https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022

Natural Hazards and Earth System Sciences Discussions

Real-time urban rainstorm and waterlogging disasters detection by Weibo users

Haoran Zhu[1], Priscilla Obeng Oforiwaa[1], Guofeng Su[1]

[1]Department of Engineering Physics, Tsinghua University, Beijing, 100086, China

Correspondence to: Guofeng Su (sugf@mail.tsinghua.edu.cn)

Abstract. With the process of urbanization in China, the urban waterlogging caused by rainstorm occurs frequently and often leads to considerable damage on natural environment, human life, and the city economy. Rapid detection of rainstorm and urban waterlogging is an essential step to minimize related losses. Weibo, a popular microblogs servicer in China, can provide many real-time Weibo posts for rapid detection. In this paper, we propose a method to identify microblogs with rainstorm and waterlogging information and apply them to waterlogging risk assessment. After pre-processing the microblog texts, we evaluate the performance of clustering (K-means) and classification (support vector machine, SVM) algorithms in the classification task. Apart from the word vector features, we also introduce the sentiment and publisher features for a more real-time and accurate results. Furthermore, we build a waterlogging dictionary to assess the waterlogging risk from the Weibo texts, and get a risk map with ArcGIS. To examine the efficacy, we collect Weibo data from two rainstorm and waterlogging disasters in Beijing city as examples. The results indicate that the SVM algorithm can be applied for real-time rainstorm and waterlogging information detection. Compared to the official authentication and personal certification users, the microblogs posted by general can better show the intensity and timing of rainstorm. The location of waterlogging points is consistent with the risk assessment results, which can be used as a reference for timely emergency response.

Key words: Urban rainstorm and waterlogging; Text classification; Risk assessment; Real-time detection

1 Introduction

In our modern society, urban rainstorm waterlogging is a frequent occurrence all over China, resulting in serious damage on city safety and resident's daily life (Yin et al. (2015)). On July 21, 2012, a heavy rainstorm swept Beijing. The rainstorm and waterlogging destroyed 10,660 houses, caused economic losses up to RMB 11.64 billion and the deaths up to 79 people. Such serious waterlogging disasters also occurred in Shanghai on September 13, 2013, in Guangzhou on May 7, 2017, in Zhengzhou on July 20, 2021. Therefore, many studies have been carried on urban rainstorm and waterlogging's risk assessment, preventing, forecasting and early warning.

Many factors work together in the formation of waterlogging, including rainstorm, terrain, drainage system, vegetation coverage rate and so on. Considering these features, many different models were constructed to simulate the submerged area and depth of different rainstorm return period. A classic type is the hydrologic model, represented by the Storm Water

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

Management Model (SWMM) ((Bisht et al. (2016); Jiang et al. (2015)). The hydrologic model depends on the pipeline data and needs huge modelling and calculating resources, which limits its application. To make up for this setback, many simplified models are developed (Quan (2014); Tao et al. (2018); Tang et al. (2018)). These models study the macroscopic relationship between factors and the waterlogging disasters instead of the detailed dynamic calculation, and obtain results of good conclusions compared with the hydrologic models. Based on the risk assessment and forecasting models, improved suggestions on waterlogging prevention was also promoted to the government. In the early warning step, researchers focus more on the improvement of government's management system and the building of physical sensor networks (Perera et al. (2020); Liu et al. (2014)). However, these early warning systems require significant time and material costs. We need a faster and more convenient early warning method.

Compared to the physical sensors, Sakaki et al. promoted a concept of social sensors for earthquake detection in 2010 (Sakaki et al. (2010)). Social networks such as Twitter and Weibo are popular all over the world. As of March 2021, Weibo's monthly active users totalled 530 million, of which 94% were mobile users. They are all real time social sensors for sharing their experience on Weibo site. Many disasters' early warning researches were carried on the analysis of Weibo's content (Choi et al. (2015); Avvenuti et al. (2016); Cao et al. (2017)). By screening the keywords with supervised or unsupervised algorithms, we can quickly select out and analyse the relevant microblog texts related to specific disasters. Most of the researchers focus on the earthquake disaster, as there is usually only one strong earthquake at a time. The detection of urban rainstorm and waterlogging disasters is more complex. Nair et al. (2017) used naïve Bayes (NB), random forest (RF) to detect the 2015 Chennai flood's effect on Twitter users. Na Xiao et al. (2018) promoted an identification method of urban rainstorm waterlogging microblogs based on three supervise algorithms in 2018, of which the classification accuracy rate came up to 84%. Unsupervised algorithms are also widely used in the Weibo content processing ((Gao et al. (2014); Wang et al. (2013)), but there are few researches discussing their applications in urban waterlogging disasters. The analysis of Weibo data also limits in the word vector in most cases, without taking the features of emotion, publisher or interaction (like, comment and transform) into account. Most of the previous studies separated the storm-related microblogs from general microblogs, while the emergency response needs to further select out the microblogs containing timely rainstorm information from the storm-related microblogs.

Based on the above discussion, this study takes two urban rainstorm and waterlogging disasters in Beijing as examples, promoting an integrated data-mining method on Weibo users. The primary objectives of our study are as follows: 1) to evaluate the role of publisher and sentiment feature in real-time rainstorm and waterlogging microblogs information extraction; 2) to select out the microblogs with real-time information; 3) to get a waterlogging risk assessment map of the microblogs with location information.

The remainder of this paper is organized as follows. Section 2 briefly introduces our study area and the data collection process. Section 3 covers all data processing steps, including data pre-processing (section 3.1), data classifying and effect evaluation (section 3.2) and risk assessment (section 3.3). Section 4 compares and analyses the results; and the last section presents the conclusion and recommendations

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

2 Study area and data collection

2.1 Study area

Beijing city, the capital of China, was selected as the study area. Beijing locates in the monsoons climate zone, and the sudden rainstorms often result in serious waterlogging disasters. The study area covers 16,410 km^2 with 16 districts, and is surrounded by mountains in the west, north and northeast. Precipitation centres are distributed along the windward slopes of these mountains. The foothills of northern of Beijing, eastern slope and piedmont area of the Taihang Mountains, and the land–sea interface of Bohai Bay contribute to the formation of short-duration heavy rainfall (SDHR) apart from the climate factors (Cheng et al. (2021), which makes the urban rainstorm and waterlogging early warning more difficult. The monsoon season in Beijing spans from June through September. After the serious rainstorm and waterlogging disaster on July 21, 2012, the government identified 64 waterlogging vulnerable points. However, with the development of urbanization, every time there is a rainstorm, there are new spots of waterlogging. Compared to relying on the experience, rapid detection of waterlogging points in a rainstorm is a better solution.

2.2 Data collection

We chose two heavy rainstorms in Beijing on August 12, 2020 (case A) and July 18, 2021 (case B) as the cases, for which the Beijing Emergency Management Bureau issued an orange alert. We extracted Weibo text about disaster situation based on keywords ‘Beijing’ and ‘rainstorm’ as a preliminary selection. In case A, meteorological information showed it would be the heaviest precipitation in Beijing in recent years, so the warning was issued long time before the precipitation. We extracted 19,791 Weibo data in the 48-hour period from 0:00 on August 12,2020 to 23:59 on August 13, 2020. In case B, we extracted 2,840 Weibo data in the 36-hour period from 12:00 on July 17, 2021 to 23:59 on July 18, 2021. The data included word text, user information, release time, and release type (post or repost).

3 Methodology

Text is a typical unstructured data with complex features. After pre-processing, we can get the basic unit of text, the words. Compared with other features, words are more ambiguous and changeable with the context. What’s more, the microblogs always have a character limitation of 140 characters. To better understand the information in microblogs in order to classify them, we introduced more features apart from the words, e.g. the publisher and the sentiment features. Some microblogs contain location information by mentioning area names in the text or choosing to show their current location when posting. These microblogs are more accurate sensors, which aids in the accurate estimation of spatial distribution of precipitation and waterlogging instead of only time distribution. The overall processing framework is shown is Figure 1.

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022

Natural Hazards and Earth System Sciences
Discussions

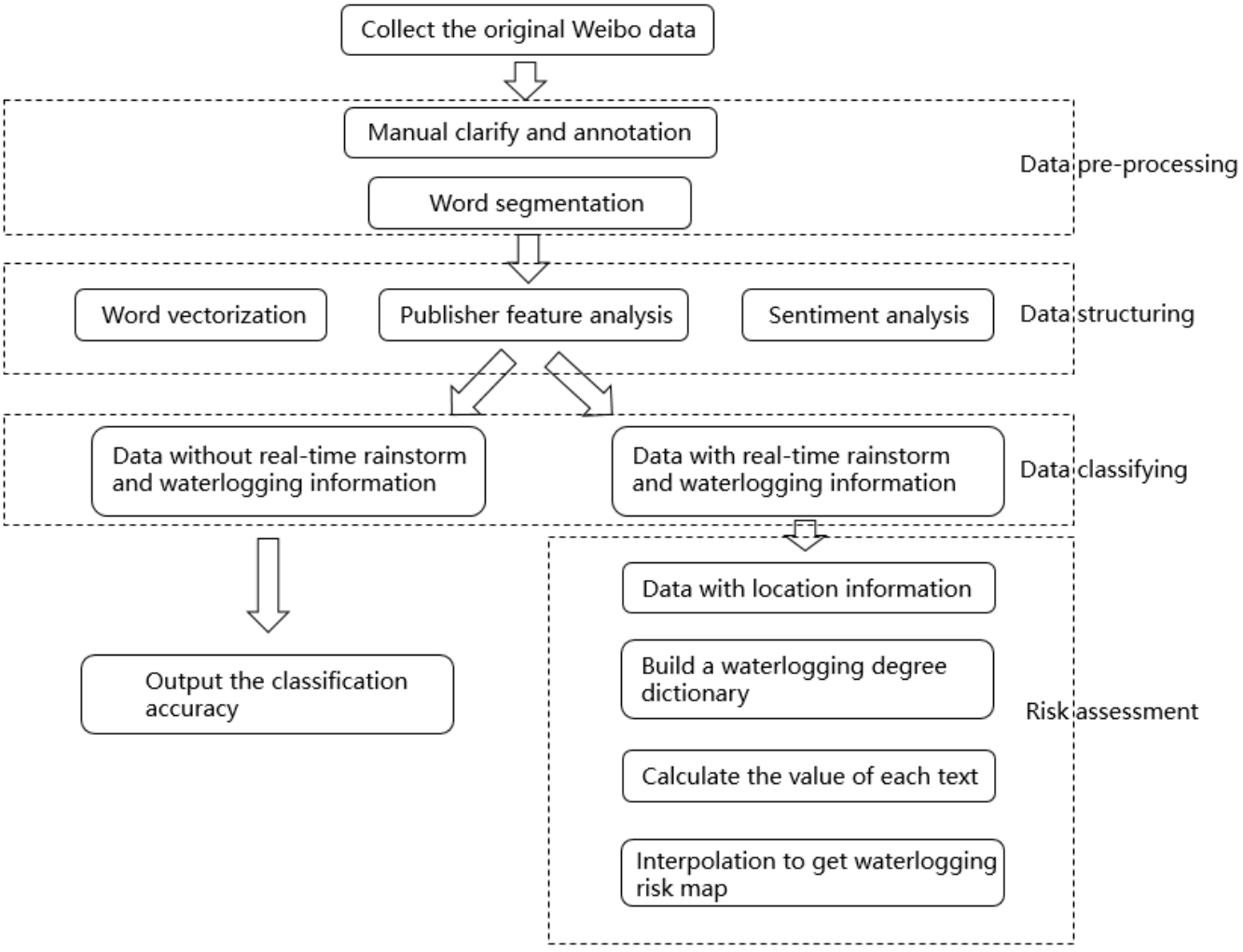


Figure 1. The overall processing framework

3.1 Data pre-processing and structuring

3.1.1 Manual clarity and annotation

As social sensors, there must be some errors and distortions in Weibo users' information. There is an assumption that if a microblog is reposted, it is more likely to carry outdated information, which cannot be used for real-time disaster assessment, so we eliminate all the repost text. In supervised algorithms, the label of each text must be given for model training. A text carrying information related to rainstorm is a broad concept, such as the forecasting and warning messages from the meteorological department before the rainstorm, the share of a particular view from the general users after the rainstorm. Microblogs with such information tend to be classified into positive category in many researches. However, these microblogs contain advanced or lagging information, which contributes little to current risk assessment and early warning. In our

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

experiments, timeliness is as important as disaster description. The microblogs with description of real-time urban rainstorm and waterlogging disaster are marked as positive categories with the label 1. Others are negative with label 0. Now we transform the text classification problem into a binary classification one.

Table 1. Examples of positive and negative Weibo texts

Category	Weibo texts
Positive	The rain rushes underground outside, and I am falling asleep listening to the rain.
Negative	Latest forecast! Today Beijing will welcome the heaviest rain in the flood season, and the main rainfall period is from noon to night
Negative	Beijing after the rainstorm. I love autumn in Beijing the most, the sky is high and the clouds are pale.

3.1.2 Word segmentation

Compared to English, words segmentation is more complex in Chinese as there is no separation between characters. The segmentation steps are as follows.

(1) Filtering the Chinese characters. Most of the segmentation modules support only one language, while there are many English or special symbols in the text. In this paper, we used a regular expression to filter the text first.

(2) Removing the Weibo topic tags. To make it more convenient for the public to participate in the discussion, Weibo gives every topic a tag with the format of '##', for example '#Beijing rainstorm#'. Many microblogs have three or four tags, and we should remove them for less distortion. Stop words are functional words that have no concrete meaning, e.g. conjunction, preposition and interjection. In many researches, stop vocabularies were created to remove these words. However, we did not eliminate the stop-words in this paper, as many stop-words could help us understand the time information in Weibo texts.

(3) Word segmentation. In this paper, we introduced Jieba segmentation module for word segmentation, which is based on word frequency statistics to segment Chinese words.

(4) Weight calculation. We got the word vector representation of every text after word segmentation, but each word had different weight. Term frequency–inverse document frequency (TF-IDF) is a popular method in determining weights. The basic idea is the fewer the words appearance, the more important it is in this document. The calculation formula is:

$$w_{ij} = tf_{ij} \times idf_{ij} = \frac{n_{i,j}}{\sum_k n_{kj}} \times lg \frac{|D|}{|\{j: t_i \in d_j\}|} \quad (1)$$

In which the $n_{i,j}$ represents the number of occurrences of word i in document j, $\sum_k n_{kj}$ represents the total number of words in document j, D represents the total number of documents, and $\{j: t_i \in d_j\}$ represents the number of documents with the word i.

(5) Sentiment analysis. Traditional sentiment analysis only focuses on sentimental words, degree adverbs and negative words, while another thinks that every word contributes to the sentiment value. We prefer the second method. In this paper, we

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

use Boson dictionary for sentiment analysis by Boson NLP Company, which gives the sentiment value of each word.

3.2 Data classifying and effect evaluation

There are two commonly used classification models: clustering methods and supervised methods. We introduce k-means algorithm representing clustering method and SVM algorithm representing supervised methods, then apply them into the classification task and evaluate the accuracy. A brief introduction of the two methods are given below.

The k-means clustering analysis is a simple and commonly used clustering algorithm based on the squared error criterion which is over 50 years old (MacQueen (1967)). This algorithm divides n samples into k categories as the squares within the group are less than the squares between groups. In the k-means method, the position of clustering centres and the number of categories determine the clustering accuracy. For the first problem, we randomly initialize the initial clustering centres multiple times and took the mean to avoid the results falling into local extremums. For the latter problem, the method of sum of the squared errors (SSE) and silhouette coefficient are popular for this task.

SVM is a linear classifier defined in feature space with the largest interval, which distinguishes it from perceptron. With the help of different kernel function, it can also handle nonlinear problem. As a supervised method, the model studied from the training set is verified with the test set to illustrate the overfitting and under fitting problems in the process. We then choose a better combination of the penalty(trial) parameter C and the type of kernel for better results.

There are four possible classification situations: predict a positive sample as positive (TP), predict a positive sample as negative (FN), predict a negative sample as positive (FP), and predict a negative sample as negative (TN). Based on these four indicators, we further give three global indicators.

Precision represents the proportion of original positive samples in the samples judged positive by the classifier, which can be calculated by:

$$\mathrm{P} = \frac{TP}{TP + FP} \tag{2}$$

Recall ratio represents the proportion of the positive samples judged correctly by the classifier in the positive samples, which can be calculated by:

$$\mathrm{R} = \frac{TP}{TP + FN} \tag{3}$$

F-score balances the precision and recall ratio by parameter β, which can be calculated by:

$$\mathrm{F} = (1 + \beta^2)\frac{P \times R}{\beta^2 \times P + R} \tag{4}$$

The value of β represents the relative importance between precision and recall ratio. If $\beta > 1$, recall ration is more important; if $\beta < 1$, we focus more on the precision. In this paper, we choose β's value as 1.

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

3.3 Risk assessment

Some microblogs contain location information by mentioning place names in the text or choosing to show their current location when posting, which makes it possible to assess the spatial distribution of waterlogging risk. The risk assessment steps are as follows.

(1) Extracting location information. As mentioned above, there are two types of location information. If the users choose to show their current location when posting microblogs, the location information can be extracted while extracting the microblog texts. In a situation where the location is mentioned the text, we build a dictionary of location names in Beijing to get the location information. However, the dictionary can only be accurate up to the street name rather than the buildings. Therefore, we mainly depend on the current location shown by the users.

(2) Assessing the waterlogging risk in a single microblog. In this paper, we proposed that the waterlogging risk consists of three types of information: the size of rain, the duration and the description of current accumulation of water. If the rain is heavier and the duration is longer then, current accumulation of water is more serious. we consider the waterlogging risk in this region as higher. We build a dictionary for extracting the three types of information out of the text and the risk assessment.

(3) Assessing the spatial distribution of waterlogging risk. Combining the risk value and location information, we get the isolated risk point on the map. With the use of interpolation, we get the spatial distribution of waterlogging risk and on the map.

4 Results and discussion

The experiment used the Windows 10 operating system and run programs in the Python 3.9 environment. We mainly applied the Jieba and sklearn package in data processing. The spatial distribution of waterlogging risk was calculated and drawn by the ArcGIS.

In the pre-processing step, we first iterated over all the microblogs text and built a bag of words with 22,692 words. Then we performed word segmentation on each microblog, and recorded the number of occurrences of each word and the location of the word. We introduced the sparse matrix in this process for a rapid calculation. Which enabled the conversion of microblog into a vector with 22,692 dimensions. Then we calculated the sentiment value of each microblog by the Boson dictionary. If the value is greater than zero, it is recorded as a positive sample with a label 1. Others recorded as negative samples with a label 0. We combined them together to get a vector with 22,693 dimensions for further classification.

4.1K-means clustering results

In case A, we extracted 19,791 Weibo data in the 48-hour period from 0:00 on August 12 to 23:59 on August 13. After filtering out the repost and repeat microblogs, 9,951 Weibo data remained. We considered the category of Weibo with timely rainstorm information to be positive, and the others to be negative. By manual annotation, there are and 7,410 negative and 2,541 positive

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022

Natural Hazards and Earth System Sciences Discussions

categories in case A. Firstly, we apply the word segmentation process in Weibo data and convert the text into word-vector representation. Then we introduced the SSE and the silhouette coefficient method for determining the optical number of categories. In the SSE method, the point with the highest decreasing rate corresponds to the optimal number of categories, so the SSE method is also called the elbow law. In the silhouette coefficient method, the point with highest value corresponds to the optimal number of categories. To avoid falling into local extremes, we randomized the initial centre points 20 times and take the mean value in both methods.

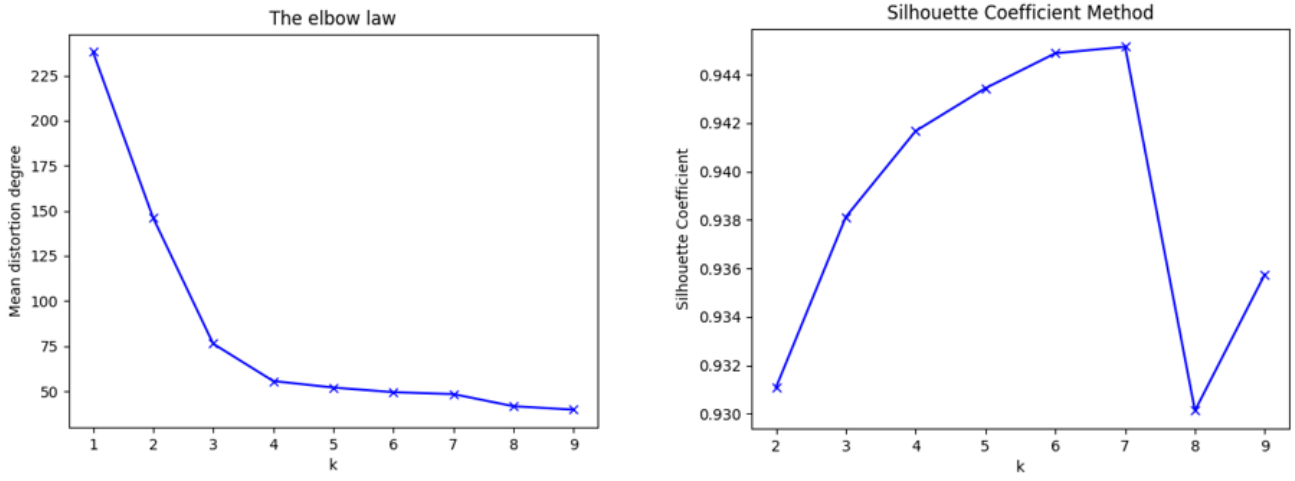


Figure 2. The evaluatuon for different number of categories in the elbow and silhouette coefficient method

As shown in the figure 2, there were some puzzles between the two methods. In the silhouette coefficient, 7 seemed to be the most appropriate number. However, the elbow law illustrated that 2 is the best, followed by 3. This situation illustrated the uncertainty of k-means clustering. Bo (2018) once applied the k-means method for the classification of earthquake microblogs, and found that the subjects of social media information were always mixed. Considering both methods, we chose the number of categories as 2.

The comparison of positive proportion between three groups are shown in Table 2.

Table 2. Comparison of positive and negative proportion between three groups (unit: %)

	Overall	Class 1	Class 2
Positive proportion	25.20	24.47	33.80

It could be seen from the comparison that the k-means method did not distinguish the text with or without the disaster information. Actually, in the classification task with multidimensional features, the k-means method usually returned as a poor performance. It is more suitable for fixed and fewer features, for example, the clustering of geographical locations of disaster information (Lu et al. (2016)).

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

4.2 Sentiment features, publisher features and the SVM classification results

There are also many types of information from rainstorm, for example, the warning news before the storm, the rescue reports after the storm. Such information could hardly be separated from the real-time rainstorm information, while they contribute little to the real-time disaster assessment and early-warning. In this paper, we improved the extraction results of rainstorm and waterlogging information in two aspects.

(1) Extract more real-time information. The real time has two concepts: extraction of the information more quickly and the information description of the situation short before its occurrence

(2) Migratory validation. The training set and test set do not come from the same disaster as the high-frequency words vary from case to case.

We introduced the sentiment analysis into the classification process. By manual annotation, we first divided the data into two categories according to whether it carried real-time rainstorm information or not. In the category with useful information, the proportion of negative microblogs was 55.42% while in the other category, the number was 44.85%. We further separated the waterlogging microblogs from the rainstorm ones. In the waterlogging category, the proportion of negative microblogs was 70.12%. This difference in sentiment values could help us better classify rainstorm and waterlogging information. In sentiment analysis, if the sentiment value was greater than 0, the text would be marked as 1. The others were marked as 0.

We imported data from case A as training set and data from case B as test set. As comparison, we also set an experiment with 80% of the samples in case A as training set, 20% of the sample in case A as test set. By manual annotation, there were 2,541 positive and 7,410 negative categories in case A, and 710 positive and 971 negative categories in case B. The experimental results are shown in table 3.

Table 3. The classification accuracy (unit: %)

	Precision	Recall rate	F value
Within case A	75.59	68.25	71.73
Migratory validation	62.06	71.90	66.62

From the table above, we could get a higher accuracy within one disaster data group. In migratory validation, the accuracy would be slightly worse, but it could be further improved with a larger training set. In the experiments, it was especially difficult to classify short texts with less than 10 words. People usually used some objects to represent rainstorm and waterlogging, such as shoes or windows, which rarely appeared in our training set. This situation reduced the accuracy of the classification. Overall, it showed the feasibility of being applied in practice.

There are two most common classification kernel in SVM. The linear kernel got a higher score in precision and selected out more correct microblogs, while the gaussian kernel performed better in recall rate index and made fewer mistakes. In our experiments, we chose the linear kernel for a better performance in precision for the larger number in correct microblogs. On the selection of the word vector weight, the term frequency (TF) performed better than the TF-IDF. There were many kinds of

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022

Natural Hazards and Earth System Sciences Discussions

expressions when talking about rainstorm information, and TF-IDF might exaggerate the weight of specific expressions and make it difficult to learn.

There are three types of publishers: official authentication, personal certification and general user. We counted the number of microblogs on rainstorm and waterlogging posted by three types of user in each hour from 11 am on August 12 to 11 am August 13. As the number of each publisher type was different, we calculated the percentage of posts in each hour for different types. The results are shown in figure 3.

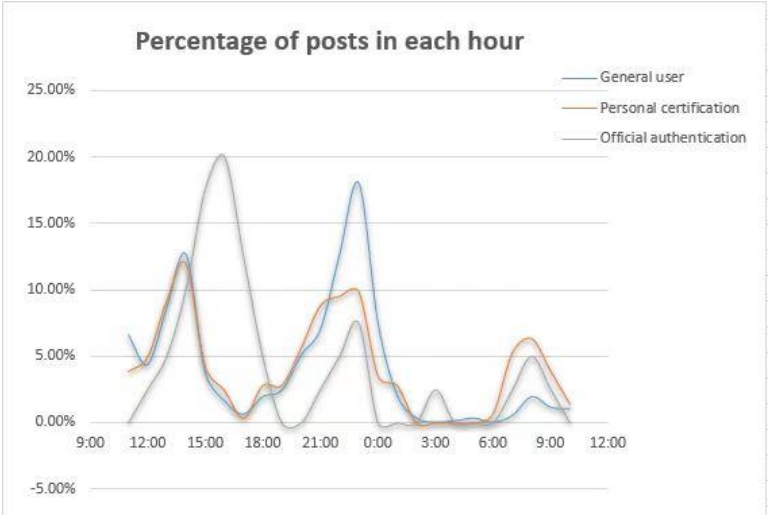


Figure 3. The percentage of posts for different types in each hour.

There were two large-scale precipitations in this period of time. One with less rainfall lasted from 11 am to 2 pm. The other heavy precipitation lasted from 9 pm to early morning the next day. For the peak position, there was obviously a lag about 1 hour for official authentication users. For the peak height, the number of microblogs posted by the general users was positively correlated with precipitation. Therefore, when we used social network data to evaluate the extent of urban rainstorms and waterlogging disasters, much attention must be paid to the microblogs of general users.

4.3 Waterlogging risk assessment and map-labelling

We were concerned about the location of waterlogging points in urban rainstorm and waterlogging disaster, which was hard to obtain. We often used precipitation to estimate the depth of water in different blocks. By analysing the word frequency of the microblogs with description of rainstorm and waterlogging disasters, we built a dictionary for waterlogging points detection. The dictionary contained three types of words shown in Table 4.

Table 4. The dictionary of waterlogging degree

Degree	Precipitation information	Time information	Waterlogging information
0	Rain, rainy, light rain…	Recent, now, minutes…	

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022

Natural Hazards and Earth System Sciences Discussions

1	Storm, stormy, heavy rain…	Hours, incessant, all night…	Stagnant water, ankle, submerged…
2	Downpour, scary, broken sky…	All the day, the last few days…	River, ocean, boat, knee, tire…

We divided the texts with precipitation and time information into three categories. From value 0 to 2, the risk of waterlogging was low, medium and high. When it came to the waterlogging information, the precipitation had exceeded the drainage capacity, so the degree started from value 1. Adding the three degrees together, we gave the waterlogging risk value in different blocks. We assumed that areas with a value greater than 3 had a high risk of waterlogging.

In case A, we selected out 269 microblogs with location information. Through the open platform of Baidu Map, we got the latitude and longitude of each point. We marked them by ArcGIS in figure 4.

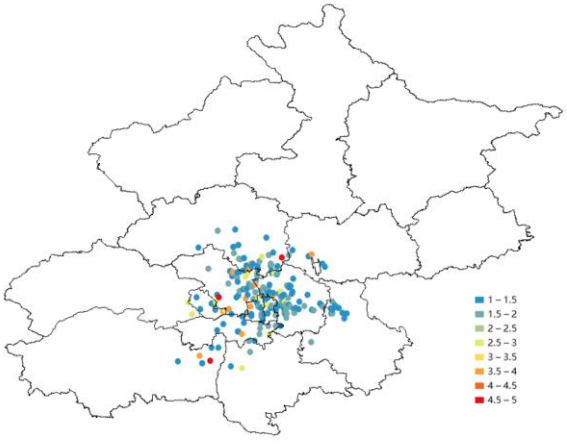
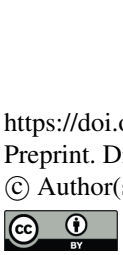


Figure 4. Annotation of the waterlogging microblogs with location information and risk value on the original map

It could be seen from the figure 4 that most microblog' locations were concentrated in the core urban area of Beijing. We chose Haidian District, Chaoyang District, Dongcheng District, Xicheng District, Shijingshan District and Fengtai District as examples where the microblogs with location information were concentrated. From the scattered points, we wanted to assess the waterlogging risk over the study area. Interpolation could contribute to this task. In this paper, we applied the inverse distance weighted method and the local polynomial method (Caruso et al. (1998)) for comparison. And we showed the interpolation results in the form of a map, the risk assessment map.

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022



Natural Hazards and Earth System Sciences
Discussions

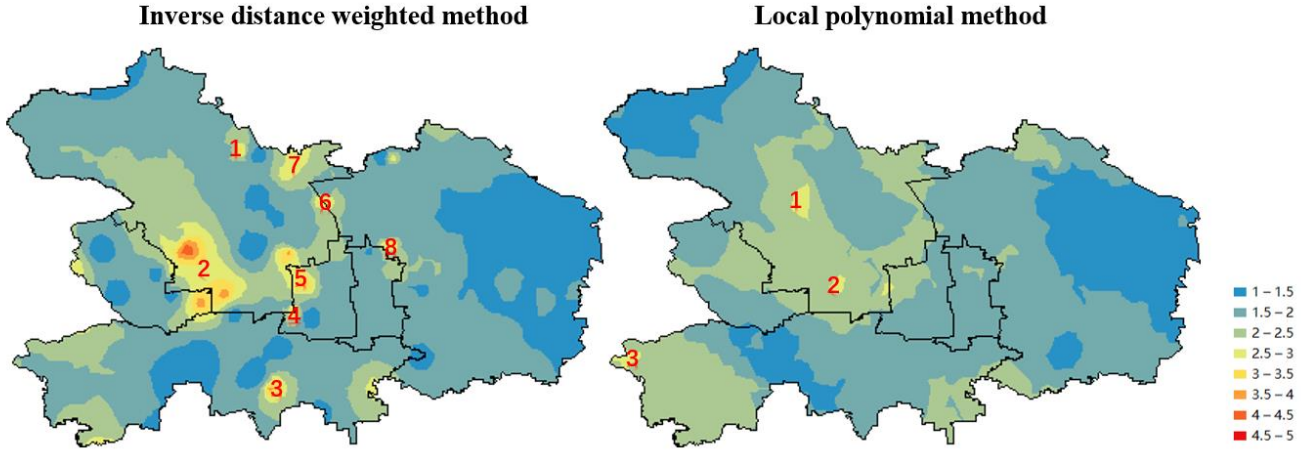


Figure 5. Waterlogging risk assessment map by inverse distance weighted method and local polynomial method

Figure 5 showed that inverse distance weighted method could get more discriminative results than the local polynomial method, which made it easier to identify the key areas for emergency response. In the map of inverse distance weighted method, there were mainly eight areas with a risk value higher than 3 and we marked them on the map from number one to eight. Compared with the local polynomial method, the high-risk areas from the inverse distance weighted method were larger and more dispersed. In order to compare the accuracy of the two methods, we found the waterlogging point information in the six districts from the microblog of the Beijing Daily (Copernicus Publications). The annotation results were shown in figure 6.

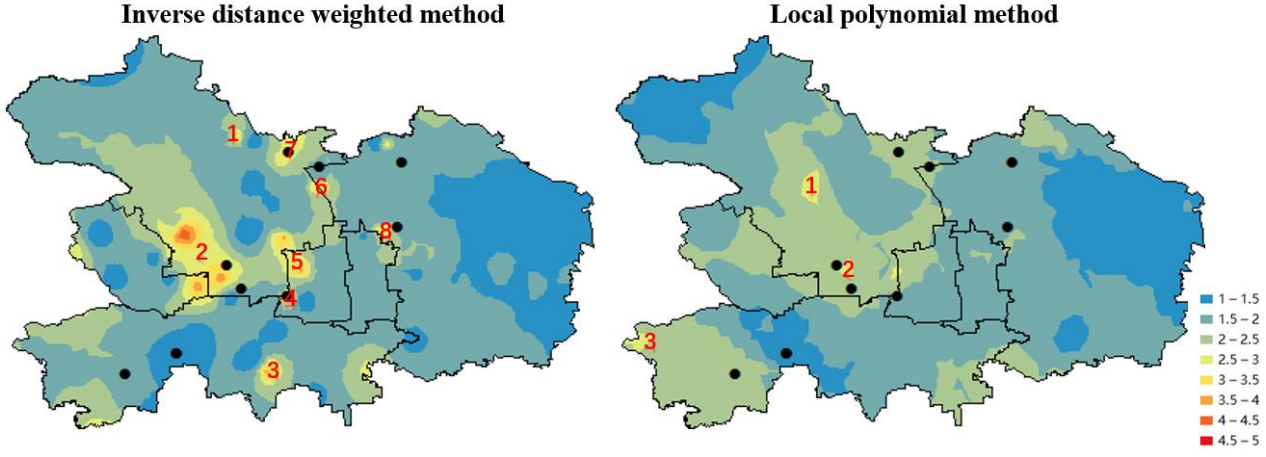


Figure 6. Annotation of the waterlogging points on the risk assessment map

We could conclude from the figure 6 that inverse distance function was better. In the total nine waterlogging points, four were in the area with a value over 3, located in Zone 2, Zone 4, Zone 6 and Zone 7. Three points were in the local extreme area, located in Zone4, Zone 7 and Zone 8. What's more, in eight areas with greater risk of waterlogging, four had waterlogging points, including Zone2, Zone 4, Zone 6 and Zone 7. This showed that the high-risk areas of waterlogging obtained from weibo texts with location information were certainly accurate. We could also notice that some waterlogging points located in the low-

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

risk area. The first reason was the conditions for the formation of waterlogging were complex, in addition to precipitation, including topography, drainage systems and other reasons as well. Another reason was the limit on the number of microblogs, which led to inaccurate assessments in marginal areas. Through annotation and different of Weibo texts, we could initially screen out high-risk areas and make references for timely early warning and emergency response.

5 Conclusion

This paper proposed a social network data processing method for urban rainstorm and waterlogging disasters' risk assessment and real-time detection. Based on the word vector, we could separate out the microblogs with timely disaster information. Combining the classification results with the publisher feature and sentiment analysis, we could better understand the time and severity of the rainstorm and waterlogging disasters. Microblogs posted by general users can better represent the intensity and timing characteristics of precipitation and microblogs posted by personal certification users are also timely. Furthermore, we built an urban rainstorm and waterlogging disaster dictionary for real-time risk assessment and early warning. With microblogs with location information, we could generate a real-time waterlogging risk map for emergency management.
In the future work, we will attempt to build a larger urban rainstorm and waterlogging text database for a higher accuracy in classification results. In the word segmentation, we focused on reducing the dimensionality of word vectors by more accurate part-of-speech tagging for a rapid classification.

Code and data availability. The data and code used in the study are available at https://github.com/zhr-thu/Real-time-urban-rainstorm-and-waterlogging-disasters-detection-by-Weibo-users.

Acknowledgement. This work was supported by the National Key Research and Development Program of China (grant nos. 2018YFC0807000).

Author contributions. ZHR and SGF conceived the research framework and developed the methodology. ZHR was responsible for the code compilation and data analysis. ZHR and Priscilla O.O. had done the first draft writing. SGF managed the implementation of research activities. ZHR and Priscilla O.O. revised the manuscript. All authors discussed the results and contributed to the final version of the paper.

Competing interests. The authors declare that they have no conflict of interest.

Special issue statement. This article is part of the special issue 'Advances in flood forecasting and early warning'.

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

References

315 Avvenuti, M., Del Vigna, F., Cresci, S., Marchetti, A., & Tesconi, M. (2015, November). Pulling information from social media in the aftermath of unpredictable disasters. In 2015 2nd International Conference on Information and Communication Technologies for Disaster Management (ICT-DM) (pp. 258-264). IEEE, https://doi.org/10.1109/ict-dm.2015.7402058

Bisht, D., Chatterjee, C., Kalakoti, S., Upadhyay, P., Sahoo, M., & Panda, A. (2016). Modeling urban floods and drainage using SWMM and MIKE URBAN: a case study. Natural Hazards, 84(2), 749-776, https://doi.org/10.1007/s11069-016-2455-
320 1

Bo, T. (2018). Application of earthquake disaster data mining and intensity rapid assessment based on social media. Institute of Engineering Mechanics, China Earthquake Administration

Cao, Y. B., Wu, Y. M., & Xu, R. J. (2017). Research about the Perceptible Area Extracted after the Earthquake Based on the Microblog Public Opinion. J Seismol Res, 40(02), 303-310.

325 Caruso C., Quarta F. (1998). Interpolation methods comparison. Computers & Mathematics with Applications, 35(12): 109-126, https://doi.org/10.1016/S0898-1221(98)00101-1.

Cheng, C., Li, Q., Dou, Y., & Wang, Y. (2021). Diurnal Variation and Distribution of Short-Duration Heavy Rainfall in Beijing–Tianjin–Hebei Region in Summer Based on High-Density Automatic Weather Station Data. Atmosphere, 12(10), 1263, https://doi.org/10.3390/atmos12101263

330 Choi, S., & Bae, B. (2015). The real-time monitoring system of social big data for disaster management. In Computer science and its applications (pp. 809-815). Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-662-45402-2_115

Gao, Y., Guo, W., Zhou, H., & Nie, Z. (2014). Improvements of personal weibo clustering algorithm based on K-means. Microcomput Appl, 33(14), 78-81.

Jiang, L. E. I., Chen, Y. A. N. G. B. O., & Wang, H. U. A. N. Y. U. (2015). Urban flood simulation based on the SWMM
335 model. Proceedings of the International Association of Hydrological Sciences, 368, 186-191, https://doi.org/10.5194/piahs-368-186-2015

Lin, T., Liu, X., Song, J., Zhang, G., Jia, Y., Tu, Z., Zheng, Z. and Liu, C., 2018. Urban waterlogging risk assessment based on internet open data: A case study in China. Habitat International, 71, pp.88-96, https://doi.org/10.1016/j.habitatint.2017.11.013

340 Liu, Y., Du, M., Jing, C., & Cai, G. (2014, September). Design and implementation of monitoring and early warning system for urban roads waterlogging. In International Conference on Computer and Computing Technologies in Agriculture (pp. 610-615). Springer, Cham, https://doi.org/10.1007/978-3-319-19620-6_68

Lu, X. S., & Zhou, M. (2016, April). Analyzing the evolution of rare events via social media data and k-means clustering algorithm. In 2016 IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC) (pp. 1-6). IEEE,
345 https://doi.org/10.1109/icnsc.2016.7479041

https://doi.org/10.5194/nhess-2022-48
Preprint. Discussion started: 14 February 2022


Natural Hazards and Earth System Sciences Discussions

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Nair, M., Ramya, G. R., & Sivakumar, P. B. (2017). Usage and analysis of Twitter during 2015 Chennai flood towards disaster management. Procedia computer science, 115, 350-358, https://doi.org/10.1016/j.procs.2017.09.089

Perera, D., Agnihotri, J., Seidou, O., & Djalante, R. (2020). Identifying societal challenges in flood early warning systems. International Journal of Disaster Risk Reduction, 51, 101794, https://doi.org/10.1016/j.ijdrr.2020.101794

Quan, R. (2014). Rainstorm waterlogging risk assessment in central urban area of Shanghai based on multiple scenario simulation. Natural Hazards, 73 (3), 1569–1585, . https://doi.org/10.1007/s11069-014-1156-x

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes twitter users. In Proceedings of the 19th international conference on World wide web (pp. 851-860), https://doi.org/10.1145/1772690.1772777

Tang, X., Shu, Y., Lian, Y., Zhao, Y. and Fu, Y., 2018. A spatial assessment of urban waterlogging risk based on a Weighted Naïve Bayes classifier. Science of the total environment, 630, pp.264-274, https://doi.org/10.1016/j.scitotenv.2018.02.172

Wang, Y., Xiao, S., Guo, Y., & Lv, X. (2013). Research on Chinese micro-blog bursty topics detection. Data Analysis and Knowledge Discovery, 29(2), 57-62, https:// 10.11925/infotech.1003-3513.2013.02.09

Xiao, Y., Li, B., & Gong, Z. (2018). Real-time identification of urban rainstorm waterlogging disasters based on Weibo big data. Natural Hazards, 94(2), 833-842, https://doi.org/10.1007/s11069-018-3427-4

Yin, J., Ye, M., Yin, Z., & Xu, S. (2015). A review of advances in urban flood risk analysis over China. Stochastic Environmental Research and Risk Assessment, 29, 1063–1070, https://doi.org/10.1007/s00477-014-0939-7

Copernicus Publications: https://weibo.com/6215401356/JfG8swIOQ, last access: 26 December 2021.