


Analysis on an Auto Increment Detection System of Chinese Disaster Weibo Text


Hua Bai

(Fujian Normal University, Fuzhou, Fujian, China)

 <https://orcid.org/0000-0002-7761-6423>, baihua1727@163.com


Hualong Yu

(Harbin Institute of Technology, Harbin, Heilongjiang, China)

 <https://orcid.org/0000-0002-5873-2912>, yuhualong@hit.edu.cn

Guang Yu


(Harbin Institute of Technology, Harbin, Heilongjiang, China)

 <https://orcid.org/0000-0001-8794-8205>, yug@hit.edu.cn

Álvaro Rocha

(University of Lisbon, Lisbon, Portugal)

Corresponding Author

 <https://orcid.org/0000-0002-0750-8187>, amr@iseg.ulisboa.pt

Xing Huang

(Southwest University of Science and Technology, Mianyang, Sichuan, China)

huangxing6213@126.com

Abstract: With the rapid development of Internet information technology, the advantages of social media in terms of speed, content, form, and effect of communication are becoming increasingly significant. In recent years, more and more researchers have paid attention to the special value and role of social media tools in disaster information emergency management. Weibo is the most widely used Chinese social media tool. To effectively mine and apply the emergency function of disaster situation microblogs, a disaster situation information discovery and collection system capable of online incremental identification and collection are constructed for massive and disordered disaster microblog text streams. First, based on the deep learning-trained word vector model and a large-scale corpus, an unsupervised short-text feature representation method of disaster situation Weibo information is developed. According to the experimental results of the feature combination test and the training set scale test, the SVM algorithm was selected for disaster microblog information classification, which realized effective identification of disaster situation micro-bloggings. Then, the temporal information similarity and geographic information similarity are used to improve the single text similarity algorithm, and a Chinese disaster event online real-time detection model is constructed. Furthermore, the disaster-affected areas can be achieved in real-time based on the detection results. By crawling and classifying the micro-bloggings from the disaster-affected areas, it is possible to realize the incremental identification and collection of online disaster situation Weibo information. Finally, the empirical analysis of disaster events such as the “Leshan Earthquake” shows that the real-time intelligent identification and collection system for disaster situation Weibo micro-bloggings developed in this paper can obtain large-scale and useful data for disaster emergency management, which proving that this system is effective and efficient.

Keywords: Social media, Disasters, Emergency management, Word2vector

Category: H.3.5, I.2.7, I.7.5, I.5.4

DOI: 10.3897/jucs.65106

1 Introduction

In the process of disaster emergency management, information communication is a key part of disaster management, specifically in reducing the number of casualties and losses resulting from disasters [Houston et al., 2015]. During a disaster, infrastructure such as electricity and telecommunications is vulnerable to damage, but the Internet infrastructure is relatively solid and stable [Giovannini, 2020]. With the rapid development of social networking services, the role of Internet technology in disaster management has become increasingly significant. Social media became emerging Internet communication technologies, which contribute significantly to improving disaster communication processes influencing much information capacity, dependability, and interactivity [Stieglitz et al., 2018]. Due to the sharp increase in the number of users on social media platforms such as Twitter and Weibo, before and after disaster events, more and more users post information related to disaster events through social media platforms [Houston et al., 2015]. Compared with traditional media and other communication channels, social media using advanced Internet technology has become an important tool for people to receive and send information about the development of disaster events [Chen et al., 2017]. Social media is conducive to the implementation of crowdsourcing work models in donations and volunteer services during emergency management.

In recent years, with the continuous acceleration of climate warming and urbanization, various natural disasters have occurred frequently, resulting in serious casualties and property losses. Many Chinese scholars have also noticed this emerging research field, and have researched in the fields of the earthquake, extreme weather, rainstorm, etc., thus promoting Weibo, the most widely used Chinese social media platform, to serve the disaster warning process [Bai et al., 2020]. However, the current research about Chinese Weibo during disasters is focusing on using Weibo as a supplementary tool for official media, playing the role of Weibo's news reporting and notification, ignoring the two-way interactive characteristics of social media, and failing to play the important role of the user-generated content platform. To deeply exploring the disaster emergency management function of Chinese Weibo, the most important basis and the premise are to obtain sufficient Weibo data. However, Chinese is different from English, and the text structure and expression of Chinese Weibo are very different from English Twitter. So far, there is a shortage of short text research on Chinese microblogs during disasters, and related algorithms and automatic systems for the characteristics of disaster environments are needed to be developed.

Therefore, to improve and support the disaster emergency management process, three goals are focused in our research: (1) to explore the particularity of Chinese microblogs short texts in disaster situations, (2) to achieve online disaster event outbreak detection by Chinese microblogs, (3) to identify and collect a large number of disaster microblogs.

In the traditional text mining process, the Bag of Words model is widely used to construct the text feature space. This method has obvious limitations of sparse features and large noise when performing short text representation on Chinese social media. In order to overcome this limitation, this paper collects large-scale Weibo (a popular microblog tool in China) data, builds an initial corpus of about 3 million Weibo instances, and uses deep learning methods to perform unsupervised training on the

initial corpus to generate a Word2vec semantic model for the short text representation of social media text during disasters. A widely accepted method of acquiring event-related microblog information is to filter keywords on a noisy microblog text stream. However, information containing disaster keywords and phrases may not be related to disaster events. For example, when a user talks about an earthquake on Weibo, he may say the actual Leshan Earthquake occurred, or the Tangshan Earthquake (a famous movie). Therefore, it is necessary to establish a two-category model for the online microblog text and to identify the microblog information that is truly related to the disaster event from the microblog text containing the disaster keywords.

Thus, facing the post-disaster microblog text stream, according to the experimental results of feature combination and training set scale test on several classic classification algorithms such as support vector machines and random forests; a suitable Weibo text classification model was established to identify Weibo instances during disaster situation effectively. Then, time and geographic information carried by the Weibo instances are used to improve the single text similarity algorithm to build a new online Chinese disaster event detection model. This model can not only detect the outbreak of sudden disaster events but also output the outbreak time and affected area of the disaster quickly, which could be in favor of the following data collection and update. The Leshan Earthquake case research confirmed that the output of the above model could effectively serve the real-time and intelligent collection of disaster microblog information in the affected disaster area, which is conducive to providing a useful real-time data source for dynamic emergency response [Yang, A., 2020].

Currently, "Twicident", "Tweet4act", "Twitter Earthquake Detector" and other real-time monitoring systems for different needs are mostly developed based on the English Twitter platform. These platforms are impossible to conduct real-time detection and information collection of Chinese social media messages. In addition, many researchers only focus on static data sets when analyzing Chinese disaster microblogs. Chinese Weibo data is dynamic. Disaster management puts forward high requirements for the freshness of emergency information. Therefore, the online disaster real-time detection and disaster information collection method developed by us for the Chinese Weibo platform can effectively break the bottleneck of dynamic information collection in the disaster management process and provide effective information support for disaster emergency management decision-making.

The rest of this study is organized as follows. Section 2 illustrates the state of the art to expose the differences between this work and other seminal works. Section 3 explains the research methods employed in this study. Section 4 analyses the Auto Increment Identity System of Disaster Situation Information based on Weibo platform, including framework (Section 4.1), disasters Weibo text representation (Section 4.2), and disaster Weibo identification (Section 4.3), and disaster events real-time detection (Section 4.4). Section 5 conducts a case study taking the Leshan Earthquake as an example to discuss key findings of the above methods in detail. Section 6 concludes the paper and presents directions for future work.

2 State of the Art

2.1 Social Media in Disaster Situations

Previous case studies on disasters, such as the Asian tsunami and the Queensland flooding, revealed that social media provide a large amount of first-hand information for disaster relief and mitigation processes [Thelwall and Stuart, 2007]. As effective platforms for exchanging messages, social media play an important role in disaster management and can significantly improve disaster emergency management processes [Murphy, 2016].

Lasswell (1948) cited three functions of communication in his work, “Communication Process and Its Function in Society.” The first of these functions is environmental monitoring, which is considered the most important function of mass communication. The mass media constantly provides information to the public about different incidents, including upcoming and ongoing disaster events. The spread of low-cost Internet access has prompted many countries to develop Web-based disaster incident monitoring systems, such as the “Did You Feel It?” program of USGS [Atkinson and Wald, 2007].

Twitter is another social media platform with plenty of information related to natural disasters, such as the Wenchuan earthquake (China, 2008-04-29), the Los Angeles earthquake (USA, 2009-01-24), and the Morgan Hill earthquake (California, USA, 2009-03-30). Ian O’Neill and Michal explored the possibility of monitoring earthquakes by using Twitter. Earle [2010] also proved that social media could detect disasters much quicker compared with traditional methods. In the case of earthquakes, the transmission rate of messages on Twitter is significantly faster than that of seismic waves, thereby allowing potential victims to prepare to escape before the onset of the disaster [Allen, 2012]. Crooks et al. [2013] also considered social media superior to the “Do You Feel It?” program of USGS in terms of its speed and capacity of information dissemination.

Social media has also been used for detecting other disasters widely. The number of studies on using social media during emergencies, such as wildfires [De Longueville, 2009], flu outbreaks [Chew, 2010], storms [Neubauer et al., 2015], has recently increased. Chew [2010] found that during the 2009 H1N1 outbreak, people used Twitter to share information from credible sources, their personal experiences, and their opinions. De Longueville [2009] analyzed the temporal, spatial, and social dynamics of Twitter activity during a major forest fire event in Southern France in July 2009 and found that Twitter can act as a wildfire sensor that receives rich, free signals from the public. Neubauer [2015] studied the trend of the social media data flow from 2013 to 2014 and successfully detected the outbreak of the Egyptian storm events. Musaev [2014] found that using social media data was more effective than using satellite remote data in monitoring dammed lakes.

Many researchers have also produced excellent outputs in this research area. To achieve a real-time tracking and reporting of earthquake information, Sakaki et al. [2010, 2012], designed a real-time earthquake detecting system based on Twitter data and then improved this system to achieve a 96% detection rate. After launching “Did You Feel It?,” USGS developed a Twitter Earthquake Detector, an earthquake detection system that operates over a filtered Tweet stream and even outperforms the “Did You Feel It?” program [Earle et al., 2012]. CSIRO [Cameron et al., 2012] developed an

Emergency Situation Awareness platform in 2009 that analyses Twitter messages that are posted during disasters and crises. This platform uses natural language processing and data mining techniques to achieve early detection of events and to extract crowdsource relevant information about a disaster.

Several other systems that monitor disaster information in real-time by using social media include Twicident [Abel et al., 2012], Tweet4act [Chowdhury et al., 2013], CrisisTracker [Rogstadius et al., 2013], Ushahidi (<http://www.ushahidi.com/>), Twitter Earthquake Detector [Earle et al., 2012], Emergency Situation Awareness [Power, 2014], and EARS [Avvenuti et al., 2014].

Social media have the advantages of timeliness and interactivity, which are typical features of user-generated content. Therefore, social media platforms allow users to utilize their collective wisdom effectively and may provide disaster emergency responders with useful information that can help them in their work [Yates and Paquette, 2011]. Natural disasters often cause serious damage to humans and their homeland and may even result in a large number of casualties and homelessness. Therefore, providing timely and reliable information about emergency services [Bird, 2012] and available shelter [Iwanaga, 2011] through social media is very important during the onslaught of disasters. People can also use social media to learn about health problems brought by these disasters [Boulos, 2011], to find missing persons and disaster-struck areas [Hjorth, 2011], and to obtain the latest information about disasters [Feldman, 2016].

The timely provision of emergency supplies can effectively reduce the number of casualties and losses resulting from disasters [Giovannini, 2020]. Murakami et al. [2012] collected tweets published around the 2011 Japan earthquake and found dynamic changes in the demands of the earthquake victims. Specifically, the demands of these victims differed across each stage of the disaster response process. Therefore, apart from information dissemination, social media also plays an important role in facilitating rescue processes and identifying the demands of victims [Gao, 2011]. The American Red Cross developed an application that detects tweets for help and plans the allocation of emergency supplies. By using NLP technology on social media data, some researchers such as Deng [2016] and Wang [2018] examined the disaster crowdsourcing process, which has introduced a new hotspot in both research, practice, and has inevitably attracted the attention of many researchers.

The location information available in social media also plays a key role in disaster relief processes. By using tweets published in Christchurch, New Zealand, Gelernter et al. [2011] separately applied the named entity recognition method (developed by Stanford University) and the manual method to extract detailed location information from the collected tweets. They also highlighted the importance of the place name dictionary and the natural language processing method in automatically identify the location information of tweets. Li [2012] built a buffer zone in the disaster impact area based on a set of collected tweets that can be used in disaster simulation research. Social media relies on three types of data, namely, registration data, automatic GPS recognition data, and local text information, to extract geographic information. The geographic information carried by disaster-related tweets provide rescuers with detailed information on the situation of disaster-struck areas [Vieweg et al., 2010] that can help them assess service risks and damages [Shklovski et al., 2010].

The importance of social media at times of disaster has been highlighted in many disaster emergency management practices and has attracted the attention of governments and research institutions. The available research methods are becoming highly diverse [Runje, 2019]. The early works on the use of social media in disaster situations have mostly relied on case studies. Many researchers take the Haiti earthquake, Japan earthquake, Hurricane Katrina, Queensland flooding, and other disaster events to study the behavioural patterns and information dissemination characteristics of social media users in different disaster situations. Most of these studies have also adopted statistical analysis methods, such as descriptive statistical analysis and logistic regression. With the continuous improvement of natural language processing technologies and data mining methods as well as the flow of noisy disaster information in social media platforms, some scholars have begun to introduce improved automatic data mining technologies to analyse location information, URLs, texts, and diffusion networks from social media. The applicability and generalizability of their findings are also gradually improving.

2.2 Text Mining Methods of Social Media Text

2.2.1 Short text representation of Word2vec

Text representation is the basis for text classification and mining. One-hot Representation is a widely used word representation method. The main idea of this method is to represent each word as a Boolean vector according to the vocabulary. The length of the Boolean vector depends on the size of the vocabulary. Each word in the vocabulary corresponds to a value in the vector. Only the dimension value representing the current word in the vector is 1, and the remaining dimension values are all 0. This word representation method is very popular in natural language processing, but it has some inherent deficiencies. First, the two words in One-hot Representation are independent of each other, so this method ignores the potential semantic connection between words; second, due to the influence of the vocabulary, words expressed in this way tend to have more High dimensions. Therefore, One-hot Representation is easy to induce "dimensional disaster" in the text processing process.

Another word representation in the field of natural language processing is called Distributed Representation, which can overcome the sparse defect of One-hot Representation to a certain extent [Hinton, 1986]. The basic idea of this method is to use words as features. By training each feature (word) can be mapped to a K-dimensional vector space (in general, the dimension K is much smaller than the length of the vocabulary in the previous method). Then, all these vectors are put together to form a word vector space, and each word is a point in space. As a result, the processing of text content is simplified to vector operations in the K-dimensional vector space. Through this conversion process, the semantic similarity of the text content can be expressed as the similarity in the word vector space, so that the text data can be represented on a deeper level. Moreover, in this case, the dimension of the word is generally smaller than that of the One-hot Representation method, and the representation is not unique (that is, the dimension scale can be adjusted).

We use the words "earthquake" and "shake" to illustrate the difference between the two representation methods. Under One-hot Representation, "earthquake" is represented as [0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 ...], and "shake" is represented as [0 0 0 0

0 0 0 0 0 0 0 0 0 1 0 ...]. However, under the distributed representation, "earthquake" is expressed as [0.012 -0.040 0.015 0.044 ...], and "shaking" is expressed as [0.040 0.007 0.062 0.005 ...].

The contribution of distributed representation is that it can bring related or similar words closer. The distance between vectors can be measured by Euclidean distance or cosine distance. This method is used to express the semantic distance. Compared with the distance between the words "earthquake" and "dancing", "earthquake" and "shake" The distance between these two words will be closer. It can be seen that Distributed Representation represents the text as a dense real vector of lower dimensions, each of which represents the hidden features of the words, and this latent feature is related to the syntactic features in the text and the meaning of the words that make up the text. Features are closely related. Therefore, this method can express the syntactic and semantic features between words with the dimensional distribution.

There are many methods for training word vectors, including Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), Latent Dirichlet allocation (LDA), and Word2vec model based on deep learning ideas. The natural language processing tasks in this paper are mainly based on Word2vec. Compared with LSA, PLSA, LDA, etc., Word2vec effectively uses the context of words, and the semantic information is more abundant [Artzai et al., 2019].

Word2vec is a word vector generation tool based on deep learning released by Google in 2013. The goal is to use high-quality document data to train high-quality word vectors, and then perform many NLP-related clustering and part-of-speech tagging [Chaw, 2019]. Word2vec is essentially a neural network structure; there are two main models, namely Continuous Bag-of-Words Model (CBOW) and Skip-Gram model. Each model corresponds to two strategies. The structure of the two models is shown in Figure 1.

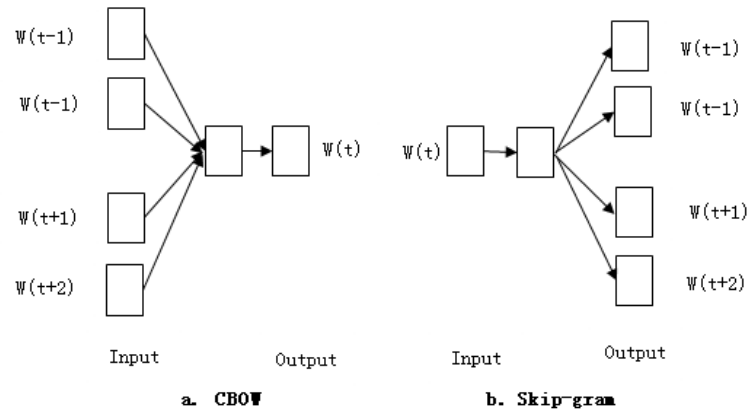


Figure 1: Word2vec model structure

The CBOW model simplifies the neural network language model, removes the hidden layer in the original structure, and saves computing time. As can be seen from Figure 1, the goal of the CBOW model is to predict the unknown word w_t with its

known context $(\cdots, w_{t+2}, w_{t+1}, w_{t-1}, w_{t-2}, \cdots)$ of w_t . Therefore, assuming that C is the input corpus, then the optimization goal of the CBOW model is the following log-likelihood function:

$$L = \sum_{w \in C} \log P(w_t / w_{t-k}, w_{t-k-1}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k}) \quad (1)$$

The main idea of the Skip-Gram model is exactly the opposite of CBOW. As can be seen from Figure 1, the goal of the Skip-Gram model is to predict the unknown context $(\cdots, w_{t+2}, w_{t+1}, w_{t-1}, w_{t-2}, \cdots)$ of with the known w_t . Then, by representing corpus C as a sequence of phrases w_1, w_2, \dots, w_T , the optimization goal of the Skip-Gram model is the following log-likelihood function:

$$L = \frac{1}{T} \sum_{t=1}^T \sum_{w \in C} \log p(w_1, w_2, \dots, w_{t-1}, w_{t+1}, \dots, w_T | w_t) \quad (2)$$

As can be seen from the above formulas (1) and (2), the conditional probability function plays a key role in the solution process. Both Hierarchical Softmax and Negative Sampling algorithms can be used to construct the CBOW model. After obtaining the specific log-likelihood function, each parameter can be solved by stochastic gradient descent, and then, we can get the word vector. Considering the training time, the first CBOW model is introduced in this paper.

The next task is how to use the word vector obtained by the above Word2vec model to reasonably represent the content features of the microblog text. A simple and popular idea is to accumulate the word vectors of all words in the text and average them as the feature vector of the microblog text [Kusner, 2015]. Assuming the text T contains k words, that is $T = \{w_1, w_2, \dots, w_k\}$. Given the word vectors set of k is $C = \{c_{w_1}, c_{w_2}, \dots, c_{w_k}\}$, then the feature vector $C(T)$ of microblogging T can be calculated by the following formula (3).

$$C(T) = \frac{\sum_{i=1}^k C(w_i)}{|k|} \quad (3)$$

2.2.2 Short text classification methods

Compared with long text such as news, Weibo text is generally shorter. Therefore, Weibo text has obvious feature sparsity. In addition, the Weibo text is very colloquial and lacks formal grammatical expression. Users often use many punctuation marks, special emoticons, and a large number of network idioms, colloquialisms, and omissions in their posts and comments. These unstructured expression features pose challenges to the mining of short text on Weibo. In the era of big data, the rapid development of machine learning technology has provided convenience for the short text mining work [Venganzones-Bodon et al., 2019]. In our research, we discussed the

following four widely used classification algorithms, which have good performance when dealing with Weibo text classification problems in previous researches.

Support Vector Machine (SVM) is derived from the theory of VC dimension in statistical learning. Based on the criterion of structural risk minimization, this algorithm constructs decision surfaces (hyperplanes) for classification based on sample information and maximum classification interval. Support vector machines were originally proposed for binary classification problems. Therefore, when solving multi-classification problems, it is generally necessary to construct multiple binary classifier combinations to complete the classification process. Because the classification process of SVM needs to be completed by solving quadratic programming, this algorithm is more suitable for a nonlinear small sample, high-dimensional data classification problems.

Random Forests (RF) is derived from the Decision Tree algorithm. It is a classification algorithm based on many decision trees. Its classification process is completed by finding the mode of many decision trees. RF has obvious advantages in the process of solving imbalanced data classification, but it is not suitable for dealing with noisy classification problems. When the sample data has obvious noise or the attribute characteristics are obviously graded, it would face overfitting problems.

The construction idea of K-Nearest Neighbour (KNN) algorithm believes that if the neighboring samples of a sample mostly belong to a certain category, this sample also belongs to a certain category. The construction idea is relatively simple and the theory is relatively mature. Similar to the RF algorithm, KNN is also a typical instance-based classification algorithm, which needs to calculate the distance between each instance. In general, the KNN algorithm has obvious advantages when processing sample data with many types of crossovers or overlaps, but due to the high computational complexity, it is not suitable for classification processes with many feature space dimensions or large data sizes.

The classification idea of Naïve Bayes (NB) algorithm is based on the design of probability knowledge in statistics. It uses Bayes theorem to calculate the probability that a sample instance belongs to a certain category, and then selects the category with the highest probability among all categories as the classification result of the sample instance. Therefore, the classification performance of this algorithm is very good and can be applied to large-scale data, but it has high requirements for the independence of the sample.

3 Methodology

3.1 Disasters Weibo Text Representation

The Word2vec text representation method has higher requirements for the training corpus. Therefore, for the task of short text natural language processing of microblogs, the text representation effect of the Word2vec model trained by the microblog corpus will be better than the training of long texts such as news and papers. In order to achieve a more reasonable representation of the disaster Weibo text, a large amount of randomly collected Weibo data is used for training Word2vec model.

Considering the requirements of the Word2vec model for the corpus and the computational complexity of the training process, based on the Weibo

platform(<https://weibo.com>), we randomly collected about 3 million Weibo messages to construct an original sample corpus(Table 1). The basic structure of this training set is shown in Table 1 below.

Weibo text has typical unstructured characteristics. And there is a lot of noisy data in the dataset. Therefore, before classifying Weibo text, we have to preprocess the text first, including word segmentation, removing URL, @, stop words, and other useless information.

In order to explore the effect of different text vector representation methods on the disaster Weibo classification process more accurately, we conducted a comparative test between Word2vec and a classic One-hot Representation model named Bag of Words.

Collection method	Collection objects	Size
Weibo API	User-generated Weibo text	3 000 000

Table 1: The original corpus structure

First, we conducted an optimal dimension test on Word2vec, and determined that the optimal dimension of the short text Word2vec based on the current corpus is 300 dimensions, so the cumulatively averaged text vector is also 300 dimensions. Then, the support vector machine is used as a classifier, and the A, F1, P, and R-value are used to evaluate the classification results under the two text representation methods. The results are shown in Table 2 below. As can be seen from Table 2, the performance of the Word2vec model is significantly better, the R-value of the Bag of Words model is basically the same as the Word2vec model, and the other three evaluation indicators are significantly lower than the Word2vec model.

According to the comparison results, the 300-dimensional Weibo short text Word2vec model is used to perform disaster Weibo short text representation work in our auto system.

Models	A	P	R	F1
Word2vec	0.903	0.889	0.887	0.874
Bag of Words	0.862	0.854	0.888	0.868

Table 2: Test results of two language classification methods

3.2 Disaster Weibo Identification

We obtained a large number of public microblogs through the "statuses / public_timeline" data interface of the Weibo API. Then, we used keyword filtering to get Weibo texts containing disaster keywords, and manually annotated these microblogs to labelling training set.

Taking the earthquake disaster as an example, from August 27, 2014 to March 31, 2015, we collected 20702 Weibo messages containing earthquake keywords. After manual screening and labelling by three experts (the known time of earthquake occurrence can provide great convenience for manual labelling), 1398 Weibo messages are marked from 20702 Weibo messages containing earthquake keywords and are related to earthquake disasters. This 1398 was taking to be a positive dataset, and given

its label as "+". Then, we randomly extracted the other 1398 irrelevant Weibo messages from the remaining data and labelled them as "-" to be a negative dataset.

Therefore, our earthquake dataset contained 2796 Weibo messages, including 1398 relevant samples and 1398 irrelevant samples. All the samples in training dataset were used for feature selection test. In order to get the best combination of features in classifiers, considering the characteristics of Weibo text, we tested eight key features, including character count, word count, user mention count, hash tag count, hyperlink count, question mark count, exclamation mark count, and unigrams. Then we used four classification algorithms to perform 10-dimensional cross-validation on all feature combinations. Each classification algorithm requires 255 experiments. It should be noted that unigrams are constructed using the Word2vec text feature vector space, which has been confirmed to be more excellent by the above experiments. The feature selection result of each classifier is shown in Table 3.

Classifiers	Best Combination of Features
SVM*	char count, link count, question mark count, exclamation mark count ,and Word2vec
KNN	exclamation mark count and Word2vec
NB	word count, question mark count, exclamation mark count ,and Word2vec
RF	all features

Table 3: The best features combinations for four classifiers

Next, the 10-dimensional cross-evaluation results of the four classifiers under the optimal feature combination are compared and analysed (as shown in Table 4). It can be seen that the four evaluation indicators of the SVM are all excellent.

Similarly, we also tested the performance of the four classifiers on the other disasters (including fires, rainstorms, floods, and typhoons) on their Weibo datasets. The experimental results showed that SVM not only performed excellent in earthquake but also superior to other classifiers in other disasters.

Based on the above experimental results, we used the 300-dimensional Word2vec model to characterize the disaster Weibo text, combined with the remaining key features selected in the feature combination traversal process to construct the feature vector space. Then, the disaster Weibo identification model is trained by SVM classifier.

Classifier	P	A	F1	R
Random Forests	0.841	0.852	0.848	0.811
KNN	0.697	0.686	0.741	0.707
SVM	0.896	0.884	0.871	0.832
Naive Bayes	0.768	0.735	0.693	0.667

Table 4: The classification results of four classifiers on training data

3.3 Disaster Events Real-time Detection

After identifying the disaster Weibo messages, in order to determine whether and where the disaster occurred for the first time, the identified disaster Weibo messages were also used for disaster event outbreak detection. For this purpose, facing the Weibo data flow, we constructed a disaster event burst detection algorithm. Its basic idea is to judge whether a disaster event has occurred according to the semantic similarity, geographic similarity, and temporal similarity of disaster relevant Weibo messages.

Cosine similarity is a classic text similarity calculation algorithm, which is widely used. Therefore, this method was also introduced in our research to calculate the similarity between short Weibo messages. According to Word2vec model, for each Weibo message, its text could be represented by a point in a d-dimensions space, $V_i = (v_1, v_2, \dots, v_d)$, d means the dimensions of Weibo text vector.

$$sim_{test}(T_i, T_j) = \frac{v_i \cdot v_j}{\|v_i\| \times \|v_j\|} \quad (4)$$

In the above formula, $v_i \cdot v_j$ means the inner product of vector v_i and vector v_j , $\|v_i\| \times \|v_j\|$ represents the length of two vectors.

For example, assuming that the vectors of Weibo X and Weibo Y are x_1, x_2, \dots, x_{100} and y_1, y_2, \dots, y_{100} , then the cosine distance between Weibo X and Weibo Y can be expressed by the cosine of the angle between them.

$$\cos \theta = \frac{x_1 y_1 + x_2 y_2 + \dots + x_{100} y_{100}}{\sqrt{x_1^2 + x_2^2 + \dots + x_{100}^2} \cdot \sqrt{y_1^2 + y_2^2 + \dots + y_{100}^2}} \quad (5)$$

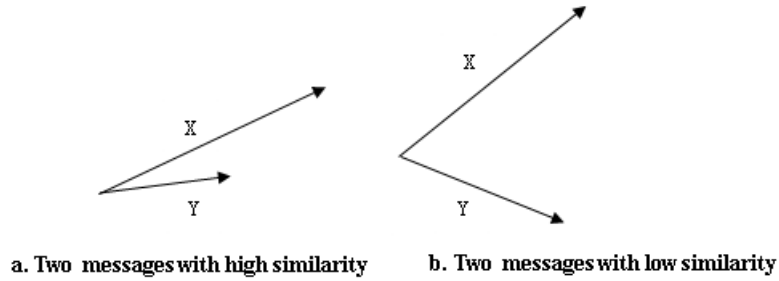


Figure 2: Cosine similarity calculation

As shown in Figure 2, when the angle cosine of two Weibo vectors (the vector X and the vector Y in the figure) is equal to 1, the two Weibo texts are repeated texts; when the cosine of the angle between the two Weibo vectors is close to 1, the two Weibo texts are similar; the smaller the cosine of the angle between the two Weibo vectors, the lower the correlation between the two Weibo messages.

After further considering the time dimension and geographic dimension, we added the time factor and geographic factor to the similarity measurement. The revised similarity calculation formula is as follows.

$$\text{sim}(T_i, T_j) = \text{sim}_{\text{test}}(T_i, T_j) * \text{sim}_{\text{time}}(T_i, T_j) * \text{sim}_{\text{geo}}(T_i, T_j) \quad (6)$$

In the above formula, $\text{sim}_{\text{test}}(T_i, T_j)$ is the semantic similarity between the Weibo texts, which is calculated using the cosine similarity method based on Word2vec.

$\text{sim}_{\text{time}}(T_i, T_j)$ is the similarity of the published time of the Weibo messages, which can describe the difference in the publication time of the messages. $\text{sim}_{\text{geo}}(T_i, T_j)$ is the geographic coordinate similarity of the Weibo messages, which describes the distance between the published geographic coordinates of the messages. Both of $\text{sim}_{\text{time}}(T_i, T_j)$ and $\text{sim}_{\text{geo}}(T_i, T_j)$ could be calculated by discrete definition. Specifically, the discrete values of the time similarity and geographical similarity of the earthquake disaster Weibo messages were set as shown in Table 5. Assuming the first disaster Weibo message (is classified to be “+”) is T^1 in a time window ω . Then, face to the next Weibo message T_i , we calculate $\text{Sim}(T^1, T_i)$. If $\text{Sim}(T^1, T_i)$ is bigger than threshold δ ; it would be put into the Disaster Warning Weibo Dataset. If the number of elements in Disaster Warning Weibo Dataset is bigger than threshold ϵ , disaster outbreak warning would be sent and the follow-up Weibo messages collection would start at the same time. Conversely, if $\text{Sim}(T^1, T_i)$ is not bigger than threshold δ , it would be considered as an independent disaster Weibo T^2 and participate in the calculation of the next disaster Weibo message. It is worth noting that time window ω , threshold δ and ϵ should be settled according to different disaster types.

Time interval	Values	Geographic interval	Values
[0—30s)	1	Same city	1
[30s—1m)	0.75	Neighboring city	0.75
[1m—5m)	0.5	Same province	0.5
[5m—30m)	0.25	Neighboring province	0.25

Table 5: The time and geographical similarity computing discrete values

3.4 System framework and information processing flow

The text stream data provided by the social media platform (Weibo) is large-scale and noisy. It is difficult and unnecessary to analyse and process the completely online platform data. In general, most studies use Keywords query method to filter sample

data from the original text stream, which is used in our research too. An interface called “statuses / public_timeline” from the Weibo API can be used for data collection.

However, although the microblogs obtained after filtering include disaster keywords, they may not be related to disaster events. Therefore, the microblogs containing the disaster keywords also need to be classified to identify the microblogs that are truly related to the target disaster event. Thus, the original Weibo text stream needs to undergo a two-category text classification process after the keywords filtering process.

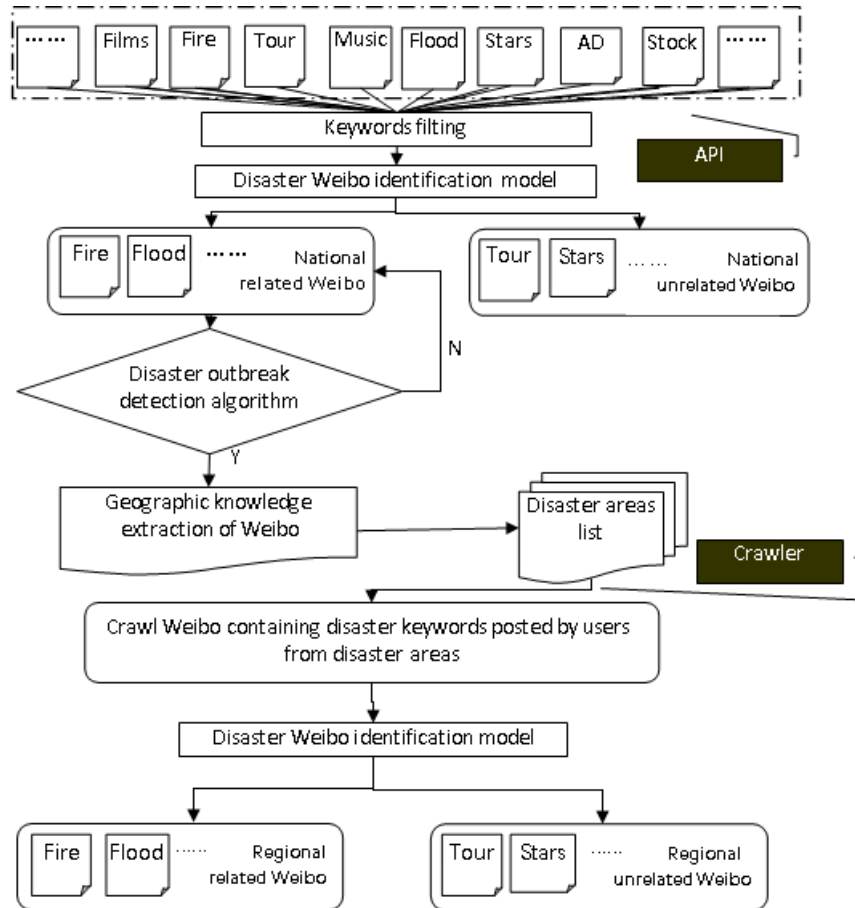


Figure 3: The framework of Disasters Auto Increment Identity System based on weibo

In addition, the emergency management process of disasters has high requirements on time. In order to effectively reduce computational redundancy, based on effective identification of disaster-related microblogs, it is also necessary to use online disaster information to realize real-time outbreak detection of disaster events. This detection is

not only helpful to start the emergency response process in time, but also helps to quickly determine the sensitive area of the disaster event during the detection process, so as to conduct targeted microblogs collection.

Specifically, according to the results of identification of public microblog information stream data captured by the Weibo API and event outbreak detection, real-time information on disaster outbreak periods and disaster areas can be obtained. Then, after real-time detection of a disaster outbreak in a certain area, we can use the "keyword search" function in the "Advanced Search" of Weibo webpage to crawl the Weibo information of the disaster areas. We could use the suitable classification model to recognize the real relevant microblogs from disaster areas that contain disaster keywords, to realize the real-time collection of disaster microblogs.

Furthermore, a multi-threaded group crawling strategy is used for network data crawling, each crawler corresponds to a disaster area, and the corresponding crawling clues are updated in real-time according to the detection results. In summary, the specific framework of Disasters Auto Increment Identity System based on Chinese Weibo text stream is designed as Figure 3.

According to the framework of Disasters Auto Increment Identity System(Figure 3), the real-time text stream data of Chinese Weibo will be processed in the proposed system following the steps shown in Figure 4. It should be pointed out that the fourth step shown in Figure 4 is not fully involved in our current work. The data cleaning part is included in the system design, and the data mining part needs to be explored for the follow-up emergency management practice requirements. Data is the basis and premise of analysis. In fact, our research achievement is that the system framework and methodology proposed above can provide sufficient, real-time data for subsequent emergency data mining.

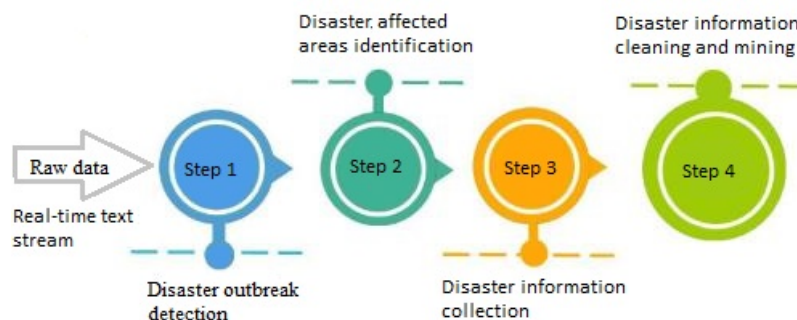


Figure 4: The flow of Chinese Weibo messages processing

4 Case Study

4.1 Leshan Earthquake

In order to confirm the effectiveness of the above methods facing to real unbalanced and noisy Weibo stream, "Leshan Earthquake" was used as a case for empirical analysis.

At 13:21 on January 14, 2015, according to the official measurement by China Seismic Network, a magnitude 5.0 earthquake occurred in Jinkouhe District, Leshan City, Sichuan Province (29.3 degrees north latitude, 103.2 degrees east longitude), with a focal depth of 14 kilometres.

4.2 Date Set

We collected 1500 Weibo messages before and after the Leshan Earthquake (from 2015-01-13 11:22:03 to 2015-01-15 15:52:44) by Weibo API “statuses/public_timeline” interface.

All the collected Weibo instances have been pre-processed with deduplication, denoising, and word segmentation. Based on the above-mentioned comparative analysis of text representation methods, the better Word2vec method is used for the microblog text representation of the Leshan earthquake data set. At last, we had a data set including 289 positive (relevant) samples and 1211 negative (irrelevant) samples for the case study.

4.3 Results

Based on the time and area information output by the detection system, more and more real-time incremental Weibo messages from disaster areas could be collected by multi-threaded crawlers and “Advanced Search” function from Weibo website. The detection time and the number of collected Weibo messages were listed in Table 7 and Figure 5.

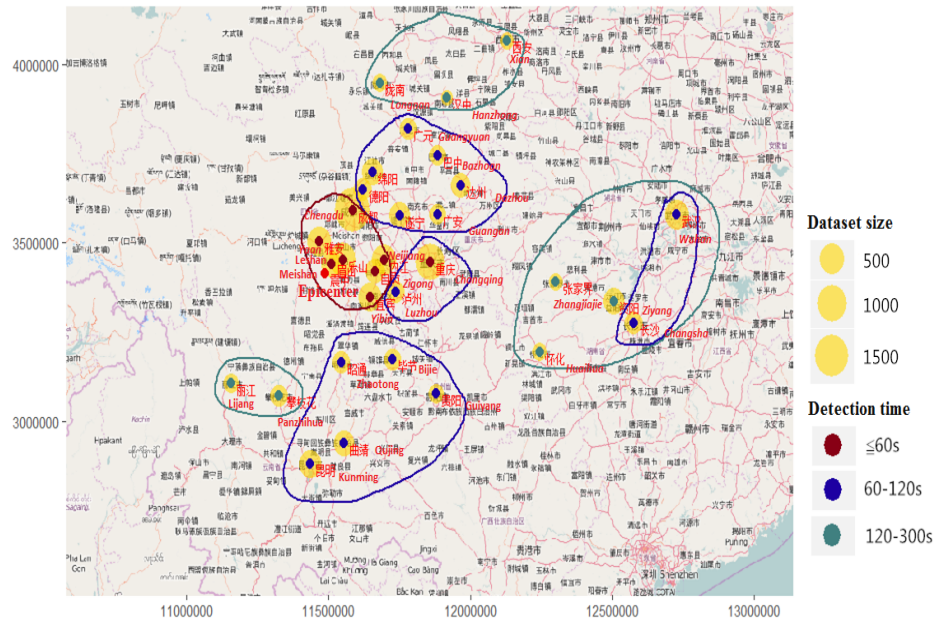


Figure 5: “Leshan earthquake” detection time and data collection size on “OpenStreetMap”

Cities	Detection time(s)	Linear distance from epicenter(km)	Dataset size
Epicenter	0-60	0	-
Chengdu	0-60	184.5	1141
Neijiang	0-60	195.8	506
Yibin	0-60	161.9	379
Leshan	0-60	74.1	764
Yaan	0-60	81.9	413
Meishan	0-60	116.1	232
Zigong	0-60	164	137
Chonging	0-60	334.3	945
Wuhan	60-120	1091.6	412
Kunming	61-120	468.6	317
Qujing	60-120	423.4	206
Luzhou	60-120	233.1	125
Dazhou	60-120	477.3	173
Changsha	60-120	927.6	116
Mianyang	60-120	295	232
Deyang	60-120	244.5	101
Guangan	60-120	368.4	71
Suining	60-120	279.1	243
Bazhong	60-120	456.7	76
Zhaotong	60-120	221.6	116
Bijie	60-120	307.6	102
Guiyang	60-120	465.9	54
Lijiang	120-300	383.7	61
Zhangjiajie	120-300	716.5	34
Huaihua	120-300	698.5	46
Guangyuan	120-300	442.3	101
Hanzhong	120-300	566.8	27
Longnan	120-300	494.8	41
Xi An	120-300	783.3	29
Guiyang	120-300	180.8	254
Panzhihua	120-300	325.5	88

Table 6: Detection time and data collection size results of “Leshan Earthquake”

It could be seen from Table 7 that the number of users has a great influence on the results of event detection and data collection. Larger cities, such as Wuhan, had a slight shock. Nevertheless, there were many users in Wuhan, so it could be detected and warned quickly. On the contrary, the cities like Zunyi and Bijie, which experienced a slight shock and few users, and then could be detected and warned slower than Wuhan.

At last, we collected 10372 Weibo messages containing disaster keywords from the above-influenced cities. After text preprocessing (such as deduplication) and screening by the disaster Weibo identification classifier, we finally obtained 7542 "Leshan earthquakes" Weibo messages, which were posted by the users experiencing "Leshan Earthquake" (see the last column of Table 6).

4.4 Evaluation and Discussion

The Accuracy (A), Recall (R), F1 value (F1), and Precision (P) are used as the performance evaluation indicators of the above four classifiers. The evaluation results of each index obtained in the Leshan Earthquake test set were shown in Table 7. By observing real data sets, we can find typical imbalances. Before the outbreak of the disaster event, there were few positive samples (basically zero). After the outbreak of the disaster event, the number of positive samples increased sharply and then showed a slow downward trend.

Classifier	A	R	P	F1	AUC
SVM	0.912	0.914	0.972	0.942	0.934
Random Forests	0.899	0.901	0.956	0.927	0.913
KNN	0.857	0.840	0.876	0.857	0.863
Naive Bayes	0.874	0.879	0.900	0.889	0.887

Table 7: The classification results of four classifiers on Leshan Earthquake dataset

Therefore, in addition to the A, R, F1, P, the ROC curve method was added to evaluate the performance of each classifier on the test set. The ROC curve is a performance evaluation indicator that is more suitable for unbalanced data sets. It is generally believed that the larger the area under the ROC curve, the better the classifier effect.

Figure 6 showed the ROC curve evaluated by four algorithms based on the Leshan Earthquake dataset, and the area under the ROC line was shown in Table 7. It can be seen from Figure 6 and Table 7 that the SVM algorithm performs best in the "Leshan Earthquake" test set, which had the largest AUC value. Moreover, SVM took much less time than RF algorithm that performs well too. Therefore, it can be proved that the classification performance of the SVM is also excellent in practice, and it is suitable for our auto incremental detection system to use this algorithm.

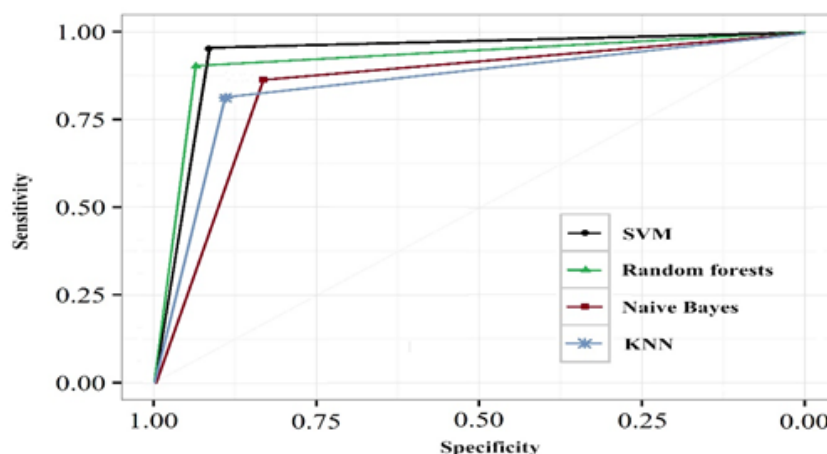


Figure 6: Roc curves

Overall, although the small "Leshan Earthquake" (magnitude 5.0) shook few cities and did not cause major casualties and property damage, the size of disaster Weibo sample dataset collected from the disaster cities is obviously bigger than the size of sample dataset initially obtained by Weibo API, which can meet the follow-up Emergency management needs better.

5 Conclusions and Future Work

A perfect disaster information dissemination tool has the characteristics of easy-to-use, mobile, reliable, and fast, and it is best to provide one-to-many communication channels and include geographic information recognition capabilities. Social media has all these advantages. For many Internet users from China, social media has become an important part of daily life. When a disaster occurs, social media becomes a key way for people to send and receive information. First-hand disaster social media information from users in disaster areas can provide real-time, large-scale information for disaster emergency response decisions. There is no doubt that these first-hand microblogs play important role in disaster emergency management. However, research on Chinese social media in disasters started relatively late. After the appearance of Weibo in China, the research on Chinese disaster Weibo information mostly revolved around the analysis of crawled historical data. In 2014, Robinson established the SWIM system. This is the first Chinese disaster information online processing system. However, due to the impact of Sina Weibo's API usage policy and data calculation scale, it is difficult to achieve disaster information crawling for the entire network data. Therefore, the related research after the establishment of SWIM system mostly uses static data. At present, there are few studies on the use of real-time and incremental disaster microblogs to serve disaster emergency management. This deficiency is closely related to the spoken and unstructured characteristics of disaster microblogs. Therefore, based on in-depth analysis of the short text features of disaster microblogs, we used a large-scale microblog corpus to train an excellent Word2vec semantic model. Furthermore, we combined Word2vec with the text features of other disaster microblogs to train the

textual representation models of the corresponding disaster microblogs for different types of disasters. Next, a short text depth representation model of microblogs and an automatic identification method of disaster information are established. A Chinese online disaster event real-time detection algorithm is proposed to realize online disaster information intelligent collection and real-time detection of disaster events. The following case study of the Leshan Earthquake confirmed the rationality and applicability of the above research model and system framework.

In addition, several useful findings are proposed by our research. First, for the short texts of disaster microblogs, the Word2vec model performs better than the bag of words model. Second, compared with the RF, KNN, and NB models, SVM can more effectively identify disaster microblogs from disaster areas from all microblogs containing disaster microblog keywords. Third, the time and geographic information carried by Weibo can be used to improve the traditional text similarity. The case study of the Leshan Earthquake confirmed that the improved similarity method can be well embedded in the online incremental disaster event detection system and perform detection work well.

However, there are some shortcomings in our research currently. Several types of disasters may be difficult to detect by using social media platforms. These disasters include those (1) occurring in less populated areas, (2) cannot be easily felt (e.g., earthquakes with magnitudes of less than three), and (3) occurring in less developed areas where most of the population do not use social media. The first two types of disasters often bring minimal damage, while the detection of the third type of disaster can be improved through the continuous development of ICT. Besides, we only explored the characteristics of Weibo text in disasters. More and more users post Vlogs (Short videos) on the Weibo platform. The Video looks more real than text and could contain more content than short text. However, in terms of technology, there are still many bottlenecks in the digital mining technology of video content. Therefore, we will also work on the fusion research of disaster video and text fact to Weibo platform in the future.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 71774041 and 71531013), Natural Science Foundation of Fujian Province of China (Grant No. 31185015), and Social Science Planning Project of Fujian Province of China (Grant No. FJ2018C008).

References

- [Abel et al., 2015] Abel, F., Hauff, C., Houben, G. J., Stronkman, R., & Tao, K. (2012, June). Semantics+filtering+search= twitcident. exploring information in social web streams. In Proceedings of the 23rd ACM conference on Hypertext and social media, pages 285-294.
- [Allen et al., 2012] Allen, R. M. (2012). Transforming earthquake detection? *Science*, 335(6066): 297-298.
- [Artzai et al., 2020] Artzai P, Aitor A. G, Unai I. et al. (2020) Why deep learning performs better than classical machine learning? *DYNA*, 95(2):119–122.

- [Atkinson et al., 2007] Atkinson, G. M., & Wald, D. J. (2007). "Did You Feel It?" intensity data: A surprisingly good measure of earthquake ground motion. *Seismological Research Letters*, 78(3), 362-368.
- [Avvenuti, et al., 2014] Avvenuti, M., Cresci, S., Marchetti, A., Meletti, C., & Tesconi, M. (2014). EARS (earthquake alert and report system): a real time decision support system for earthquake crisis management. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1749-1758.
- [Bai et al., 2020] Bai, H., Yu, H., Yu, G. & Huang, X. (2020). A novel emergency situation awareness machine learning approach to assess flood disaster risk based on Chinese Weibo. *Neural Comput & Applic* . <https://doi.org/10.1007/s00521-020-05487-1>
- [Bird et al., 2012] Bird, D., Ling, M., & Haynes, K. (2012). Flooding Facebook-the use of social media during the Queensland and Victorian floods. *Australian Journal of Emergency Management*, 27(1), 27.
- [Boulos et al., 2011] Boulos, M. N. K., Resch, B., Crowley, D. N., Breslin, J. G., Sohn, G., Burtner, R., ... & Chuang, K. Y. S. (2011). Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *International Journal of Health Geographics*, 10(1), 67.
- [Cameron et al., 2012] Cameron, M. A., Power, R., Robinson, B., & Yin, J. (2012, April). Emergency awareness from twitter for crisis management. In Proceedings of the 21st International Conference on World Wide Web, pages 695-698.
- [Chen et al., 2017] Chen, J., Chen, H., Wu, Z., Hu, D., & Pan, J. Z. (2017). Forecasting smog-related health hazard based on social media and physical sensor. *Information Systems*, 64, 281-291.
- [Chaw, 2019] Chaw, H. T, Kamolphiwong, S., & Wongsritrang, K. (2019). Sleep apnea detection using deep learning. *Tehnicki Glasnik*, 13(4), 261-266.
- [Chowdhury et al., 2013] Chowdhury, S. R., Imran, M., Asghar, M. R., Amer-Yahia, S., & Castillo, C. (2013). Tweet4act: Using incident-specific profiles for classifying crisis-related messages. In Proceedings of the 10th International ISCRAM Conference, pages 1-5.
- [Chew et al., 2010] Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS One*, 5(11), 1-10.
- [Crooks et al., 2013] Crooks, A., Croitoru, A., Stefanidis, A., & Radzikowski, J. (2013). # Earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1), 124-147.
- [De Longueville et al., 2009] De Longueville, B., Smith, R. S., & Luraschi, G. (2009). Omg, from here, I can see the flames! A use case of mining location based social networks to acquire spatio-temporal data on forest fires. In Proceedings of the 2009 International Workshop on Location Based Social Networks, pages 73-80.
- [Deng et al., 2016] Deng, Q., Liu, Y., Zhang, H., Deng, X., & Ma, Y. (2016). A new crowdsourcing model to assess disaster using microblog data in typhoon Haiyan. *Natural Hazards*, 84(2), 1241-1256.
- [Earle et al., 2010] Earle, P., Guy, M., Buckmaster, R., Ostrum, C., Horvath, S., & Vaughan, A. (2010). OMG earthquake! Can Twitter improve earthquake response?. *Seismological Research Letters*, 81(2), 246-251.

- [Feldman et al., 2016] Feldman, D., Contreras, S., Karlin, B., Basolo, V., Matthew, R., Sanders, B., ... & Serrano, K. (2016). Communicating flood risk: Looking back and forward at traditional and social media outlets. *International Journal of Disaster Risk Reduction*, 15, 43-51.
- [Gao et al., 2011] Gao, H., Barbier, G., & Goolsby, R. (2011). Harnessing the crowdsourcing power of social media for disaster relief. *IEEE Intelligent Systems*, 26(3), 10-14.
- [Gelernter et al., 2011] Gelernter, J., & Mushegian, N. (2011). Geo-parsing messages from microtext. *Transactions in GIS*, 15(6), 753-773.
- [Giovannini, 2020] Giovannini, M. (2020). Solidarity economy and political mobilisation: Insights from Barcelona. *Business Ethics: A European Review*, 29(3), 497-509.
- [Hjorth et al., 2011] Hjorth, L., & Kim, K. H. Y. (2011). The mourning after: A case study of social media in the 3.11 earthquake disaster in Japan. *Television & New Media*, 12(6), 552-559.
- [Houston et al., 2015] Houston, J. B., Hawthorne, J., Perreault, M. F., Park, E. H., Goldstein Hode, M., Halliwell, M. R., ... & Griffith, S. A. (2015). Social media and disasters: a functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1), 1-22.
- [Iwanaga et al., 2011] Iwanaga, I. S. M., Nguyen, T. M., Kawamura, T., Nakagawa, H., Tahara, Y., & Ohsuga, A. (2011, November). Building an earthquake evacuation ontology from twitter. In 2011 IEEE International Conference on Granular Computing, pages 306-311.
- [Li et al., 2012] Li, R., Lei, K. H., Khadiwala, R., & Chang, K. C. C. (2012, April). Tedas: A twitter-based event detection and analysis system. In 2012 IEEE 28th International Conference on Data Engineering, pages 1273-1276.
- [Musaev et al., 2014] Musaev, A., Wang, D., & Pu, C. (2014). LITMUS: a multi-service composition system for landslide detection. *IEEE Transactions on Services Computing*, 8(5), 715-726.
- [Murphy et al., 2016] Murphy, R. R. (2016). Emergency informatics: using computing to improve disaster management. *Computer*, (5), 19-27.
- [Murakami et al., 2012] Murakami, A., & Nasukawa, T. (2012, April). Tweeting about the tsunami?: Mining twitter for information on the Tohoku earthquake and tsunami. In Proceedings of the 21st International Conference on World Wide Web, pages 709-710.
- [Neubauer et al., 2015] Neubauer, G., Huber, H., Vogl, A., Jager, B., Preinerstorfer, A., Schirnhofer, S., Schimak, G. & Havlik, D. (2015, March). On the volume of geo-referenced tweets and their relationship to events relevant for migration tracking. In International Symposium on Environmental Software Systems , pages 520-530.
- [Power et al., 2014] Power, R., Robinson, B., Colton, J., & Cameron, M. (2014, October). Emergency situation awareness: Twitter case studies. In International Conference on Information Systems for Crisis Response and Management in Mediterranean Countries, pages 218-231.
- [Rogstadius et al., 2013] Rogstadius, J., Vukovic, M., Teixeira, C. A., Kostakos, V., Karapanos, E., & Laredo, J. A. (2013). CrisisTracker: Crowdsourced social media curation for disaster awareness. *IBM Journal of Research and Development*, 57(5), 4-1.
- [Runje, 2019] Runje, B., Kondic, Z., Horvatic Novak, A. & Keran, Z. (2019). Estimation of process capability based on continuous and attribute data. *Tehnicki Glasnik*, 13 (2), 162-164. <https://doi.org/10.31803/tg-20190514132701>

- [Vieweg et al., 2010] Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010, April). Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1079-1088.
- [Sakaki, et al., 2010] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web (pp. 851-860).
- [Sakaki, et al., 2012] Sakaki, T., Okazaki, M., & Matsuo, Y. (2012). Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4), 919-931.
- [Shklovski et al., 2010] Shklovski, I., Burke, M., Kiesler, S., & Kraut, R. (2010). Technology adoption and use in the aftermath of Hurricane Katrina in New Orleans. *American Behavioral Scientist*, 53(8), 1228-1246.
- [Stieglitz et al., 2018] Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics—challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39, 156-168.
- [Thelwall et al., 2007] Thelwall, M., & Stuart, D. (2007). RUOK? Blogging communication technologies during crises. *Journal of Computer-Mediated Communication*, 12(2), 523-548.
- [Venganzones-Bodon, 2019] Venganzones-Bodon M. (2019) Machine learning challenges in big data era. *DYNA*, 94(5):478–479.
- [Wang et al., 2018] Wang, R. Q., Mao, H., Wang, Y., Rae, C., & Shaw, W. (2018). Hyper-resolution monitoring of urban flooding with social media and crowdsourcing data. *Computers & Geosciences*, 111, 139-147.
- [Yang, A., 2020] Yang, A., Liu, W, & Wang, R. (2020). Cross-sector alliances in the global refugee crisis: An institutional theory approach. *Business Ethics: A European Review*, 29(3), 646-660.
- [Yates et al., 2011] Yates, D., & Paquette, S. (2011). Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake. *International Journal of Information Management*, 31(1), 6-13.