hashtags, thus proving its availability by comparing it with other keyword-based methods [31]. Based on geo-tagged Flickr photos, Yan et al. proposed a workflow to discover spatiotemporal knowledge about post-disaster tourism recovery [32]. Yet another research topic dealt with the visualization of events. Nakaji and Yanai proposed a method of visualizing real-word events by showing event-related photos selected from geo-tagged social media [33]. Gao discussed different visualization methods to present the influenza activities in the United States [34].

As can be seen, geo-tagged social media data contains rich event information, and the current research work is mainly focused on when, where, and what happened, as well as related visualization methods. The immediate perceptions people have of the event tend to prevail, whereas their reflections on the concept, judgment, and inference of events remain less affected. In this study, we attempt to shift the research focus onto how social events have happened and how they affect our lives by taking advantage of widely accessible geo-tagged social media platforms and more advanced data-mining approaches.

3. Methods

The analysis framework described in Figure 1 includes four steps after social media data retrieval: data preprocessing, event-related information extraction, social event impact analysis, and visualization. Data preprocessing, which is described in Section 3.1, aims at improving the quality of data and adapting the data to the subsequent treatment. The extraction of event-related information from geo-tagged social media data is described in Section 3.2. The analysis of social event impact is divided into two parts: social sentiment analysis and social opinion mining. The former examines the public perception of events (Section 3.3), and the latter emphasizes the rational knowledge that people have of events (Section 3.4). Finally, visualization methods for data and analysis results are discussed in Section 3.5.
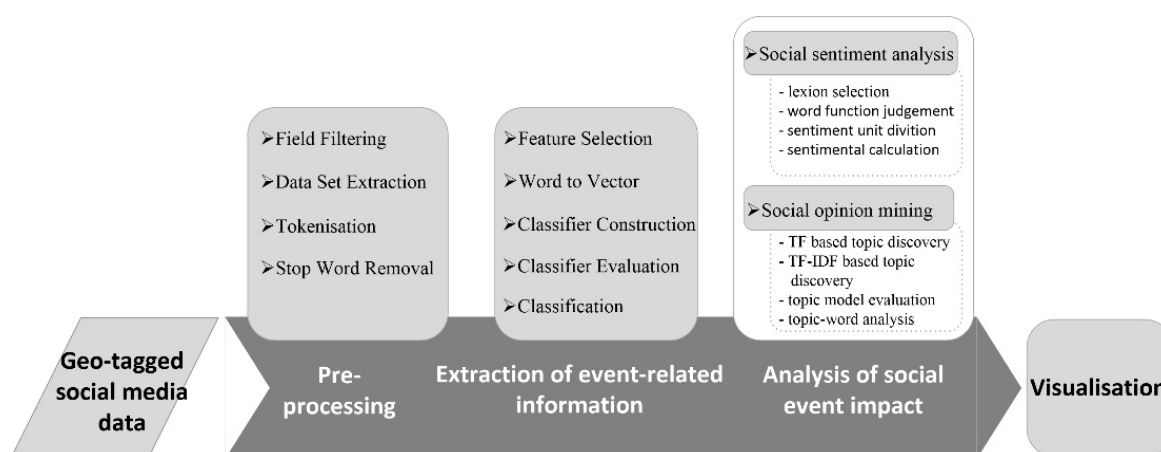

Figure 1. The analytical framework for event knowledge discovery.

3.1. Preprocessing

This section includes field filtering, dataset extraction, text tokenization, and stop word filtering. Bearing in mind that this study deals with the spatial, temporal, and semantic dimensions of geo-tagged social media data, related fields such as release time, location, and text messages are extracted from the raw data, and other unnecessary information are filtered away.

Regarding event-related information extraction, the hashtag is a good reference to label event-related information by the public. First, we manually select event-related hashtags as a reference to construct an event-related information dataset. Then, in order to extract more useful data, we approach this as a binary classification problem. Each record will be divided as being either event-related or event-irrelevant. In constructing the binary training dataset, the extracted dataset of event-related information based on hashtags will be the dataset for event-related training, whereas the dataset of event-irrelevant training may be selected from data posted with a similar background but before the event took place.

In order to facilitate the subsequent classification and mining of texts, a tokenization process is needed. Cohesive strings from social media posts should be split up into single words using word segmentation technology. Words in the English language are separated by spaces. However, in Chinese text, a semantic unit is often composed of one or more Chinese characters, and there is no

obvious separator between words. To solve this problem, a Python package used for Chinese text segmentation called "jieba" is used and performs well. By building a directed acyclic graph based on prefix dictionary structure, it can find potential word combinations efficiently. The most common and frequently occurring words lacking valuable information, which are called "stop words", are then excluded to reduce noise among the remaining tokens. In this study, we used the standard stop word list released by the Information Retrieval Laboratory of the Harbin Institute of Technology to remove stop words.

3.2. Extraction of Event-Related Information

Extracting unlabeled event-related information from geo-tagged social media datasets could be formulated as a binary classification problem and solved with machine learning methods. As shown in Figure 2, the whole process starts with the construction of training data, continues with feature selection and the construction and evaluation of classifiers, then terminates with the extraction of event-related information.
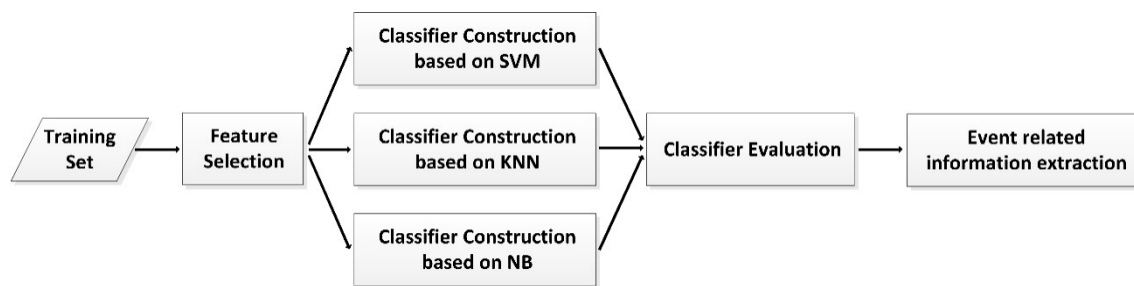

Figure 2. The workflow for the extraction of event-related information.

At first, two training datasets are constructed: an event-related dataset, and an event-irrelevant one. Hashtags serve as markers for an event-related training dataset, whereas a dataset posted under similar conditions but before the event happened is selected as an event-irrelevant training dataset. Subsequently, we apply the Vector Space Model (VSM), which uses the vectors of identifiers to represent the text documents. In addition, the term frequency (TF) is used as the semantic weighting factor.

A text-based binary classification is required in order to extract more event-related information from the original data. With respect to a specific event, it is usually difficult to obtain large labeled data for deep learning in a short period of time, so this paper considers three robust machine learning algorithms (Naive Bayes, the k-nearest-neighbors algorithms, and Support Vector Machine) as alternatives [35–38]. The Naive Bayes (NB) method is a probabilistic classifier based on Bayes' theorem and assuming conditional independence. The k-nearest-neighbors (KNN) algorithm is a non-parametric method with the k closest training examples in the feature space as input. Support Vector Machine (SVM) constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space to do classification, regression, or other tasks. These three methods are perhaps insufficient for complex classification tasks, but they perform well on the binary classification of texts. Hyperparameters tuning is indispensable to classifiers building, and cross validation is required to avoid overfitting. In the end, we use three common measures to compare the relative performance of the classifiers: precision, recall, and F1_score, as shown in Equations (1)–(3):

$$\text{Precision} = \frac{t_p}{t_p + f_p} \tag{1}$$

$$\text{Recall} = \frac{t_p}{t_p + f_n} \tag{2}$$

$$\text{F1\_score} = 2 \times \frac{\text{Precision·recall}}{\text{Precision} + \text{recall}} \tag{3}$$

where t_p is the number of correctly classified positive items, f_p is the number of wrongly classified positive items, and f_n is the number of wrongly classified negatives. Precision is the fraction of positive items among the retrieved items. Recall measures the proportion of actual positives that are correctly identified as such, and F1_score is the harmonic average of the precision and recall. Among the three aforementioned classifiers, the one with the best performance is adopted to extract event-related data for further analysis.

3.3. Social Sentiment Analysis

This section describes how to analyze changes in public sentiment based on social media messaging. The common sentiment analysis task is to extract sentiment polarity or intensity from text, facial expression, body movement, or music [39]. Regarding short text-based sentiment analysis, common methods can be categorized as supervised methods and lexicon-based methods. Since supervised approaches are usually domain-specific, while lexicon-based methods are more general, the lexicon-based methods are a better fit for our context.

We consider sentimental words, stop words, degree adverbs, and antonym words in analyzing the level of sentiment in the social media message. Three tasks are involved: judgment of the word function, construction of a sentimental unit, and calculation of the sentimental value. With regard to the first task, each word is determined to be a sentimental word, a degree word, or an antonym word. The sentimental units are then constructed according to the location of the sentimental word. Each sentimental unit contains a sentimental word as well as possible degree words and antonym words before it. Finally, we calculate the sentimental score for each sentimental unit and summarize all scores as the final sentimental score for the whole text, as shown in Equation (4):

$$P(T) \ = \ \sum_{i=1}^{n} P(U_i) \tag{4}$$

where $P(U_i)$ means the sentimental value of the ith sentimental unit, and $P(T)$ is the sentimental value of the text.

The degree wordlist is provided by HowNet (Chinese knowledge base) [40]. It has 219 words and is divided into six levels. Referring to [41], we assign these degree levels different weights in a descending order (2, 1.5, 1.25, 1.2, 0.8, 0.5) and attach the weights to the corresponding sentimental words. The degree factor γ for the ith sentimental word w_i is defined in Equation (5):

$$\gamma(w_i) = \prod_{k=1}^{m} d_{ki} \tag{5}$$

where d_{ki} is the weight of the kth degree word for the ith sentimental word.

Antonym words are important for the judgment of the sentimental polarity in a sentence. If an antonym word precedes a sentimental word, the semantics for the word will be reversed, which will affect the sentimental polarity of the whole sentence. Since there is not a standard Chinese antonym wordlist, we extracted 44 antonym words to construct such a wordlist based on several related research works [42,43]. The impact factor of antonym words τ on the sentimental words w_i is defined as:

$$\tau(w_i) = (-1)^n \tag{6}$$

where n is the number of antonym words in the ith sentimental unit. Taking all the above characteristics into consideration, we obtain the sentimental value of the ith sentimental unit:

$$P(U_i) = P(w_i) \times \gamma(w_i) \times \tau(w_i) \tag{7}$$

where $P(w_i)$ is the sentimental value of the ith sentimental word. The sentimental wordlist provided by the Dalian University of Technology has a more detailed intensity division for sentimental words [44].

It contains 27,466 sentimental words and each word has a polarity and a sentimental intensity. Sentimental intensity is divided into five levels (1,3,5,7,9) in an ascending order.

3.4. Social Opinion Mining

Social media users acting as smart social sensors are able to articulate their personal perception to social events in addition to relying on their own judgments and inferences to reveal the hidden relations between a focus event and other events. In this section, we will use the topic discovery model and the topic evaluation model to mine social opinions from social media messages. The analysis of the topic words obtained from topic discovery can assist us in discovering the attitude of the public and the implicit relationships among events.

3.4.1. Topic Model: LDA

The topic model is a type of statistical model used to discover the latent semantic structures of a text body. It helps us to organize and gain insights into large collections of unstructured text bodies [45]. Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) are the prevailing methods used for selecting representative terms in text collections [46,47]. In our experiments, we chose LDA since it is the generalization of PLSA and can create results to better explain the text semantics. LDA is a Bayesian graphical model and has three layers of "document–topic–word" [48]. The graphical model of LDA is shown in Figure 3. Its generative process is presented as follows:

- For each document m, pick a multinomial distribution ϑ_m from a Dirichlet distribution with parameter α;
- For each topic $k = z_{m,n}$, pick a multinomial distribution φ_k from a Dirichlet distribution with parameter β;
- For the nth word in the mth document where $m \in \{1, \cdots, M\}$, and $n \in \{1, \cdots, N_m\}$

  - Sample the word $w_{m,n} \sim Multinomial\left(\varphi_{z_{m,n}}\right)$
  - Sample the topic $z_{m,n} \sim Multinomial(\vartheta_m)$
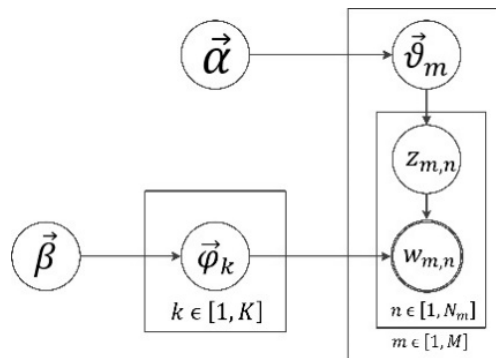

Figure 3. Graphical model of Latent Dirichlet Allocation (LDA).

The $w_{m,n}$ is the only observable variable, and other variables are latent variables. We can use Gibbs sampling [49], and the word-topic matrix ϑ and φ could be calculated as follows:

$$\vartheta_{k,t} = \frac{n_k^{(t)} + \beta_t}{\sum_{t=1}^{V}\left(n_k^{(t)} + \beta_t\right)} \tag{8}$$

$$\varphi_{m,k} = \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}\left(n_{m,-i}^{(k)} + \alpha_k\right)} \tag{9}$$

$$p(z_i = k \mid \vec{z}_{-i}, \vec{w}) \propto \frac{n_{m,-i}^{(k)} + \alpha_k}{\sum_{k=1}^{K}\left(n_{m,-i}^{(k)} + \alpha_k\right)} \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t=1}^{V}\left(n_{k,-i}^{(t)} + \beta_t\right)} \tag{10}$$

where $i = (m, n)$ indicates a two-dimensional subscript. $z_i = k$ is the assignment of the nth word in document m to topic k. $w_i = t$ is the nth word in document m and $-i$ means not including word w_i. $\vec{n}_m = \left(n_m^{(1)}, \cdots, n_m^{(k)}\right)$, $n_m^{(K)}$ demotes the number of words belonging to the kth topic in document m. $\vec{n}_k = \left(n_k^{(1)}, \cdots, n_k^{(V)}\right)$, $n_k^{(t)}$ demotes the number of word t generated by topic k.

3.4.2. Topic Evaluation Based on Topic Coherence

Regarding topic discovery based on LDA or PLSA, the number of topics needs to be predetermined. How to find the right number of latent topics in a given corpus, however, remains an open question. Among the proposed evaluation methods, those that are based on topic coherence have revealed a desirable performance [50].

Moreover, a topic evaluation model can be either intrinsic or extrinsic. Intrinsic methods do not use any external sources or tasks from the dataset, but for extrinsic tasks, a reference corpus is needed for creating a distributional semantic model. Since no good external Chinese domain datasets for sudden hot events currently exist, we chose the intrinsic measure UMass introduced by Newman 2010 [51] to evaluate the topic coherence and define an optimal number of topics. For each topic, computing the sum:

$$Coherence = \sum_{i<j} score\left(w_i, w_j\right) \tag{11}$$

of pairwise scores on the topic words $w_1, w_2, \cdots, w_n$, usually the top n words by frequency $p(w|k)$. The UMass measure is expressed as a pairwise score function:

$$score_{UMass}\left(w_i, w_j\right) = \log \frac{D\left(w_i, w_j\right) + 1}{D\left(w_i\right)} \tag{12}$$

$D(w_i)$ describes the count of documents containing the word w_i, $D\left(w_i, w_j\right)$ indicates the count of documents containing both words w_i and w_j. It is the empirical conditional log-probability $\log p\left(w_j | w_i\right) = \log \frac{p\left(w_i, w_j\right)}{p\left(w_j\right)}$ smoothed by adding one to $D\left(w_i, w_j\right)$.

3.5. Visualization

Visualization provides an effective means of data exploration and knowledge representation [52]. This section introduces some visualization methods for temporal, spatial, and textual patterns, which are suitable for demonstrating our analytical results.

Time is a necessary property of geo-tagged social media data. Charts, such as stacked graphs and bar charts, are a good conventional method of visualizing linear time. A clock-like time axis could be adopted to emphasize the cyclic character of time. Color or connection is generally suitable for the interpretation of the relative time of geo-tagged social media data, whereas the axis-based design could be used to present absolute time.

With regard to locations, point-based visualization helps to endow individual points within the spatial context. Each point represents an object, and its visual variables (e.g., color, size) carry related information. Heatmap is a good practice in expressing spatial hot spots that result from clustering. Line-based visualization turns the discrete points into a fitted curve and can also be used to depict locations along trajectories. Region-based visualization is a good way to depict the aggregated information over regions of a predefined granularity. For spatio-temporal visualization, space-time cube or time-series snapshots could be used to express the dynamic spatial phenomenon following specific temporal sequences.

Regarding text visualization, tag cloud is a good choice to represent the word frequency in the text. Word networks could be used to reflect the internal structure and semantic relationship in the texts. For example, the contextual relationships of words can be illustrated in a suffix tree, while the hierarchical relationship of the topics can be packed in circles [53].

4. Case Study and Results

The proposed method has been implemented and tested using Sina Weibo data in Shanghai to gain insight into a particular event—the Shanghai Stampede Tragedy that happened on New Year's Eve of 2015.

4.1. Data

The stampede tragedy, which occurred in Shanghai, began at about 23:35 local time and lasted almost 15 min. In total, 36 people were killed and 49 were injured [54]. The research data used herein was collected via the Chinese social media platform Sina Weibo. The official application programming interface was used to access and download public content from Sina Weibo. The detailed process of data acquisition was described in [55]. The whole dataset includes Weibo messages recorded from 29 December, 2013 to 8 April, 2015—11,784,344 records over Shanghai. Each record includes 33 attribute fields. Only location-, text-, and time-related information are used for analysis in the current study.

4.2. Results

4.2.1. Extraction of Event-Related Information

During the preprocessing of Weibo data, five related fields (latitude, longitude, text, created time, and created Unix timestamp) were preserved and other fields were excluded. After data preprocessing, two training datasets were constructed. Since the stampede event occurred during the 2015 New Year's Eve holiday, we extracted data released from Sina Weibo during the month following the event on the basis of hashtags (#Shanghai The Bund Stampede Event (上海外滩踩踏事件)#,# The Bund Stampede (外滩踩踏)#,# Shanghai The Bund New Year's Eve Stampede (上海外滩跨年踩踏)#,# The Bund New Year's Eve Stampede Event (外滩跨年踩踏事件)#) as the event-related dataset. Correspondingly, Weibo data posted on New Year's Day of 2014, a year before the stampede, were selected to construct the event-irrelevant dataset. Table 1 summarizes two datasets.

Table 1. Test datasets.

	The Number of Records	Date	Description
Event-related dataset	2093	January 2015	Posted with event-related hashtags
Event-irrelevant dataset	61,081	1 January 2014	Posted in the New Year 2014

For text-based binary classification, term frequency–inverse document frequency (TF-IDF) features were used, and five commonly classical methods were considered—namely MultinomialNB, BernoulliNB, KNN, SVM with linear kernel, and SVM with radial basis function (RBF) kernel. In order to avoid over-fitting, we used grid-search with 5-fold cross validation on the whole dataset to obtain the optimal hyperparameters for each classifier. Then, in order to select the optimal classifier from the five candidate classifiers, we selected 90% of the dataset for training and 10% of the dataset for testing with a fixed random state (random state = 100). The optimal hyperparameters and classifiers evaluation are summarized in Table 2. This whole process was implemented by using a scikit-learn (a Python module for machine learning) [56].

Table 2. The optimal hyperparameters and evaluation of classification methods.

Method	Parameters	Precision	Recall	F1_Score
MultinomialNB	alpha = 0.5	1.000	0.685	0.813
BernoulliNB	alpha = 0.001	0.995	0.972	0.984
KNN	K = 3, weights = Euclidean distance	0.972	0.319	0.481
SVM (Linear Kernel)	C = 1	0.991	0.995	0.993
SVM (RBF Kernel)	C = 100, gamma = 0.01	0.991	0.995	0.993

As shown in Table 2, although KNN and MultinomialNB achieve high precision, the recall scores and F1_scores are relatively low. BernoulliNB and SVM perform well, while both SVM methods achieve the best performance at the same time. This result fits the characteristics of the SVM explained in [57]. Considering that SVM with linear kernel has fewer parameters and faster computing speed, we selected SVM with linear kernel to classify three months of Weibo data after the event occurred, and 2425 event-related data records were obtained, as seen in Figure 4. We found that these data were mainly posted during the first week after the event. The first day of Weibo data posted up to 1729, and then, from the second day on, the attention to the event gradually decreased. After one week, little attention was paid to this event, but on the 13th day and the 21st day, the concern increased significantly. On the 13th day, a related news item was released, according to which leaders of the Huangpu district abused public funds to eat and drink in a nearby restaurant when the stampede happened. On the 21st day, the Shanghai municipal government announced the investigation results of the stampede. Obviously, these two related events were the main reasons for the increased attention to the tragic event.
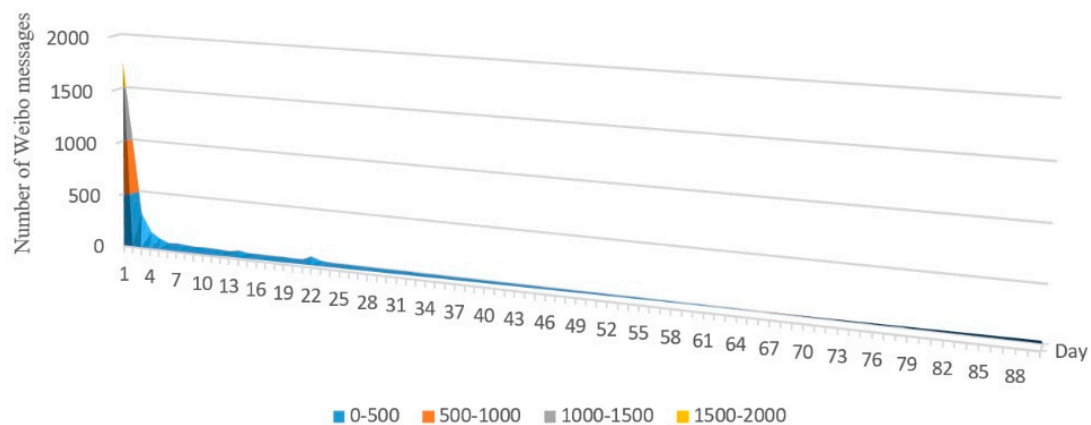

Figure 4. The number of relevant Weibo messages three months after the stampede event.

Figure 5 illustrates the Weibo numbers of the 48 h after the event. We can see that although this event happened at midnight, it stirred up some attention in the first few hours. A number of Weibo messages were posted from seven o'clock in the morning, peaking at nine o'clock. Early morning is a period during which all kinds of news travels fast. After 9 o'clock, attention to the event slowly decreased and reached the minimum at four o'clock the next morning before another wave of attention ensued.
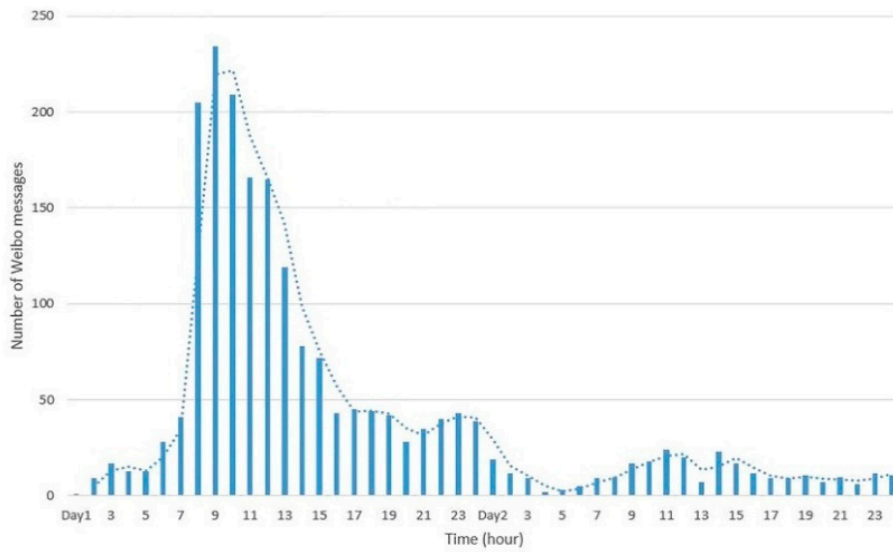
Figure 5. The number of related Weibo messages in the 48 h after the stampede event.

In Figure 6, a time series of snapshots shows the evolution of Weibo messages. A 3D spatial–temporal dot map in Figure 7 provides an alternative view of the distribution of event-related Weibo records in the first 24 h. The black spot indicates the location of the stampede event. As can be seen from these two figures, a few Weibo data were posted near the event location in the first few hours, and the distribution became slowly dispersed. Seven hours later, people started to wake up and Weibo data grew rapidly and spread out. Twelve hours later, the news spread to almost the entire Shanghai city, causing the continued spread of the information over the whole area. Comparing the distributions from one week later to three months later, we find almost no further change.
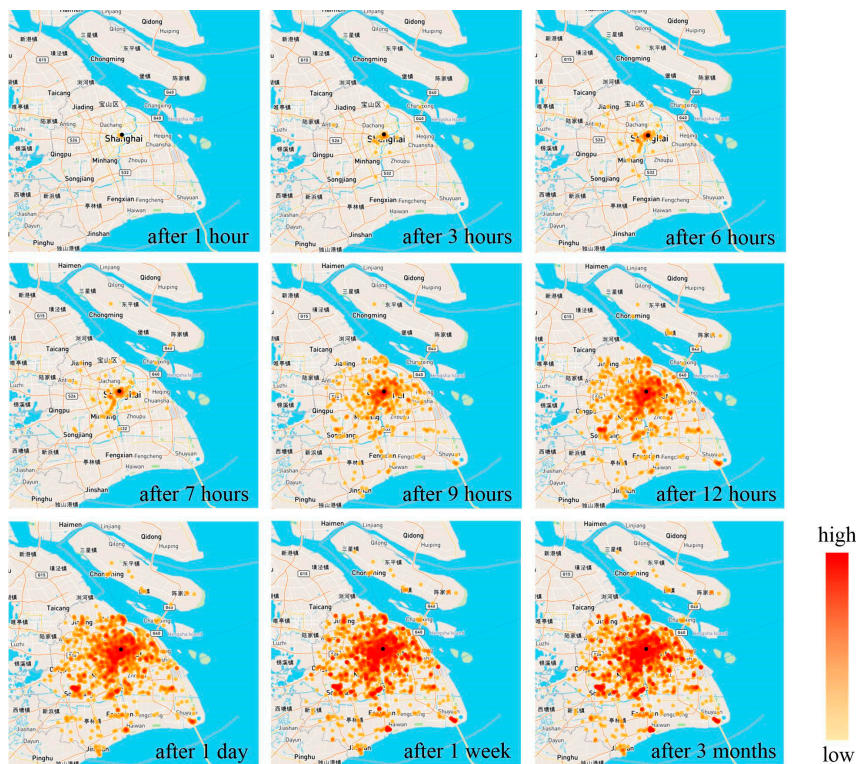

Figure 6. A time series showing the distribution of event-related Weibo records.

We used the torque-aggregation-function (torque-resolution = 2) provided by CartoDB to aggregate the geo-location of the event-related Weibo data [58]. The aggregation result is visualized as a heatmap in Figure 8. We can see clearly some hotspots (i.e., yellow areas) distributed in different areas of the city. Then, we used the density-based spatial clustering of applications with noise (DBSCAN) algorithm to select the most clustered areas. Based on the knowledge about the size of the Shanghai area and the number of event-related data, we tried with several different parameters and used one percent of the total number of event-related data (i.e., 24) as the minimum features per cluster and 1 km as the search distance. The subareas of the seven largest clusters are obtained and enlarged in Figure 8. Subarea 1 has the highest concentration of data and contains a black spot indicating the location of the stampede event. Subareas 2 and 7 are residential areas, and subarea 5 contains the Shanghai Hongqiao International Airport. The remaining three subareas are campus regions of universities and student dormitories. Students and intellectual groups tended to be more concerned about the event than other citizens.
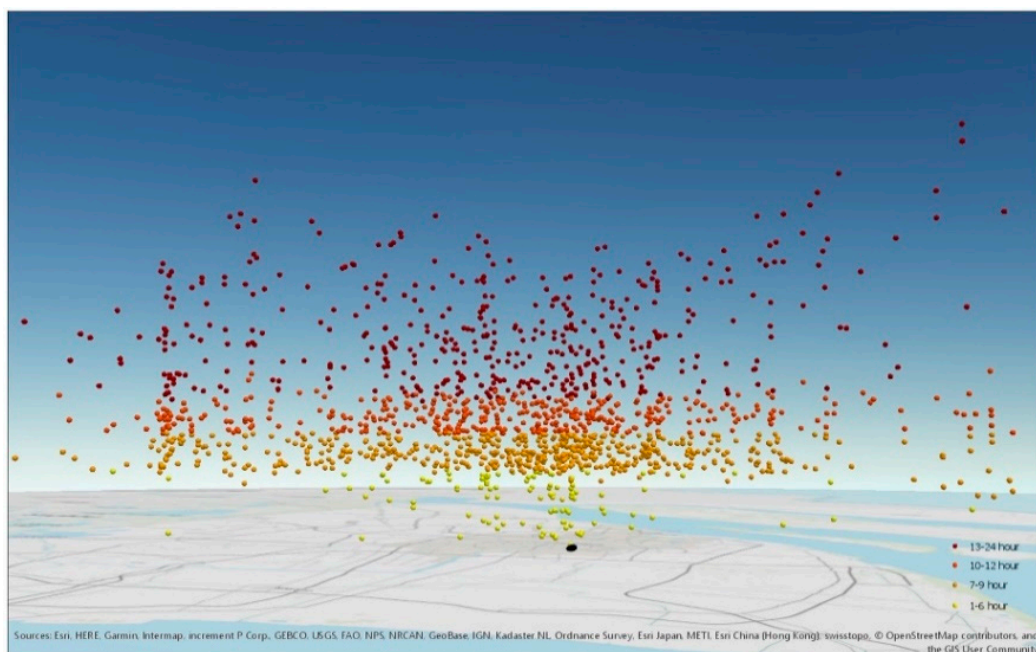

Figure 7. A spatial–temporal dot map of event-related Weibo records in the first 24 h.
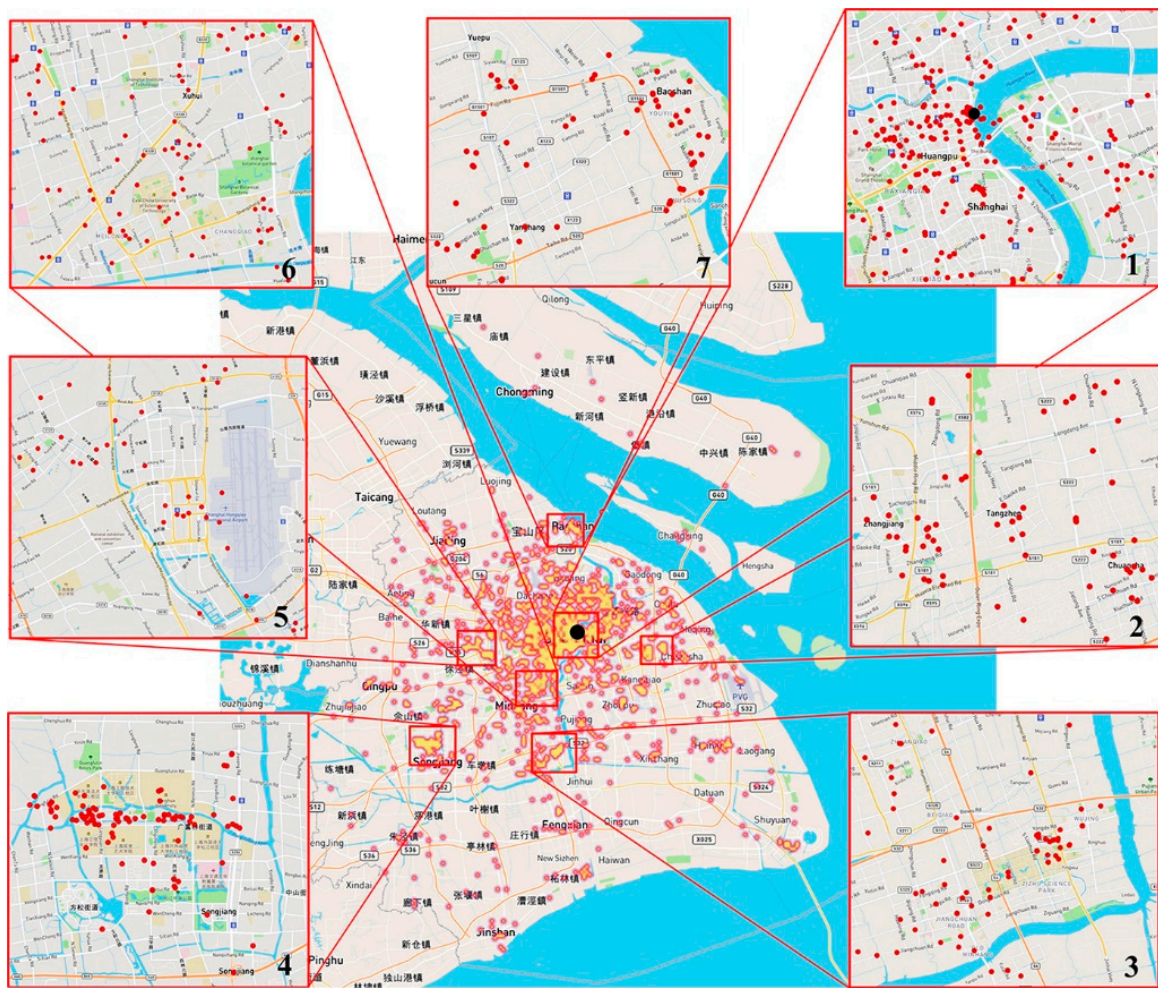
Figure 8. Spatial cluster analysis of event-related information.

4.2.2. Social Sentiment Analysis

Figure 9 illustrates the results of our statistical analysis on the daily average sentimental changes in different datasets around the week before and after the incident. Blue bars represent the daily average sentimental value of the event-irrelevant information, red bars indicate the daily average sentimental value of all Weibo messages, and gray bars indicate the daily average sentimental value of the event-related information. Some fluctuation of the daily average sentiment in all Weibo data can be perceived. The daily average sentimental value declined gradually from 25 December and went up after 28 December. It peaked on 31 December but dropped back in the following days. A plausible reason for this change is that 25 December is Christmas day, so the public mood is slightly lighter. By the last day of the year, the public's mood peaks due to the coming new year. As can be seen, the Shanghai public's sentiment is very positive in daily life.
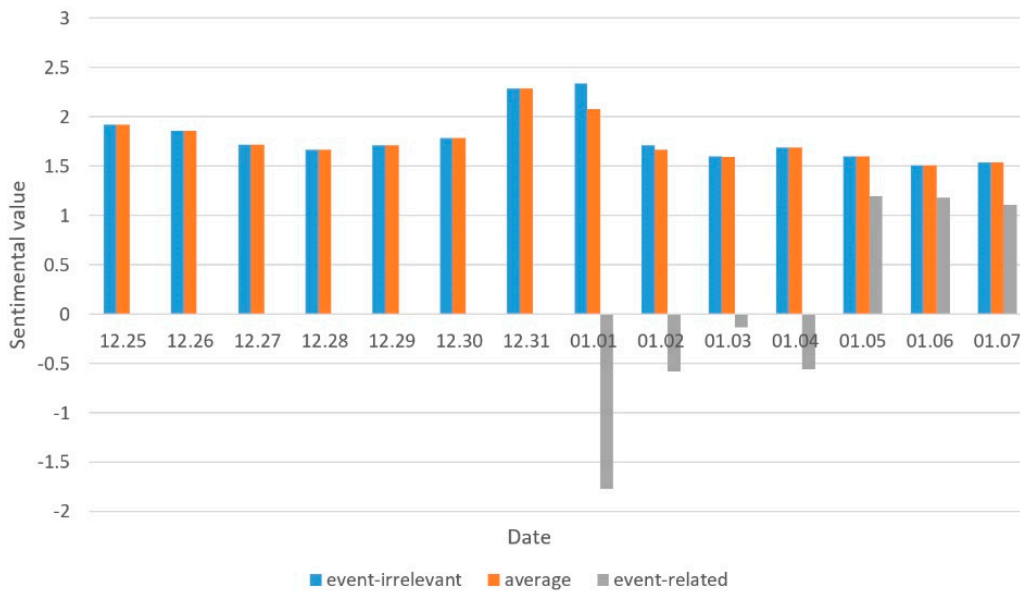
Figure 9. Daily average sentimental values in three datasets.

Since the stampede took place on New Year's Eve, the average daily sentimental value in the event-related dataset was very low on 1 January, reaching around −1.8; the average daily sentimental value of event-irrelevant data and comprehensive data are more than two. Later, as time went on, it gradually eased up and became slightly positive on the fifth day after the event, but remained below the average daily sentimental value of comprehensive data.

The daily average sentimental value of event-irrelevant data and that of comprehensive data were equal before the event. However, on the first day after the event, the daily average sentimental value of comprehensive data was 11% lower than that of event-irrelevant data, 2% lower on the second day, and 0.6% lower on the third day. Then, public sentiment was slowly recovered to the level prior to the event. This indicates, to some extent, that people's knowledge of social events had undergone a process from perceptual knowledge to rational knowledge. The comparative analysis of the sentiment changes can support us in understanding the impact of events on the public mood, which may be used as the basis for judging the happiness of the public.

Based on the sentimental intensity and the locations where Weibo data were posted, a sentimental map was created, as shown in Figure 10, where black, red, and blue, represent negative, positive, and neutral sentiment, respectively; the color intensity corresponds to the intensity of the sentimental feeling. There are obviously more black spots than red spots, which indicates a very negative impact of the stampede event on the public. As for the positive text content, we can see from the following social opinion mining that this is mainly comprised of the public expressing blessings for the victims.
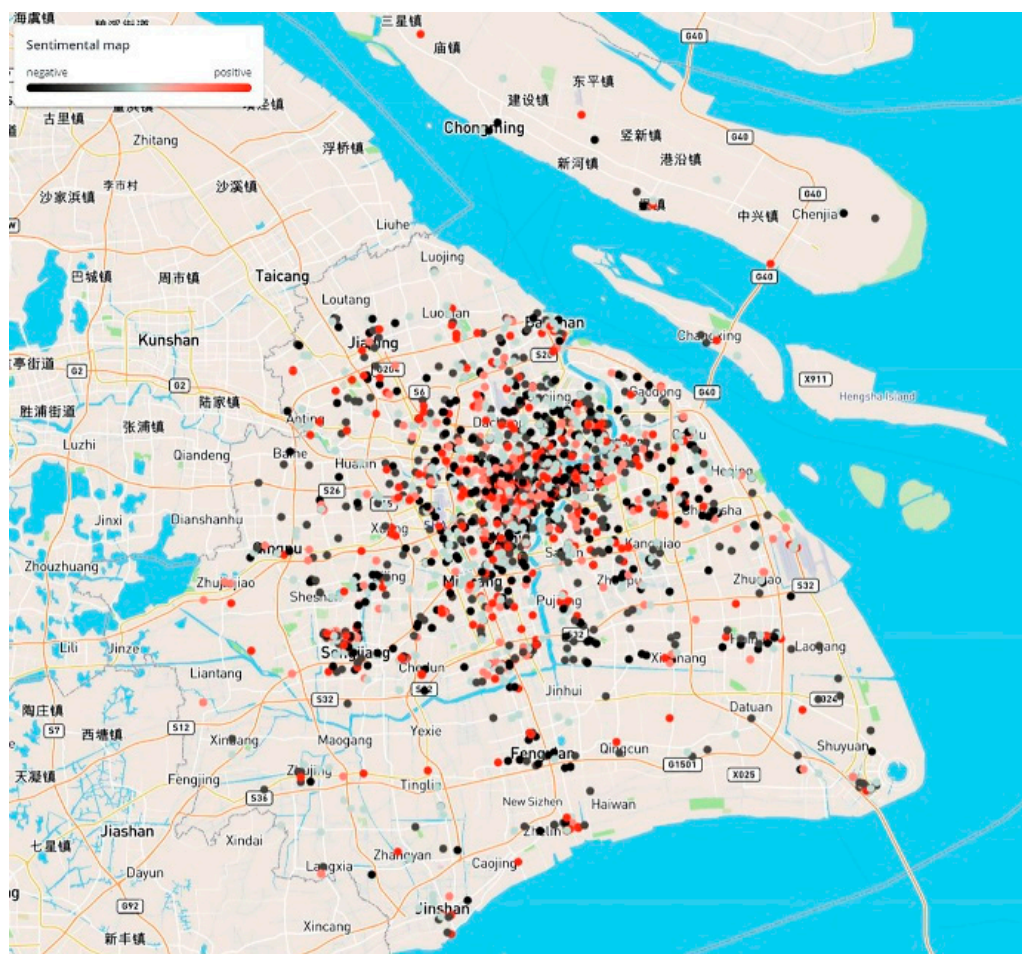
Figure 10. Sentimental map of the 2014 Shanghai stampede event.

4.2.3. Social Opinion Mining

Based on the term frequency (TF), we performed the LDA calculation of event-related information with topic numbers from one to 20. The 15 most frequently occurring words for each topic were selected and the UMass measure was adopted for the topic evaluation. The result is shown in Figure 11a. One topic is outstanding in comparison to other event-related topics. The LDA model was run with 50 iterations for this outstanding topic. The results for the 15 most frequently occurring words on the topic are summarized in Figure 12, where the font size represents the probability of the word appearing on the topic. Topic words such as "stampede", "event", and "accident" described the type of event. "Shanghai" and "the Bund" indicate the location of the event. "New Year's Eve" reveals the time of the event. "Candle", "life", "silence", and "the deceased" express people's wishes for the dead. All these keywords reflect the characteristics of the event. They belong to the perceptual knowledge about the event including human feelings and representation of the event.
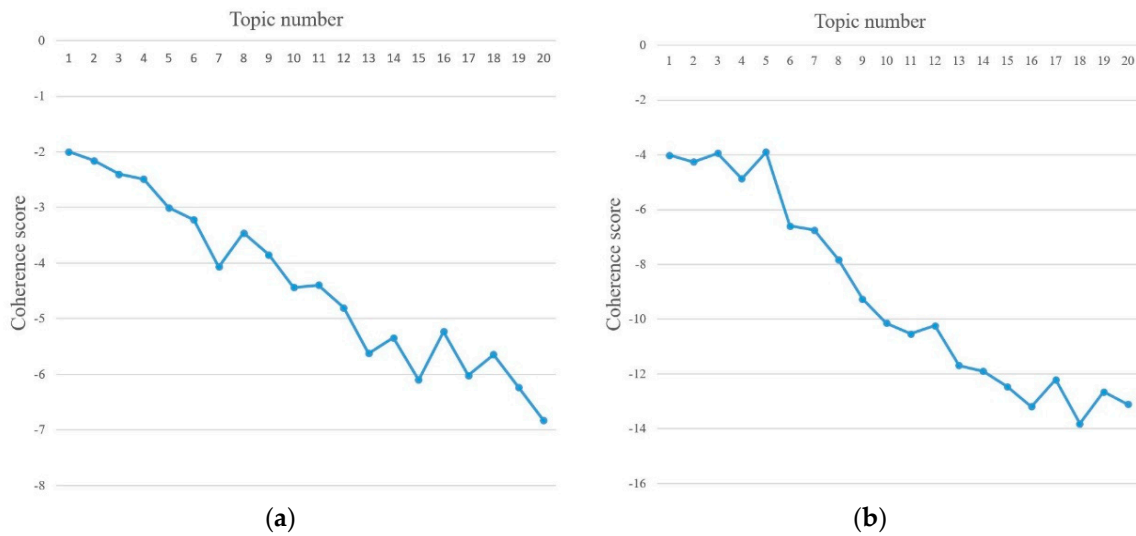
Figure 11. (**a**) Topic evaluation diagram based on the term frequency (TF); (**b**) topic evaluation diagram based on the term frequency–inverse document frequency (TF-IDF).

Figure 12. The weight chart of topic words.

In order to mine further hidden knowledge, TF-IDF is used as the word feature. This method decreases the weight of words with high frequencies in all documents, such as "stampede" and "Shanghai", and increases the weight of words which appear frequently in parts of the documents but not the whole records. In this way, some hidden knowledge could be discovered. The whole process is as above, except that the word factor changes from TF to TF-IDF. The result of topic evaluation for different topic numbers is shown in Figure 11b, where five outstanding topics for event-related information are perceivable. We used the LDA model with 50 iterations for these five topics and obtained the results. Each Weibo text has a corresponding probability for each topic. Based on the maximum probability, we divided each Weibo message into one topic and calculated the average sentimental value of the Weibo messages included in each topic. The analysis results regarding these five topics are shown in Table 3.

Topic	Topic Word Content	Interpretation	Relationship	Sentimental Value
1	candle, event, report, event investigation, happened, Shanghai, people, leader, accident, that night, all, officers, went, eat, restaurant	Local officers ate dinner at nearby restaurants during the night of the incident	Related event	−0.85
2	report, event investigation, survey result, event, candle, strong, cancel, light show, crowd, happened, that night, people, Shanghai, all, the deceased,	Because of the stampede event, the light show was canceled in case of another security incident.	Causal relationship	−1.87
3	survey result, leader, people, restaurant, event, die, Shanghai, district mayor, dismissed, go, Huangpu district, courage, happened, accident, the Bund,	The government handled the incident, dismissed the mayor, this may be related to a previous restaurant event	Causal relationship	−3.20
4	accident, Shanghai, the Bund, people, event, center, crowd, happened, mad, share, quick, come to see, candle, view, all	1. Overcrowding caused a stampede 2. The public spreads event-related information through social media	1. Reason 2. Public concentration	−0.55
5	candle, event, people, accident, Shanghai, all, go, the deceased, rest in peace, life, happened, the victim, strong, the Bund, silence	The reaction of the people to the incident, wishing that the dead rest in peace, hoping people can be strong	Public reaction to death people	−1.14

As shown in Table 3, excluding the description of the event itself, we can obtain some interesting events or knowledge from these topic words separately. These results include both the public attitude and people's judgments and inferences about this stampede event.

In topic one, "event investigation", "that night", "officers", "have dinner", and "restaurant" show that the public were aware of a restaurant event which officers attended that night. Official reports show that some local officers used public funds to dine at a nearby restaurant that night, which caused public dissatisfaction. The stampede event also connected the public's attention with this restaurant event.

In topic two, "crowd" shows the reason for this stampede event. "Cancel", "light show", and "crowd" illustrate that, possibly due to concerns related to crowding, the lantern show was canceled. This is a causal event.

In topic three, "district mayor", "dismissed", "Huangpu district", "restaurant", and "event" show that the local mayor was dismissed, which may be due to this event and the restaurant event. This should be a causal event.

In topic four, "people" and "crowd" describe the cause of the stampede. Additionally, "share", "quick", "come to see", and "view" are some words used by the public to spread news of this stampede event to others. This topic shows the public dissemination of this stampede event.

In topic five, "the deceased" and "victim" refer to death in this stampede event. "Rest in peace", "life", "strong", and "silence" are the reactions of the public to these victims. This topic shows the public's wish of peace for the victims of the incident.

It can be seen that the average sentimental value for each topic is negative. This result further illustrates the negative impact of this stampede tragedy on the public. Meanwhile, we can see that the average sentimental values of different topics are significantly different. This implies that the public's emotional responses to different topics on the same event are different. On topic 4, the public did not show strong negative emotions when discussing the cause of the incident and disseminating information about the incident. However, on topic 3, when the discussion involved the government's punishment of the relevant officials for the stampede tragedy and the related restaurant event, the public expressed strong negative emotions.

5. Discussion

It is easy to extract event-related information according to related hashtags for events that receive a high level of attention or have a great impact. Together with event-irrelevant information, we tuned the hyperparameters and performed cross-validation to select the optimal classifier from several commonly used classification methods to obtain other unlabeled related data. Both methods based on SVM achieved the best performance at the same time, while KNN and MultinomialNB did not perform well for the supervised learning of small-scale sparse matrices. However, for an event without a hashtag, we can refer to [34] to filter the data through keywords, and an event-related dataset could then be constructed through manual annotation.

As seen in Figures 4–7, event information can spread rapidly through the internet. As the 2014 Shanghai stampede occurred at midnight and most people were sleeping at the time, the event-related information spread slowly in the first few hours until most people gradually woke up and communication of the event proliferated rapidly. It took just two hours for the event-related information to cover the entire Shanghai area, indicating both the wide reach of the internet and the great impact of this event. As can be seen from Figures 6 and 7, the impact on the everyday lives of the public caused by the event is centered on the site of the incident and continues to spread outwards in space. To a certain extent, this reflects the law of network transmission of emergencies, which is spatially diffused from the incident site to the outside.

A comparative analysis of the average daily sentimental value of the event-related information, event-irrelevant information, and comprehensive information could be used to assess the impact of the event on the public. As shown in Figures 4 and 9, on the first day after the stampede event, the average daily public sentimental value dropped by 11% due to this event. As time went on, the public sentiment gradually regressed and concern for this event declined. After about one week, the public sentiment returned to the normal state. The comprehensive statistical analysis of different events can help the government and related agencies to assess and predict the public sentimental changes and the recovery cycle so that the government can formulate timely targeted response policies if a similar situation reappears.

As shown in Figure 12 and Table 3, using the LDA model based on TF, we can extract a few keywords to summarize the social event. In addition, using the LDA model based on TF-IDF, we can determine the public reaction and related hidden events. However, implicit knowledge needs to be obtained by parsing the topic words of each topic. The topic evaluation model based on the coherence score can help us determine the appropriate number of topics. In addition, the combination of topic modeling and sentiment analysis helps to understand the public's sentimental attitudes toward different topics on the same event. In response to the Shanghai stampede, the public's sentimental attitude towards each topic was negative. With respect to more controversial events (e.g., policy development, marketing activities), this approach can help deepen the understanding of diverse views held by the public.

6. Conclusions

In summary, this article presents a generic framework for analyzing the impacts of social events from geo-tagged social media. Following a discussion of the necessary data preprocessing, an event-related information extraction method combining machine learning and hashtags is adopted to extract unlabeled event-related information from the original dataset. Social sentiment analysis and social opinion mining are then used to explore the public's understanding of and feelings about the event. Finally, various visualization methods are discussed in order to represent the event-related knowledge obtained from the perspective of time, space, and semantics. The case study based on the 2014 Shanghai stampede tragedy revealed why it happened, in what way in affected local citizens, and for how long. Compared with traditional sociological methods, the proposed approach combines the methods of social sensing, machine learning, and natural language processing and can thus facilitate efficient data acquisition and analysis [6–17]. Compared with some of the existing social network-based

public opinion mining, this paper considers the dimensions of time and geo-space in observing the impact of social events in addition to sentiment analysis and opinion mining [19,20]. Meanwhile, various visualization methods make knowledge representation more vivid. The proposed analytical method could be easily adapted to various social events (e.g., riots, festivals, exhibitions) and could serve governmental agencies, enterprises, and citizens in tracing the consequences of events as well as taking necessary measures to amplify/reduce the positive/negative impacts.

Our study has some limitations. Firstly, due to the limited accessibility of Weibo data, we may have missed some knowledge aspects in our case study. Secondly, with respect to the sentimental analysis of social media messages, we have not yet included the network vocabulary in order to enhance analytical quality. Thirdly, we are restricted to intrinsic measures of evaluating topic coherence due to the lack of an external Chinese dataset for sudden events. More reliable results can be achieved when external reference datasets are available and accessible.

In the current research, we chose the 2014 Shanghai stampede as a representative case and used geo-tagged Weibo data in Shanghai region to study its social impacts. In the next step, we plan to extend this framework to conduct a comparative analysis between different kinds of events (e.g., political events, festivals, social security events) of varying scales (e.g., international, national, and city level) in order to explore the spatiotemporal behaviors of existing events and their cross-cultural impacts.

Author Contributions: Conceptualization, R.Z.; Data curation, M.J.; Methodology, R.Z. and D.L.; Supervision, L.M.; Visualization, R.Z. and C.Z.; Writing—original draft, R.Z.; Writing—review and editing, D.L., M.J., L.D. and L.M.

Abbreviations

NLP	Natural Language Processing
WTI	West Texas Intermediate
VSM	Vector Space Model
TF	Term Frequency
SVM	Support Vector Machine
NB	Naive Bayes
MultinomialNB	Naive Bayes classifier for multinomial models
BernoulliNB	Naive Bayes classifier for multivariate Bernoulli models
RBF	Radial Basis Function
KNN	K-Nearest Neighbors
PLSA	Probabilistic Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
TF-IDF	Term Frequency-Inverse Document Frequency
DBSCAN	Density-Based Spatial Clustering of Applications with Noise

References

1. Getz, D.; Page, S.J. *Event Studies: Theory, Research and Policy for Planned Events*, 3rd ed.; Routledge: London, UK, 2016; pp. 1–21, ISBN 978-1-138-89916-2.
2. Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social sensing: A new approach to understanding our socioeconomic environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [CrossRef]
3. Shi, Z.Z. *Intelligence Science*, 2nd ed.; World Scientific Publishing Company: Singapore, 2012; pp. 135–166, ISBN 981-4360-77-5.

4.   Valkanas, G.; Gunopulos, D. How the live web feels about events. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, San Francisco, CA, USA, 27 October–1 November 2013; pp. 639–648.

5.   Ali, R.; Solis, C.; Salehie, M.; Omoronyia, I.; Nuseibeh, B.; Maalej, W. Social sensing: When users become monitors. In Proceedings of the 19th ACM SIGSOFT symposium and the 13th European Conference on Foundations of Software Engineering, Szeged, Hungary, 5–9 September 2011; ACM: New York, NY, USA, 2011; pp. 476–479.

6.   Ohmann, S.; Jones, I.; Wilkes, K. The perceived social impacts of the 2006 Football World Cup on Munich residents. *J. Sport Tour.* **2006**, *11*, 129–152. [CrossRef]

7.   Fredline, E.; Faulkner, B. Variations in residents' reactions to major motorsport events: Why residents perceive the impacts of events differently. *Event Manag.* **2001**, *7*, 115–125. [CrossRef]

8.   Fredline, E.; Faulkner, B. Residents' reactions to the staging of major motorsport events within their communities: A cluster analysis. *Event Manag.* **2001**, *7*, 103–114. [CrossRef]

9.   José Miguel, P.M. Do Football Victories Affect Social Unrest? Evidence from Africa. Master's Thesis, Pontificia Universidad Católica de Chile, Santiago, Chile, 2017.

10. Scholtens, B.; Peenstra, W. Scoring on the stock exchange? The effect of football matches on stock market returns: An event study. *Appl. Econ.* **2010**, *41*, 3231–3237. [CrossRef]

11. Barreda, A.A.; Zubieta, S.; Chen, H.; Cassilha, M.; Kageyama, Y. Evaluating the impact of mega-sporting events on hotel pricing strategies: The case of the 2014 FIFA World Cup. *Tour. Rev.* **2014**, *72*, 184–208. [CrossRef]

12. Healy, A.J.; Malhotra, N.; Mo, C.H. Irrelevant events affect voters' evaluations of government performance. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 12804–12809. [CrossRef] [PubMed]

13. Crude Oil Market and Geopolitical Events: An Analysis Based on Information-Theory-Based Quantifiers. Available online: https://arxiv.org/abs/1704.04442 (accessed on 20 February 2018).

14. America Rebounds: A National Study of Public Response to the September 11th Terrorist Attacks. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.458.6605&rep=rep1&type=pdf (accessed on 20 November 2017).

15. Arvanitidis, P.; Economou, A.; Kollias, C. Terrorism's effects on social capital in European countries. *Public Choice* **2016**, *169*, 231–250. [CrossRef]

16. Breitsohl, J.; Garrod, B. Assessing tourists' cognitive, emotional and behavioural reactions to an unethical destination incident. *Tour. Manag.* **2016**, *54*, 209–220. [CrossRef]

17. Li, H.; Song, W.; Collins, R. Post-event visits as the sources of marketing strategy sustainability: A conceptual model approach. *J. Bus. Econ. Manag.* **2014**, *15*, 74–95. [CrossRef]

18. Calabrese, F.; Pereira, F.C.; Di Lorenzo, G.; Liu, L.; Ratti, C. The geography of taste: Analyzing cell-phone mobility and social events. In Proceedings of the International Conference on Pervasive Computing, Helsinki, Finland, 17–20 May 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 22–37.

19. Ratkiewicz, J.; Conover, M.; Meiss, M.R.; Gonçalves, B.; Flammini, A.; Menczer, F. Detecting and tracking political abuse in social media. In Proceedings of the International Conference on Weblogs and Social Media, Barcelona, Spain, 17–21 July 2011; AAAI: Palo Alto, CA, USA, 2011; pp. 297–304.

20. Sobkowicz, P.; Kaschesky, M.; Bouchard, G. Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Gov. Inf. Q.* **2012**, *29*, 470–479. [CrossRef]

21. Goldthorpe, J.H. *On Sociology: Numbers, Narratives, and the Integration of Research and Theory*, 1st ed.; Oxford University Press: Oxford, UK, 2000; pp. 45–65, ISBN 978-0-19-829572-3.

22. Savage, M.; Burrows, R. The coming crisis of empirical sociology. *Sociology* **2007**, *41*, 885–899. [CrossRef]

23. Mobasheri, A.; Sun, Y.; Loos, L.; Ali, A.L. Are Crowdsourced datasets suitable for specialized routing services? Case study of OpenStreetMap for routing of people with limited mobility. *Sustainability* **2017**, *9*, 997. [CrossRef]

24. Goodchild, M.F.; Glennon, J.A. Crowdsourcing geographic information for disaster response: A research frontier. *Int. J. Digit. Earth* **2010**, *3*, 231–241. [CrossRef]

25. Crooks, A.; Croitoru, A.; Stefanidis, A.; Radzikowski, J. # Earthquake: Twitter as a distributed sensor system. *Trans. GIS* **2013**, *17*, 124–147. [CrossRef]

26. Krumm, J.; Horvitz, E. Eyewitness: Identifying local events via space-time signals in twitter feeds. In Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 3–6 November 2015; ACM: New York, NY, USA, 2015.

27. Sugitani, T.; Shirakawa, M.; Hara, T.; Nishio, S. Detecting local events by analyzing spatiotemporal locality of tweets. In Proceedings of the 27th International Conference on Advanced Information Networking and Applications Workshops, Barcelona, Spain, 25–28 March 2013; IEEE: New York, NY, USA, 2013; pp. 191–196.

28. Cheng, X.; Yan, X.; Lan, Y.; Guo, J. Btm: Topic modeling over short texts. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2928–2941. [CrossRef]

29. Zhang, C.; Liu, L.; Lei, D.; Yuan, Q.; Zhuang, H.; Hanratty, T.; Han, J. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, 13–17 August 2017; ACM: New York, NY, USA, 2017; pp. 595–604.

30. Zhou, X.; Xu, C. Tracing the spatial-temporal evolution of events based on social media data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 88. [CrossRef]

31. Murzintcev, N.; Cheng, C. Disaster Hashtags in Social Media. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 204. [CrossRef]

32. Yan, Y.; Eckle, M.; Kuo, C.L.; Herfort, B.; Fan, H.; Zipf, A. Monitoring and assessing post-disaster tourism recovery using geotagged social media data. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 144. [CrossRef]

33. Nakaji, Y.; Yanai, K. Visualization of real-world events with geotagged tweet photos. In Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Melbourne, VIC, Australia, 9–13 July 2012; IEEE: New York, NY, USA, 2012; pp. 272–277.

34. Gao, Y.; Wang, S.; Padmanabhan, A.; Yin, J.; Cao, G. Mapping spatiotemporal patterns of events using social media: A case study of influenza trends. *Int. J. Geogr. Inf. Sci.* **2018**, *32*, 425–449. [CrossRef]

35. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [CrossRef]

36. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185.

37. Yang, L.; MacEachren, A.M.; Mitra, P.; Onorati, T. Visually-Enabled Active Deep Learning for (Geo) Text and Image Classification: A Review. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 65. [CrossRef]

38. Feng, Y.; Sester, M. Extraction of pluvial flood relevant volunteered geographic information (VGI) by deep learning from user generated texts and photos. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 39. [CrossRef]

39. Cambria, E.; Schuller, B.; Xia, Y.; Havasi, C. New avenues in opinion mining and sentiment analysis. *IEEE Intell. Syst.* **2013**, *28*, 15–21. [CrossRef]

40. Dong, Z.; Dong, Q.; Hao, C. Hownet and its computation of meaning. In Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, 23–27 August 2010; ACL: Stroudsburg, PA, USA; pp. 53–56.

41. Wang, Y.; Zhang, S.X. Research of sentiment analysis for Chinese micro-blog topic. *J. Fuyang Norm. Univ. (Nat. Sci.)* **2017**, *34*, 50–56. (In Chinese)

42. Wen, Z. A Study on Negation in Modern Chinese. Ph.D. Thesis, Fudan University, Shanghai, China, 2003. (In Chinese)

43. Dang, L.; Zhang, L. Method of discriminant for Chinese sentence sentiment orientation based on HowNet. *Appl. Res. Comput.* **2010**, *27*, 1370–1372. (In Chinese)

44. Xu, L.; Lin, H.; Pan, Y.; Ren, H.; Chen, J. Constructing the affective lexicon ontology. *J. China Soc. Sci. Tech. Inf.* **2008**, *27*, 180–185. (In Chinese)

45. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]

46. Hofmann, T. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, 30 July–1 August 1999; Morgan Kaufmann Publishers: San Francisco, CA, USA, 1999; pp. 289–296.

47. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022. [CrossRef]

48. Griffiths, T.L.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101* (Suppl. 1), 5228–5235. [CrossRef] [PubMed]

49. Parameter Estimation for Text Analysis. Available online: http://www.arbylon.net/publications/text-est. pdf (accessed on 15 November 2017).

50. Röder, M.; Both, A.; Hinneburg, A. Exploring the space of topic coherence measures. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 31 January–6 February 2015; ACM: New York, NY, USA, 2015; pp. 399–408.

51. Newman, D.; Lau, J.H.; Grieser, K.; Baldwin, T. Automatic evaluation of topic coherence. In Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, CA, USA, 1–6 June 2010; ACL: Stroudsburg, PA, USA, 2010; pp. 100–108.

52. Zheng, Y.; Wu, W.; Chen, Y.; Qu, H.; Ni, L.M. Visual analytics in urban computing: An overview. *IEEE Trans. Big Data* **2016**, *2*, 276–296. [CrossRef]

53. Tang, J.; Liu, Z.; Sun, M. A Survey of Text Visualization. *J. Comput. Aided Des. Comput. Graph.* **2013**, *25*, 273–285. (In Chinese)

54. Investigation Report of 2014 Shanghai Stampede Event. Available online: http://www.shjcw.gov.cn/2015jjw/n2230/n2237/u1ai51007.html (accessed on 4 November 2017).

55. Jendryke, M.; Balz, T.; Liao, M. Big location-based social media messages from China's Sina Weibo network: Collection, storage, visualization, and potential ways of analysis. *Trans. GIS* **2017**, *21*, 825–834. [CrossRef]

56. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Vanderplas, J. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

57. Joachims, T. *Learning to Classify Text Using Support Vector Machines*, 1st ed.; Springer: Boston, MA, USA, 2002; pp. 45–74, ISBN 978-1-4613-5298-3.

58. How Spatial Aggregation Works. Available online: https://github.com/CartoDB/torque/wiki/How-spatial-aggregation-works (accessed on 2 November 2017).