A multimodal data analysis approach for social media during natural disasters

mengna zhang (541186273@qq.com)
Guizhou University https://orcid.org/0000-0001-7023-170X
qisong huang
Guizhou Minzu University https://orcid.org/0000-0002-8500-9973

Research Article

Keywords: Multi-modal data, LDA, Bert, VGG-16

Posted Date: March 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1033015/v1

A multimodal data analysis approach for social media during natural disasters

Mengna Zhang **Qisong Huang**

Abstract

During natural disaster, social media can provide real-time or low delayed disaster perception information to help government managers carry out disaster response efforts efficiently，therefore, it is of great significance to mining social media information accurately. Distinguished from previous studies, this study proposed a multi-modal data classification model for mining social media information. This model conducted LDA(Late Dirichlet Allocation) to identify subject information from multi-modal data and then analyzed by Bert (bidirectional encoder representation from transformers) and Vgg-16 (visual geometry group 16), text data and image data were classified separately, resulting in real mining of topic information during disasters. This paper used Weibo data during the 2021 Henan heavy storm as the research object, comparing with previous data experiment results, this study proposed a model that can make natural disaster related topic classification of social media data more accurately, resulting in a real-time understanding of different themed natural disasters, it was helpful in making disastrous decisions.

Keywords Multi-modal data· LDA · Bert· VGG-16

1 Introduction

The flood disaster is a high-frequency natural disaster (Wang and Hu et al., 2020), it is occurring worldwide and has a profound impact on national and social development(Ragini, 2018). Since the 21st century, with the rapid development in China, the flood disaster has caused immeasurable economic losses (Chen and Wang, 2020). According to"the water drought disaster Bulletin of China", during the decade 2010-2019, the total value of direct financial loss due to flood disasters has exceeded 234.31 billion RMB in the country. During 2010-2019, the flood disasters occurred in 62% of cities nationwide, 137 cities have experienced more than three episodes of flood disasters. Heavy human injuries, economic losses and traffic are often caused by heavy rainfall because of the clustering of crucial

Mengna Zhang
Zhang_mengna1@163.com
Qisong Huang
Qshuang1976@126.com

Studying in Guizhou University, College of Management, working in Guizhou University of Finance and Economics, Guiyang,550025, China
Guizhou Minzu University, Guiyang, 550025, China

infrastructure such as population, resources, and transportation in the city Congestion, power disruption and other issues.

This paper has the following three contributions：

1、This paper presents a multi-modal system to classify and process multi-modal data from social media.

2、In this paper, the heavy storm disaster topic is more meticulously divided using the LDA theme model, which realizes an automatic near real-time extraction of information on serious storm disasters from social media.

3、Based on Sina Weibo, a multi-classification model is constructed using convolutional neural networks to extract storm-related disaster information such as weather, traffic, and rescue from a large number of social media text streams. At the same time, it visualizes and analyzes different rainstorm disaster themes in terms of relative quantity and spatial attributes, explores the time trend of disaster development and spatial distribution characteristics of rainstorm events.

2 Background information

After a natural disaster, rescue organizations needed to use extensive data information in the initial phase as a decision basis to make low-risk decisions quickly(Sadiq, 2020), still the natural disasters could cause signal interruption, so obtaining useful data information became an urgent problem. In the past, because of the lack of data information, experts made emergency decisions mainly rely on their own knowledge experience(Xu, 2020). In recent years, with the rapid development of social network, it had provided an essential platform for information dissemination(Nguyen, 2017), after the occurrence of natural disasters, hundreds of millions of people would release their ideas and hearing on the network in the form of words, pictures, audio, video, and so on, forming a vast amount of data information, so the valuable information it contains could be used as a basis for making emergency decisions in the literature(Abid and Li et al., 2020). (Behl and Rao et al., 2021)proposed that the sudden and urgent nature of emergencies requires that crisis managers must keep current and meet the critical information needs of the public, so researchers use social media as a source of information for crisis management. (Kitazawa, 2021)proposed that the rapid development and application of social media in crisis communication, the effective of which improved the efficiency of crisis communication, and then enhanced emergency response methods, reduced the cost of disasters, and increased transparency and democratizism in decision-making in the literature.(Kitazawa, 2021) proposed that in an emergency event, social media contains a large amount of subject matter, spatiotemporal and other emergency information, by classifying the real-time and massive emergency

information, which is able to identify the subject matter information such as the fact, rescue and impact of the event, so that it is beneficial to understand the status of the emergency event.

Most of the previous studies on social media data focused on a single data form (eg, text or visual data), and literature.(Piatyszek, 2012)adopted a logistic regression algorithm to classify text data and thus detect damage and injuries caused by Sri Lankan flooding, still, the overall accuracy of this classifier was only 0.647 due to the overall sample size and too small subject sample size, (Yu, 2019)used the CNN classifier to classify text data and detect damages and damages caused by hurricanes by using the text data of hurricanes Sandy, Harvey as the research objects, resulting in relevant recommendations for donation and assistance.

To increase the precision, we needed to analyze the other modal data included while classifying textual information(Nguyen, 2017). The analysis of seismic image information, from which human body parts are examined from debris, provided an adequate basis for developing seismic rescue measures with a precision of 0.8037, still, seismic rescue efforts need to obtain geographic location information in addition to accurate image information, effective combination of both kinds of information, to get precise information on people trapped in earthquakes. (Aznar-Crespo, 2021)classified disaster social media images into three categories, severe, mild, and no damage, to analyze the effects caused by natural disasters, to develop related assistance measures and mine their corresponding text information while classifying image information, which could further improve the accuracy of classification.

Recently, the form of people's expression views in social media platforms had significantly changed, and people prefered diverse expression ways such as text, images and video, so the multi-modal data contains richer information and can more accurately described the natural disaster situation(Kumar, 2020). Therefore, we needed to establish a model for multi-modal data analysis. Multi-modal data analysis was a very challenging task at the moment(Rasiwasia, 2010). A unified cross-media word bag model was constructed for both text and image, the model acquired the representation of text and image and used a logistic regression classifier, and the experimental results showed that the preparation rate of information classification for analyzing both text and image modal data on performance was 4% higher than the text-based method(Yu, 2016). (Kaplan and Haenlein, 2010)also used 2CNN structures, which extracted text data features and image data features separately, performed significantly better than existing models that used only text or visual content.

In natural disasters, relevant studies at home and abroad proved that social media data could be applied to real-time monitoring and trend prediction of disaster events (Hao and Wang, 2020). Disaster related text data were usually thematically classified, as in the literature(Ghafarian and Yazdi, 2020),Combined LDA and SVM to construct a theme classification model, which divided the typhoon

"" moranti "" related microblogs into 4 themes "" early warning information "", "" disaster information "", "" irrelevant information "" and "" rescue information "", (Ghosh and Srijith et al., 2017)proposed an LDA subject based event detection model in which multi-modal information was used to increase the number of acquired event descriptions, the multi-modal data were combined into the subject model, which all achieved a classification accuracy of 81%, in addition, (Wood and Sanders et al., 2021)designed a system named m-trend, based on the tweet containing geographic information, to construct and visualize the spatiotemporal variation trends of the display theme, explore the trend and spatial distribution law of disaster events,(Ghafarian and Yazdi, 2020),Thematic model analysis of user developed text information on microblogs generated before and after the onset of a heavy storm in Beijing in 2012 revealed differences in the temporal and spatial distribution of microblogs across themes.

3 Multimodal classification model

This article takes the heavy rain in Henan in 2021 as an example. Since the night of July 17 in 2021, Henan Province has experienced heavy rains and locally heavy rains. The average rainfall in the province is 73.1 mm. As of noon on August 2, a total of 150 counties (cities, districts), 1663 townships and 14.531,600 people are affected in Henan Province. The whole province has organized emergency avoidance of 933,800 people and relocation of 1,470,800 people. The heavy rain resulted in the collapse of 89001 houses, the area affected by crops is 16.356 million mu, and the direct economic loss is 114.269 billion yuan. In heavy rain, 302 people are killed and 50 people are missing. Therefore, this article combines the Weibo API and web crawler to obtain a total of 28,099 pieces of data from 0:00 on July 18 in 2021 to 23:00 on July 30 in 2021, using "Henan rainstorm" as the keywords.

The research of this paper focuses on automatically locating and mining natural disaster information from images and text data in social media. A real-time classification and positioning model of emergency topics based on social text and images is proposed. The flow chart is shown in (Fig.1). The model consists of the following modules: The data processing module is responsible for data collection and processing. Because the correlation between images and text data is weak, this article treats text and image data separately. The topic mining module is responsible for the mining of hidden topics. By mining topic information, we can fully understand the situation of emergencies. The topic analysis module is responsible for analyzing the text and image data.

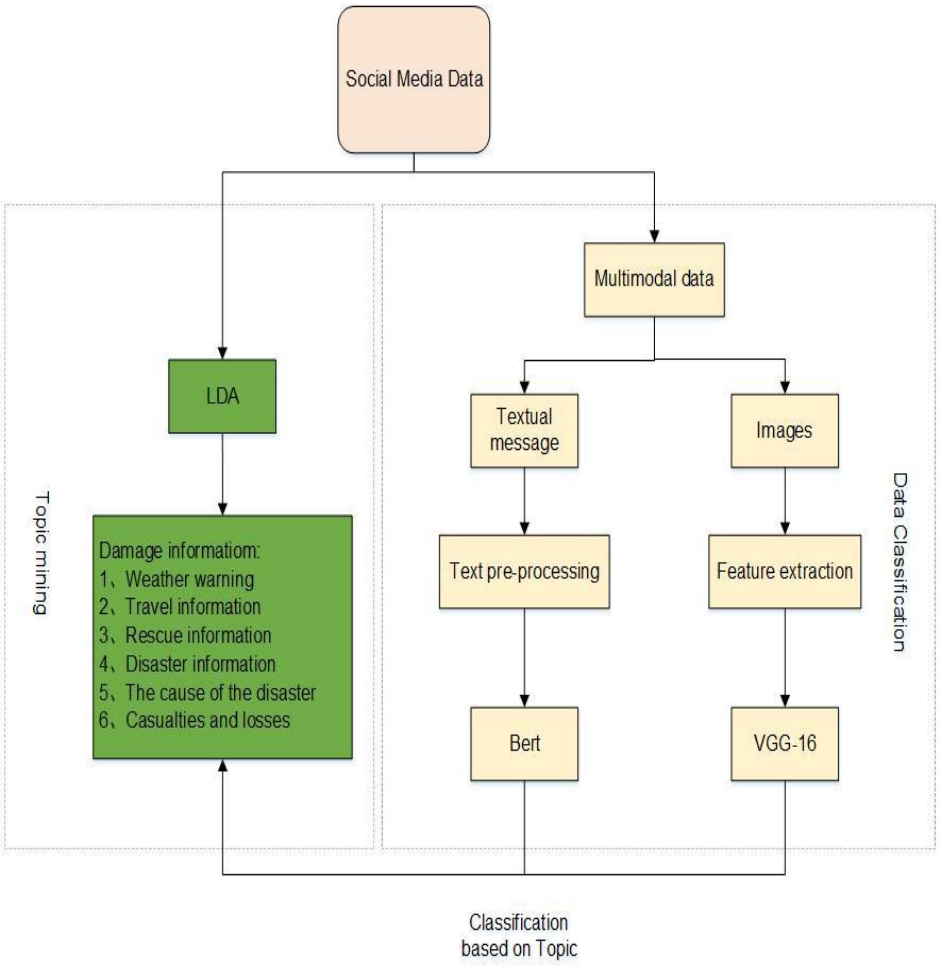

Fig.1 A structure diagram of Multi-modal data classification

3.1 Acquisition and preprocessing of data

This article uses Weibo as the research object. Weibo is a popular social media platform. Especially during natural disasters, there are hundreds of thousands of Weibo data about natural disasters every day. During natural disasters, people report disaster information through Weibo, express urgent needs and seek help. As a result, Weibo data has became an important data source for disaster management. We can use text, images and geographic location data to learn more about natural disasters and to provide data basis for natural disaster management.

3.2 Topic mining

A topic model is an essential tool for data mining of social media and has attracted extensive attention in recent years. Empirical research has found that the release of social data information is closely related to the cycle of disaster occurrence, Emergency rescue is usually divided into three stages: pre-disaster preparation, emergency response, and post-disaster recovery. People discuss

different topics on social media at different stages. In the early days of a disaster, people discussed content mainly about disaster preparedness and weather warning. After the disaster, people focused on disaster discussion and emergency rescue. In the later stage of the disaster, people released content mainly focused on post-disaster recovery and reconstruction. So different themes in different periods.

Latent Dirichlet Allocation (LDA) is the most representative topic model(Gupta and Katarya, 2021). The LDA model is a typical generative model, which is mainly used in the text processing(Kang and Ren et al., 2021). Due to the emergence of the BOW model, it is currently widely used in the area of image labeling. The text uses the LDA model for topic mining of image data and text data(Park and Yoon et al., 2021). The core idea of the LDA model is: regard topics as the probability distribution of text words, different topics correspond to other text word distributions.In the field of image annotation, we need to first extract the low-level features of the image and perform clustering. Then, we use the clustering algorithm to vectorize the low-level features of the image into visual words, finally we use the BOW model to convert the image into a set of visual words. The LDA probability graph model is shown in (Fig.2). The symbols used in the model and their meanings are shown in (Table 1).

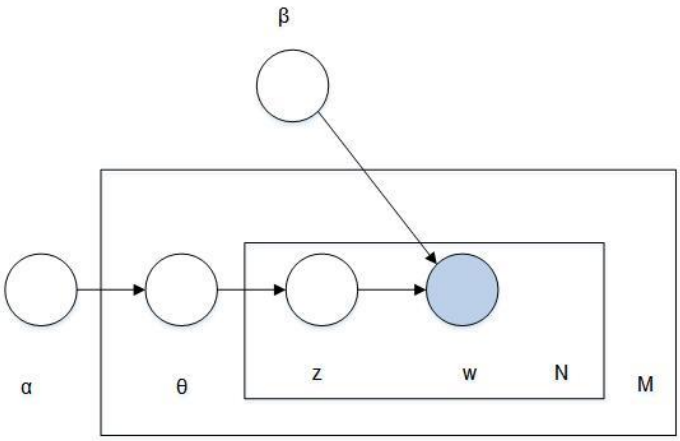

Fig.2 Probabilistic graphical model of LDA model

Table1 Symbols and their meanings in Fig. 1

symbol	Symbolic meaning	symbol	Symbolic meaning
M	Training set size	N	Number of words
K	Number of topics	ω	words
z	Potential topic	Θ	Theme ratio
α	Model parameters	β	Model parameters
γ	α Variational parameters	φ	B Variational parameters
P_{dir}	Dirichlet distribution	Mult	Variational parameters

3.3 Text data classification model

The use of social media texts for natural disaster assessment is divided into three areas. First, we

need to identify whether the tweets are related to rainstorm damage, this is a two-classification task. Secondly, the rainstorm damage is divided into multiple categories according to the theme, this is a multi-classification task. Finally, we classify the text data according to the theme and generate a damage report.

The process of subject classification is shown in (Fig.3), First, we use Bert to build a relational classification model to identify text segments corresponding to two aspects, namely entity tags. Then we use the predicted relationship and text to build an entity extraction model using Bert, that is, a multi-classification task. We divide each sentence into a three-tuple of description object, damage description and damage result. Corresponds to relationship.

Bert is a huge pre-trained model that shows excellent performance when generating entity embeddings of text(Qiao, 2021). The Bert model improves the performance of text classification through entity embedding, so it can identify tweets and generate corresponding damage reports, finally we classify the damage reports into corresponding topics.

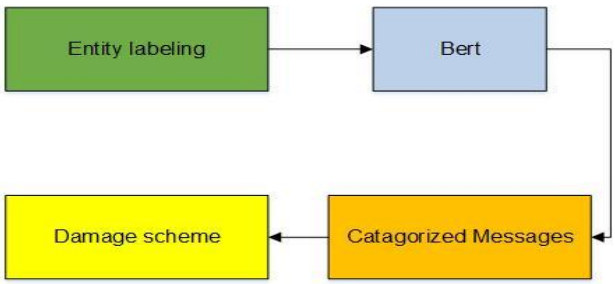

Fig.3 The framework of topic classification

3.3.1 Damage relationship definition

Topic classification based on social media text data requires identifing tweets with a harmed relationship, which can be abstracted into a triad with three aspects describing the object. Common victims in heavy rain include roads, people, houses, water and electricity supply, etc. The description of the damage is related to the feature words corresponding to the description object, and the damage result describes the ultimate severity of the damage. For example, 10600 houses collapsed, this sentence describes the object as houses, the damage is described as 10600 houses, and the result of the damage is collapsed. So it is possible to classify damage reports into the corresponding. We can identify whether the tweets are related to the rainstorm based on the damage report, secondly classify the damage report by topic, finallyclassify the damage report into the corresponding topic.

The word collocation in this paper is based on the Chinese word collocation database SogouR, and we summarize and summarize the collocations of the Weibo texts of the rainstorm event. We randomly selected 5000 pieces of text information from the rainstorm event, analyzed the grammatical characteristics of the disaster information, and finally obtained the lexical rules shown in (Table 2),

thereby obtaining the collocation relationship between the expression description object and the damage result.

Table 2 The lexical rule pattern

Pattern rule	Text word
v-n	Shattered glass everywhere
n-v	The whole village was blown to the ground
a-n	Broken window glass in one place
n-a	The road has been blocked
d-vi	Soon the community will no longer supply water
v-vi	About to stop power supply
r-v	Saw him smashed by a tree
v-r	he branch was blown off by the wind just hit him
vi	Power outage for one day today

Note：v is verb；n is noun；a is adjective；d is adverb；r is pronoun；v is Intransitive verb.

We present examples of identifying whether a tweet is a heavy rain related one based on a damage report.

Negative example: heavy rains in Henan in 2021, the tribute to the most cute man! The tribute to people's younger cousins!

Positive example: the father was washed away by a flood at 2 pm on 20 July 2021 at sukangcun Shi River, Takayama Town, Xingyang, from south to North and in the direction of the downstream fenggou. During falling, the upper body was covered with a white spot under the curve and the lower body wore sport pants.

Although the negative example contains the description of the heavy rain in Henan, it does not damage the description and the related content of the damage, so the tweet cannot be regarded as a tweet related to the heavy rain,.The positive example includes the description of the subject’s father, also it includes the description of the damage in Baohe, Zhonggang Village, Gaoshan Town, Xingyang City. The result of the damage is washed away and can constitute a complete damage report. Therefore, the tweet is a tweet related to torrential rain.

3.3.2 Constructing word pairing rules

Based on lexical rules to extract feature words in small-scale annotated corpus, we use feature word collocation as the original word pair to construct the classification knowledge base. Based on this, the word vector model and the extended version of "Synonyms Clin" are used to enrich the collocation information of characteristic words to realize the diversity of Chinese expression.

In this paper, based on the results of subject mining, the original words are defined as six aspects: weather warning, traffic situation, rescue information, disaster information, disaster cause, casualty and damage, according to 6 aspects, a dictionary of description objects and damage results is established as shown in (Table 3).

Table 3 The dictionary of description objects and damage results

Description Object	Damage Description	Damage Result
Weather	Southern North China，Henan	Rainstorm，Moderate to heavy rain，Continuous heavy rainfall
traffic situation	Railway Line 2,Platform,road	Pause, Adjustment operation, Temporary closure, Blocked
rescue information,,	Xiao Pengpeng, Huang xinrui, Genghuang Central Primary School	Lost contact, Lost contact, bedding
Infrastructure	Civic Center Station, Escalator, College Road, East Coach Station	service paused, Temporary closure, Sever diplomatic relations
Hazard Factor	Worldwide, Western Pacific Subtropical, Dongfeng	high temperature, high pressure, Rapids
Water or Power Supply	Anyang, village, outdoor	Water and power outages, Fetch water
traffic situation	Railway Line 2,Platform,road	Pause, Adjustment operation, Temporary closure, Blocked

In the field of natural language processing, the word vector model is used to calculate the distance between words usually, two words that are close in distance are also highly correlated, thus realizing the expansion of feature word collocation the commonly used word vector model contains cbow and skip gram models. For data with less than 100 million words, the performance of the Skip-gram model is better(Hung and Yamanishi, 2021), so this paper uses the skip gram model to calculate the phase between words Relevance. Its model structure is shown in (Fig.4). Frontal context information is predicted by the current word $W(t)$ for the etymological sequence in which the word resides.

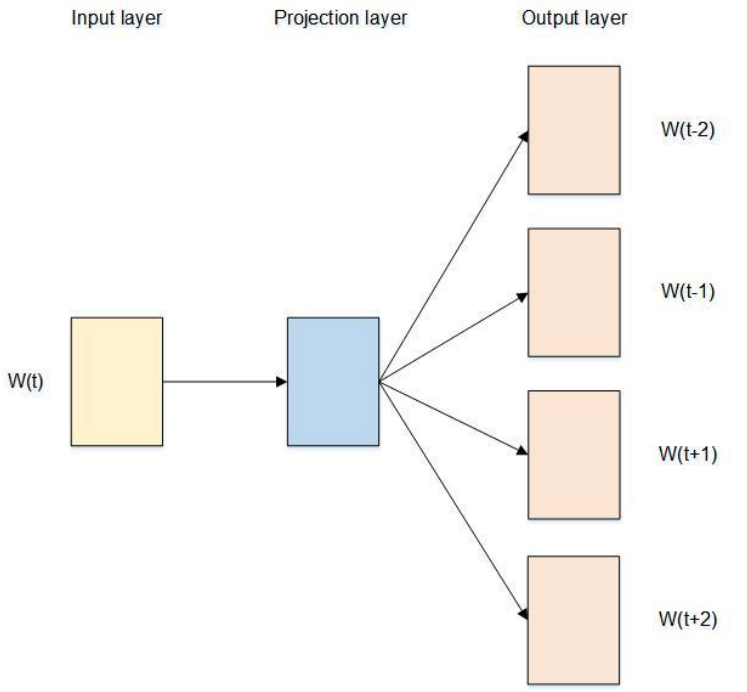

Fig.4 The model structure diagram

3.4 Image data classification model

Compared with social media text data, images convey information is more objective and valuable, still, few studies have utilized graphical data for natural disaster damage assessment, so this paper uses image and text data for multi-modal data analysis to make damage assessment more objective and accurate.

Processing image data is performed by converting the image into a digital feature vector and then using the classifier for image classification. The classifiers are responsible for different tasks that define the semantic hierarchy.

In this paper, Vgg-16 is used to extract image feature. VGG-16 is a classic network of convolutional neural networks(Mehmood and Attique Khan et al., 2022). VGG-16 convolutional neural network use the small convolution kernel 3×3 and the largest pooling layer 2×2 . The model stacks convolutional layers several times the number of layers of the standard CNN model. With the maximum pooling layer, not only can the parameters be reduced, the amount of calculations can be reduced, and the model can also improve the model's ability to express nonlinear data(Basalamah and Rahman, 2022).

The structure of the image feature extraction model based on Vgg-16 is shown in（Fig.5. The input of the Vgg-16 network was fixed in size 224×224 . Opencv is an open-source computer vision library that utilizes the resize() function, The function uniformly scales the images of the data set to

224×224 which can as input to the VGG-16 network.

Keep the remaining network structure of the original VGG-16 model encapsulated in the Keras deep learning library except for the fully connected layer. In order to realize the feature-based transfer learning method in the homogeneous space, We choose the VGG-16 model parameters pre-trained by ImageNet as the initial value of the feature extraction model, that is, the weight parameter wights of VGG-16 is set to "image net". Adding the Flatten layer is used to make the multi-dimensional input one-dimensional, and then input the one-dimensional vector into the Dense layer using ReLu as the activation function, and finally add the dropout layer to obtain the image feature extraction model based on VGG-16.

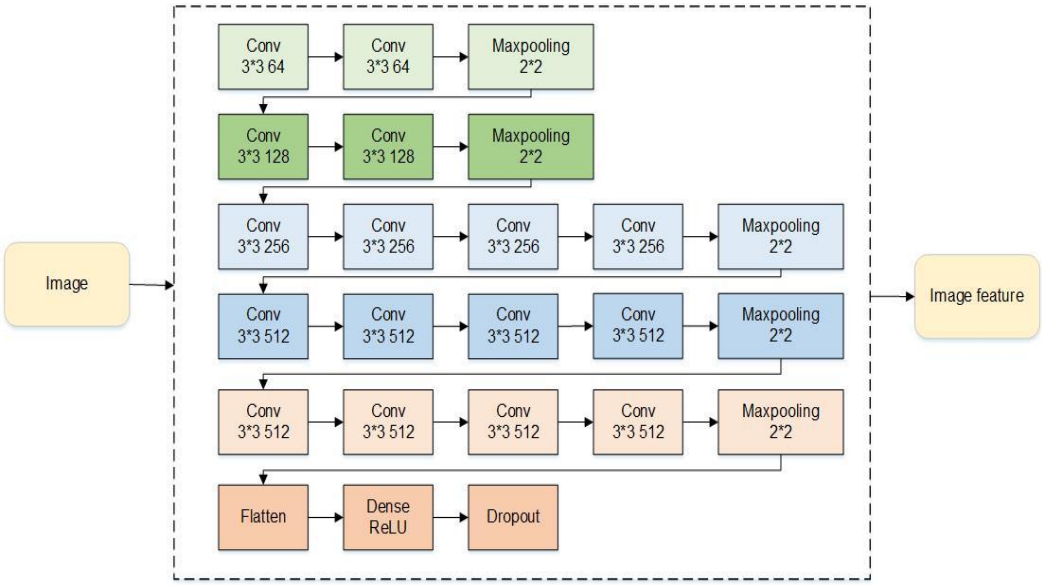

Fig.5 The image feature extraction model structure diagram

3.4.1 adaboost classifier

In this paper, we switched the softmax classifier from the original model of Vgg-16 into an AdaBoost classifier with better classification performance. The working principle of the Adaboost classifier is to train multiple different weak classifiers from a training set, and retrain each time by combining the last training sample with the new sample to obtain a new classifier, finally form a stronger classifier for the model Classification. The AdaBoost iteration algorithm is divided into three steps.

1) Initializing the distribution of the weights for the training data. If there are N samples, each is given the same weight at the very beginning： $\frac{1}{N}$ 。

$$D_1 = (w_{11}, w_{22} \cdots w_{1i} \cdots, w_{1N}), w_{1i} = \frac{1}{N}, i = 1,2,\cdots, N \quad (1)$$

2) Training weak classifiers. In the specific training process, If a sample point has been accurately classified, its weight will be reduced in the construction of the next training set; on the contrary, if a sample point has not been accurately classified, its weight will be increased. Then the sample set with updated weights is used to train the next classifier, and the whole training process proceeds iteratively in this way.Performing multiple iterations, use m = 1,2, ..., M to indicate the number of iterations.

3) Combining the weak classifiers obtained from each training into a robust classifier. After the training process of each weak classifier is over, increasing the weight of the weak classifier with a small classification error rate to make it play a more significant decisive role in the final classification function and reduce the weak classifier with significant classification error rate. The weight makes it play a more minor decisive role in the last classification function. In other words, a weak classifier with a low error rate occupies more significant weight in the final classifier, otherwise, it is smaller.

4 Research result

In this paper, the number of social media per day during the heavy rains in Henan is counted, a statistical graph of the number of tweets based on time is shown in (Fig.6). Since the night of July 17 in 2021, heavy rains began to attract widespread public attention, on July 20 in 2021, the emergency response level was elevated from level IV to level II by the Henan Provincial command on drought resistance due to the very severe prevention situation, according to Statistical (Fig.6) we can see that the number of microblogs started to rise significantly on July 20 in response to the increasing severity of the forms of flooding, and the mobile communication network of the province was fully restored and the supply of water was restored in most areas starting on July 25, thus the public concern for heavy rains in Henan gradually declined as typhoons "" fireworks "" on July 27.Affect Henan, so the number of microblogs showed an obvious upward trend. Thus, the temporal trend of microblog data largely coincides with the real time of the event, suggesting that Sina's original information on microblogs has a usage value when a major emergency occurred.

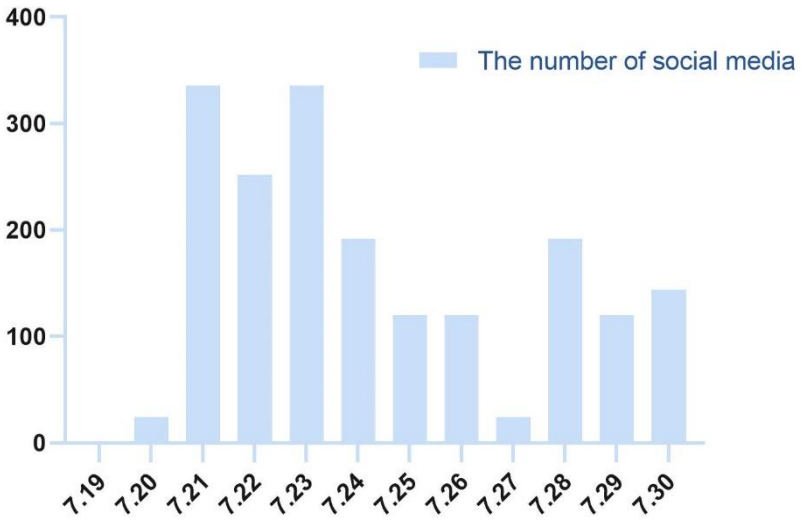

Fig.6 The count chart of the tweet

4.1 Topic mining

This paper extracts 2219 pictures related to Henan rainstorm from Weibo, randomly selects 20% pictures as the test set, and the remaining 80% pictures as the training set. The image annotation results obtained by the LDA model are shown in (Table 4). The image size is set to 192×168 , obtaining an average number of annotated words per picture of 4.5, an average number of annotated images for each annotated word is 58.6, a total of 170 annotated words in the image set, and annotated words with fewer annotated images are eliminated, and the others 120 words form the label vocabulary.

Table 4 The theme extraction of picture

Image	Topic model
	People, umbrella, bicycle, rainstorm, tree
	Houses, tree, flood, people

Deduplication, Chinese word segmentation, stop words removal and emoji preprocessing were performed on the 25880 Weibo data sets obtained from Weibo, and the vocabulary expression of each Weibo was obtained, and the data was manually labeled to obtain the corresponding Weibo Vocabulary collection and emergency themes. In order to verify the accuracy of the model classification, 20% of the samples were randomly selected as the test set, and the remaining 80% were used as the training set.

Using the LDA model as the topic classification model, the topic distribution of the sample documents and the respective feature vocabulary distribution of all topics were obtained. Some themes are shown in (Fig.7).

#Topic4		#Topic17		#Topic18		#Topic27		#Topic29	
Airport	0.021267	Stagnant wat	0.096204	Early warning	0.042271	Rescue	0.077723	Die	0.132123
Hour	0.052066	Drain	0.039654	Rainfall	0.040495	Encounter	0.060422	Victims	0.049918
Stay	0.027527	Serious	0.037401	Area	0.040141	Assault	0.048871	Disaster	0.039559
Subway	0.022227	Transportatio	0.023929	Maximum	0.035745	Come on	0.036526	Confirm	0.034913
Traveler	0.021267	Paralysis	0.018862	Rain	0.023859	Strongest	0.033091	Identity	0.032775
Sit	0.014818	Road section	0.017362	Hour	0.021661	[Microphone]	0.028187	Die	0.032523
Train	0.014723	No	0.016191	Part	0.020818	Unable to	0.027953	People	0.030909
Real	0.014431	Center	0.015243	Predict	0.019912	To	0.027324	Extra large	0.029438
Late	0.014421	Rainwater	0.014202	Reach	0.019417	Fire fighting	0.025341	Lethal	0.023085
Station	0.014053	Road	0.011039	Rainfall	0.019362	Serious	0.024981	To	0.022143
No one	0.013371	Cause	0.013339	Last	0.015673	On site	0.022374	Announce	0.022075
Line	0.011875	Overpass	0.011919	Weather stati	0.015664	Begging	0.020939	Disaster	0.021528
Once in a hundred years	0.011562	Pavement	0.011039	Citywide	0.015511	Dispatch	0.019004	Drowning	0.021184
Influence	0.010943	Long	0.009032	Urban area	0.015369	Condition	0.018881	Discover	0.021051
Train	0.010693	Map	0.008984	Orange	0.014722	Go to	0.018615	Happen	0.018821
Bus	0.010651	Expert	0.008719	Blue	0.014547	Suffer	0.018542	Houses	0.018143
Capital Airport	0.010156	Municipal	0.007733	Yellow	0.014382	Trapped	0.017739	List	0.017801
Out of service	0.010138	Report	0.007567	Average	0.013896	Diffusion	0.017288	Deceased	0.014983
Ask	0.009949	Department	0.007351	Mudslide	0.013396	Submerged	0.015447	Collapse	0.014968
Arrive	0.009792	Wish to be happy and prosperous	0.007043	Influence	0.013292	Personnel	0.014372	Electric shock	0.014137
Frequent visitor	0.009504	River	0.006357	Signal	0.013012	Police force	0.013492	Lightning strike	0.013532
. . .		...		...		...		...	

Fig.7 The part of the theme distribution

Through thematic classification of text data, we finally got 40 thematic categories, as shown in(Fig.7), through lexical analysis of thematic distribution, both topic29 and topic33 discussion themes were about the losses and impacts caused by heavy rains in Henan, therefore, we combined similar themes, and finally would get 40 thematic categories combined to get "" weather warning "", "" traffic situation "",Six emergency information related topics including "" rescue information "", "" disaster information "", "" disaster cause "", "" casualties and losses "", etc. the classification of their topics is shown in (Table 5).

Table 5 Social media classification scheme

	Class	description	example
1	Weather warning	Warning given about the change of the weather	According to the latest weather forecast by the Meteorological Bureau, it is expected that there will be a heavy rainfall in zhengzhou from July 22th to 22th
2	traffic condition	The traffic obstruction and the damage to vehicles	At 4 pm on July 20, 2021, a lot of rain poured into the platform layer of the Huiji District Government Station of Zhengzhou Metro Line 2
3	Rescue information	Provide goods and services needed by victims	On July 20, 2021, Gongyi, Henan was hit by heavy rain, and the Yichuan Condor rescue team rushed to the disaster area overnight for rescue
4	Disaster information	The information about the level and duration of the rainstorm	Continuous heavy rainfall has caused the flooding of roads, subways and other public facilities in many places in Henan
5	The cause of the disaster	The discussion of the cause about the rainstorm	More rainfall in the north this year the most important reason is the abnormally northerly subtropical high
6	Casualties and Damage	Information about casualties or infrastructure damage	Wang Yufeng walked to Sizhuang Village after about 2:40 and lost contact

4.2 Disaster type classification as well as severity information

The Weibo data were classified according to the subject classification results, which ultimately yielded the natural disaster situation for each region of Henan. (Fig.8) presents the number of subject social media during heavy rains in Henan,(Fig.9) presents the change of subject microblogs over time during heavy rains in Henan. It can be concluded by (Fig.8)and(Fig.9)that relatively little attention has been paid to weather warning during heavy rains, but it began on July 19, Henan experienced heavy storm, people started to release weather warning information via Weibo, so the number of Weibo with weather warning on July 19 was significantly more than that of other subjects, and on July 27 the typhoon "fireworks"Affected Henan, so the number of microblogs regarding weather warning obviously increased on July 27. With the development of catastrophes, the number of microblogs regarding disaster information rapidly increased on July 19.On July 24, as rainfall declined, people's concern about disaster information gradually decreased. Heavy rains caused huge damage to Henan, therefore, it can be seen through (Fig.8) that people discuss, the topic is mainly focused on rescue

information and casualty loss. From July 20, there was an explosion of social media about rescue information and casualty loss, and the concern about rescue information and casualties and loss was much higher than other topics throughout the storm. On July 26, as rainfall decreased, the concern about rescue information and casualties and loss began, The degree showed a decreasing trend, People's attention to traffic information and the causes of disaster situations during heavy rains was generally low, there were a few discussions about disaster information and traffic information during the period of storm disaster emergencies from July 21 to July 22. Therefore, during the heavy rains in Henan, people paid more attention to rescue information and the relationship between casualties and loss situations.

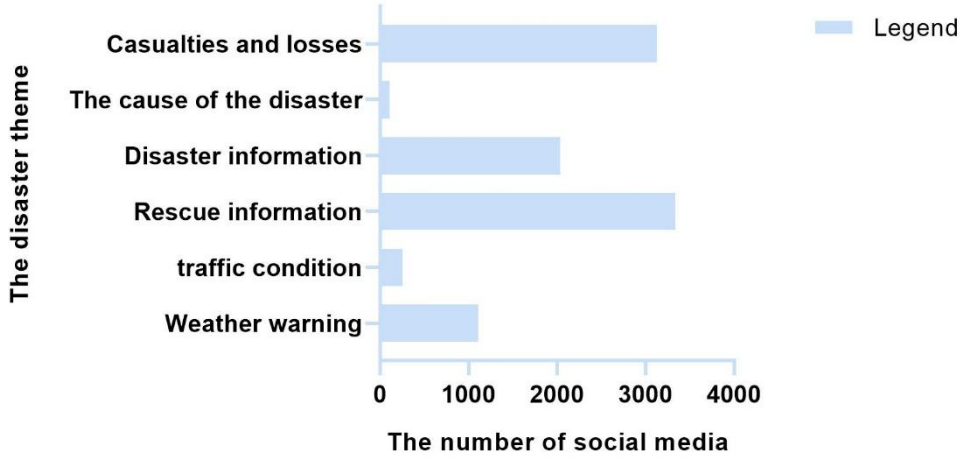

Fig.8 The histogram of natural disaster classification

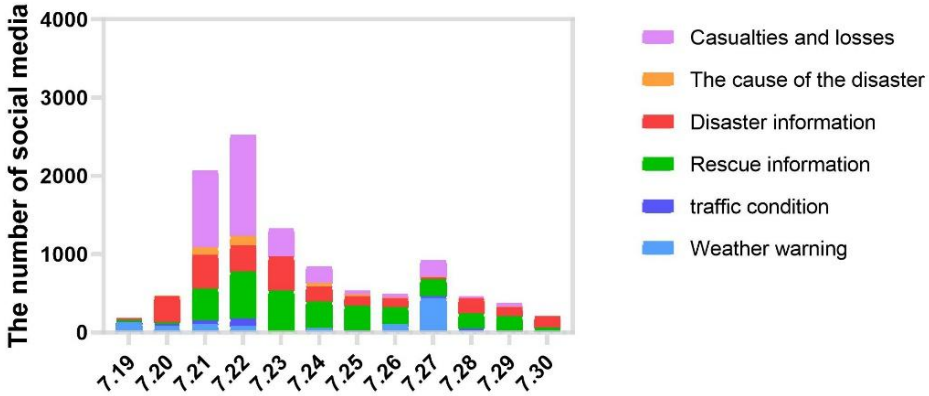

Fig.9 The Number of weibo in every topic

When natural disasters occur, we need to focus on the geographical distribution of disaster occurrence as well as on the disaster severity, therefore, we fully mined the geographical location information in Weibo tweets, shown in (Fig.10) as a regional distribution map of social media as well as a map of social media quantity distribution. It can be seen from (Fig.10)that there were relatively more heavy rain related Weibo numbers released from Zhengzhou and Xinxiang regions during the

heavy rains in Henan, secondly, Hebi, Anyang, and Luoyang released a certain number of Weibo about heavy rain.(Fig.11) calculates the damage reporting ratio of all tweets in each city. Based on the data contrasted in (Fig.10)and (Fig.11), on the one hand, the number of Weibo posted in each area is related to the severity of the rainstorm, on the other hand, it is related to the population density of the city. Zhengzhou is subjected to heavy rains during heavy rains. There are heavy rains affecting 10.352 million people at the same time as Zhengzhou, thus, Zhengzhou released the largest number of microblogs, at the same time, Xinxiang City was severely affected by heavy rains, but the population of Xinxiang was 6.043 million people, thus, the number of microblogs in Xinxiang with regard to heavy rains is less than that of Zhengzhou, At the same time, Anyang City, Hebi City, and Luoyang City were all seriously affected by heavy rains. However, the population of Anyang City is 5.192 million, that of Luoyang City is 6.69 million, Hebi City is 1.609 million. Therefore, The population of Hebi City is significantly less than that of Hebi City and Anyang City. Although Hebi City, Anyang City, and Luoyang City have almost the same number of microblogs about rainstorms, it can be inferred that Hebi City is more affected by rainstorms than Anyang City and Luoyang City.

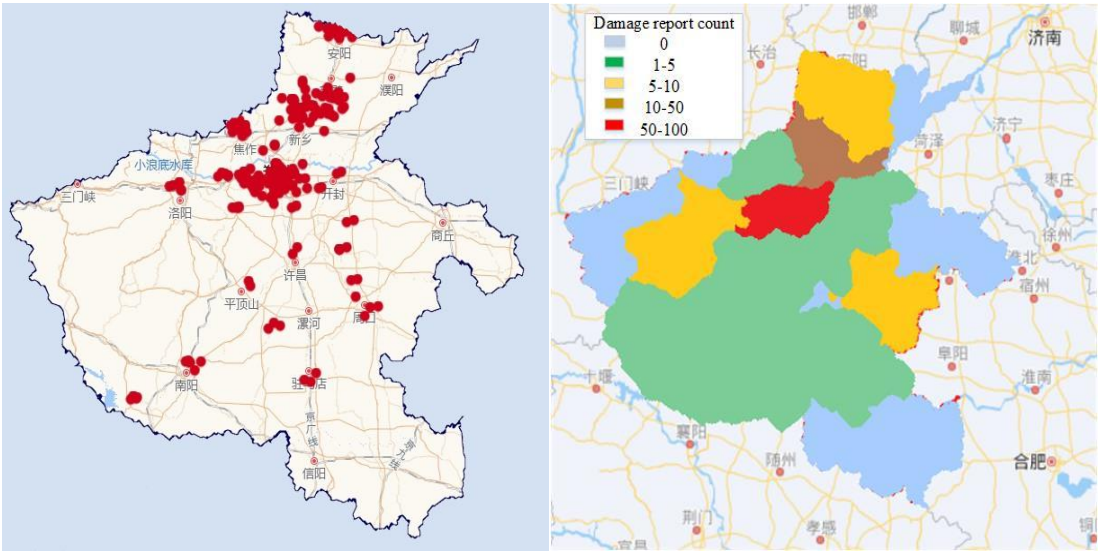

Fig.10 The space distribution map and the number of loss reports by region

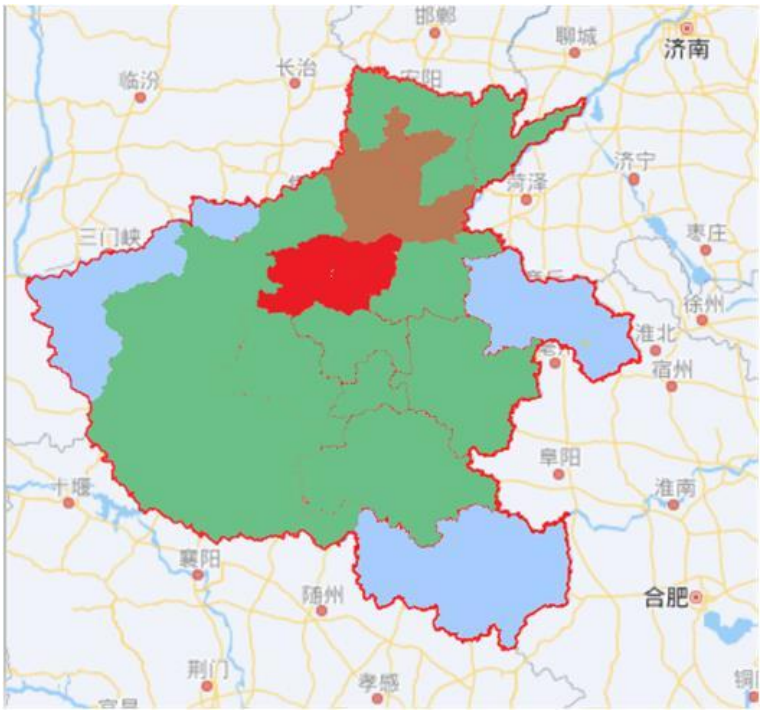

Fig.11 The loss reported rate by region

Because more attention has been paid to rescue information and casualty and loss information during heavy rains, the geographical location part of the rescue information and casualty and loss information is discussed separately in this paper. (Fig.12)shows the regional distribution plots of microblogs with respect to rescue information during heavy rains, as well as the distribution plots of microblog numbers. (Fig.12)shows the microblog area distribution map and the number of microblogs about rescue information during the heavy rain, the number of rescued information microblogs are larger, and there are few discussions about rescue information.The population size of Zhengzhou is 10.352 million, the population size of Xinxiang is 6.043 million, the population size of modifcation was 6. 922 million, the population size of Anyang is 5.19 million, the population size of Zhoukou is 8.8 million, the population size of Zhu madian is 6.89 million, the population of Hebi City is 1.6 million. According to the ratio of the number of microblogs related to rescue information in (Fig.13), we can infer that Zhengzhou City, Xinxiang City, Hebi City and Anyang City are compared, although the number of microblogs related to rescue information in Hebi City Slightly less than Anyang City, but the population of Hebi City is significantly less than Anyang City, so the rescue demand in Hebi City is higher than Anyang City, at the same time, Luoyang City, Kaifeng City, Zhoukou City, and Zhumadian City have issued a certain number of Rescue information, we need to pay attention to the rescue needs of the area

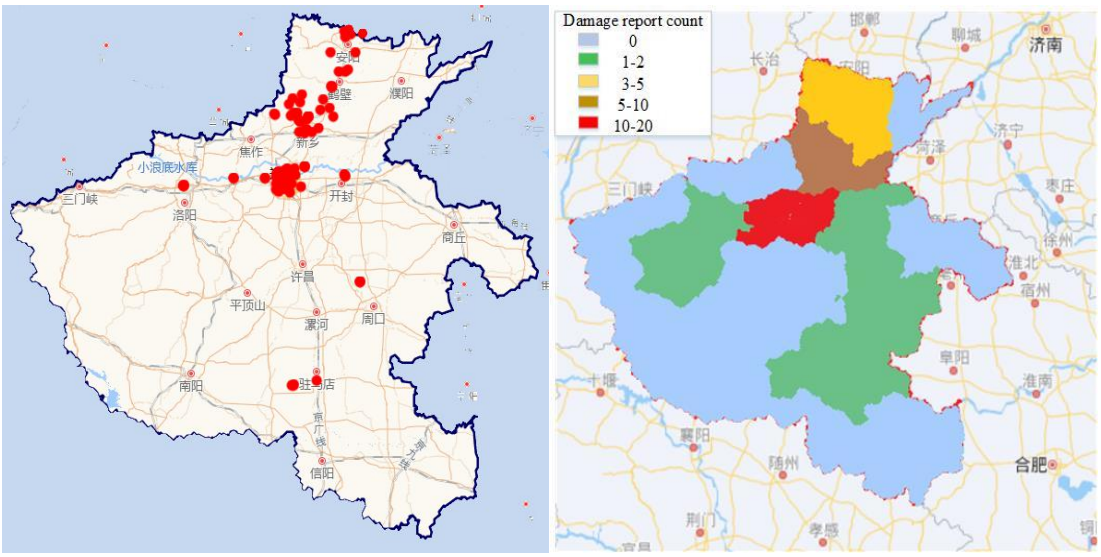

Fig. 12 The area distribution map and number distribution map of rescue information

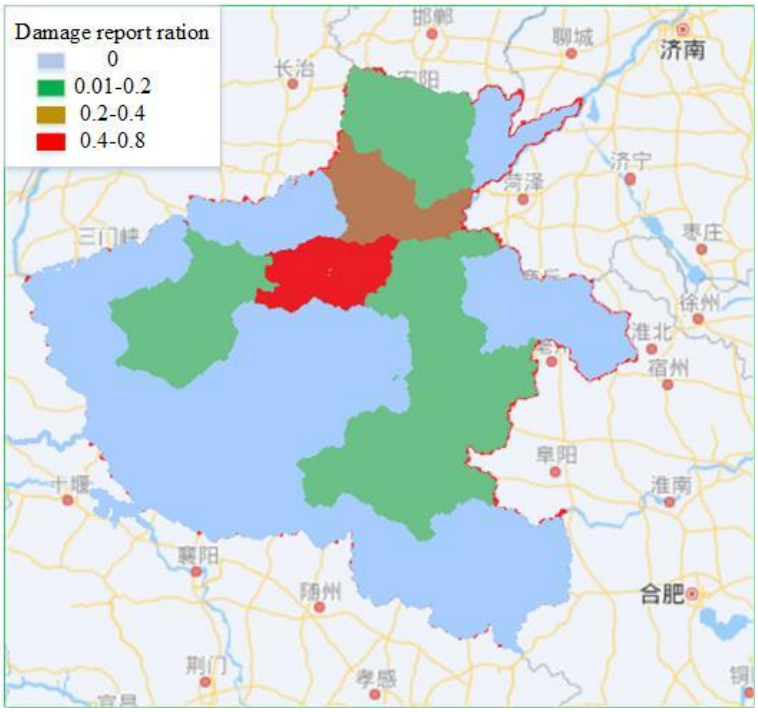

Fig.13 The loss reported rate by region

(Fig.14)shows the microblog area distribution map regarding casualties and losses during the heavy rain and the distribution map of the number of microblogs. It can be seen from (Fig14) that Weibo about casualties and losses are mainly concentrated in the Zhengzhou area, followed by a certain number of Weibo posts in Xinxiang and Zhoukou City, while Hebi and Luoyang also posted some Weibo about casualties and losses. By comparing the casualty and loss ratio chart in (Fig15), we

can infer that the Weibo postings of casualties and losses are mainly concentrated in Zhengzhou City, because Zhengzhou City has a greater impact on heavy rains and Zhengzhou has a large population. At the same time, Hebi, Anyang, A small number of casualties and loss microblogs were posted in Zhumadian and Luoyang. It can be seen that the number of casualties and loss microblogs is related to the severity of the heavy rains in each city and the population density. At the same time, compared with the number of rescue information microblogs, the distribution of casualties and losses is more concentrated in Zhengzhou. This is because during heavy rains, people use microblogs to seek help and find missing persons, and such microblogs will attract more public attention. A large number of reposts are generated. Therefore, the Weibo location information of casualties and losses is more accurate and more concentrated.

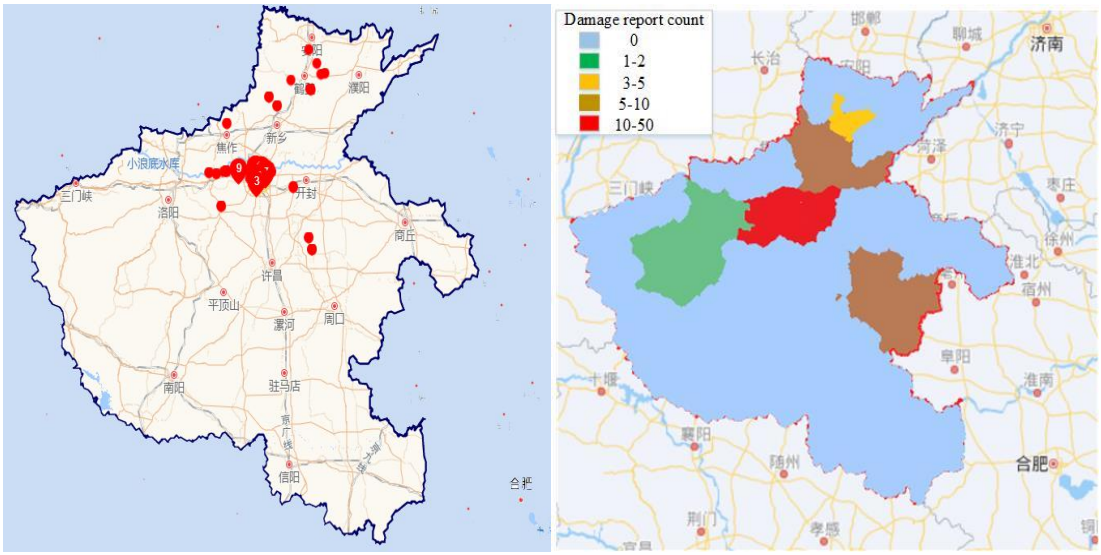

Fig.14 The area distribution map and number distribution map of casualties and losses

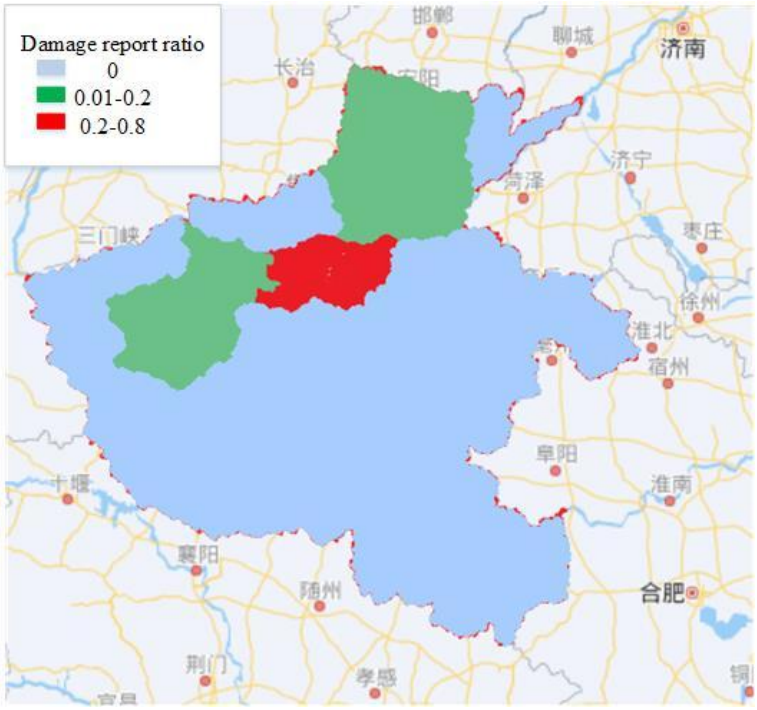

Fig.15 The casualties and losses reported rate by region

4.3 Evaluating indicator

Classification performance was evaluated by precision, recall and F1-score. Three indicators are used to measure the accuracy of the classification method proposed in this article. The accuracy of calculation for each category is shown in (Table 6). From the perspective of accuracy, recall rate and F1-score, most of them are above 0.8, indicating that most of the disaster themes can be correctly identified, indicating that the method used in this article has a comparative advantage. Good classification effect, but the poor performance of the weather warning category's recall rate is that a large number of weather warning information is misidentified. This may be due to two reasons. On the one hand, the number of Weibo related to weather warning is small, on the one hand, the content of Weibo related to weather warnings is relatively complicated, and may contain various information, resulting in misidentification from the text.

Table 6 The text disaster information accuracy assessment results

Category	Pre	Rel	F1_score
Weather warning	0.99	0.07	0.13
Traffic condition	0.97	0.89	0.90
Rescue information	0.95	0.91	0.93
Disaster information	0.88	0.87	0.85
The cause of the disaster	0.98	0.86	0.83
Casualties and losses	0.89	0.89	0.89
Overall	0.93	0.78	0.85

In order to prove the effectiveness of the model proposed in this article, the classification model of this article is compared with the classic text classification model SVM and CNN models. At the same time, (Xue and Hong et al., 2020)proposed a topic-based event classification model KGE-MMSLDA in 2019. The proposed model is compared with previous models. In contrast, its classification performance is measured by three indicators: precision, recall, and F1-score, and the accuracy is shown in (Table 7). From (Table 7), we can see that the accuracy of the classification model used in this article is significantly higher than other models, indicating that the performance of the classification model proposed in this article is more stable and the classification result is more accurate. The recall value of the model in this paper is higher than the LR and SVM models, and lower than the CNN model, but the F1 value of the model in this paper is higher than other models, indicating that the overall performance of the model proposed in this paper is better than other models.

Table 7 Compared with traditional classification models

	Pre	Rel	F1_score
SVM (text only)	0.72	0.70	0.70
CNN (text only)	0.81	0.80	0.80
Multi-modal	0.81	0.80	0.81
Overall	0.93	0.78	0.85

5 Conclusion

In recent years, the acquisition and analysis of disaster information has been a key issue for government and scientific research institutions. Social media data can enable officials and victims to be the truth and disseminators of natural disaster information at the same time, and social media data has the advantages of real-time and low latency, so ,the social media has become an important source of

natural disaster information. With the development of technology, scholars ,the methods of studying natural disaster information are also constantly being optimized.

This article uses a classification model based on LDA and a multi-classification model based on Bert and Vgg-16, which is suitable for short-term social media and other types of disaster events that have caused a large-scale sensation.

In this article, first use the web crawler combined the Weibo API to obtain text and graphic data for subsequent processing and classification. Use the LDA model to classify topics and identify topics related to emergencies. Based on the data characteristics of text and images, this paper constructs a network framework suitable for microblog text and image disaster extraction. After optimization operations such as control over-fitting and grid parameter optimization, the accuracy of the model on the test set has been improved, and the classification accuracy has reached more than 80%. The results of verification on the newly acquired Henan torrential rain dataset in 2021 further show that the application of the model to disaster information classification has a certain degree of accuracy. Finally, through the visualization and statistical analysis of the data, it is found that the disaster information is consistent with the actual disaster development stage, which shows that the method proposed in the article is effective in monitoring Henan rainstorm disaster events and can effectively help the official disaster decision-making.

Funding This paper is provided by National Social Science Foundation of China (Grant No.20AZZ006); Innovation and Entrepreneurship Foundation of Guizhou(Grant No S202110671039); Foundation of Guizhou University of Finance and Economics (Grant No .2020XQN04).

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

References:

"Social media data and housing recovery following extreme natural hazards.".

"Social media data mining and knowledge discovery under wireless network.".

Abid, F. and C. Li, et al. (2020). "Multi-source social media data sentiment analysis using bidirectional recurrent convolutional neural networks." Computer Communications **157**: 102-115.

Aznar-Crespo, P. A. A. M.(2021). "Adapting Social Impact Assessment to Flood Risk Management.".

Basalamah, A. and S. Rahman(2022). "An Optimized CNN Model Architecture for Detecting Coronavirus (COVID-19) with X-Ray Images." Computer Systems Science and Engineering **40**(1): 375-388.

Behl, S. and A. Rao, et al. (2021). "Twitter for disaster relief through sentiment analysis for COVID-19 and natural hazard crises." International Journal of Disaster Risk Reduction **55**: 102101.

Chen, C. C. and H. Wang(2020). "Using community information for natural disaster alerts." Journal of Information Science: 016555152097987.

Ghafarian, S. H. and H. S. Yazdi(2020). "Identifying crisis-related informative tweets using learning on distributions." Information Processing & Management **57**(2): 102145.

Ghosh, S. and P. K. Srijith, et al. (2017). "Using social media for classifying actionable insights in disaster scenario." International Journal of Advances in Engineering Sciences and Applied Mathematics **9**(4): 224-237.

Gupta, A. and R. Katarya(2021). "PAN-LDA: A latent Dirichlet allocation based novel feature extraction model for COVID-19 data using machine learning." Computers in Biology and Medicine **138**: 104920.

Hao, H. and Y. Wang(2020). "Leveraging multimodal social media data for rapid disaster damage assessment." International Journal of Disaster Risk Reduction **51**: 101760.

Hung, P. T. and K. Yamanishi(2021). "Word2vec Skip-Gram Dimensionality Selection via Sequential Normalized Maximum Likelihood." Entropy **23**(8): 997.

Kang, A. and L. Ren, et al. (2021). "Stakeholders' views towards plastic restriction policy in China: Based on text mining of media text." Waste Management **136**: 36-46.

Kaplan, A. M. and M. Haenlein(2010). "Users of the world, unite! The challenges and opportunities of Social Media." Business Horizons **53**(1): 59-68.

Kitazawa, K. H. S. A.(2021). "Social media and early warning systems for natural disasters: A case study of Typhoon Etau in Japan.".

Kumar, A. S. J. P.(2020). "A deep multi - modal neural network for informative Twitter content classification during emergencies.".

Mehmood, A. and M. Attique Khan, et al. (2022). "Human Gait Recognition: A Deep Learning and Best Feature Selection Framework." Computers, Materials & Continua **70**(1): 343-360.

Nguyen, D. T. O. F.(2017). "Damage Assessment from Social Media Imagery Data During Disasters.".

Park, I. and B. Yoon, et al. (2021). "Technological Opportunities Discovery for Safety Through Topic Modeling and Opinion Mining in the Fourth Industrial Revolution: The Case of Artificial Intelligence." IEEE Transactions on Engineering Management **68**(5): 1504-1519.

Piatyszek, E. K. G. M.(2012). "A model-based approach for a systematic risk analysis of local flood emergency operation plans: a first step toward a decision support system.".

Qiao, B. Z. Z. H.(2021). "A joint model for entity and relation extraction based on BERT.".

Ragini, J. R. A. P.(2018). "Big data analytics for disaster response and recovery through sentiment

analysis."

Rasiwasia, N. C. P. J.(2010). "A New Approach to Cross-Modal Multimedia Retrieval.".

Sadiq, A. M. A. H.(2020). "Human Sentiment and Activity Recognition in Disaster Situations Using Social Media Images Based on Deep Learning.".

Wang, R. and Y. Hu, et al. (2020). "Tracking Flooding Phase Transitions and Establishing a Passive Hotline With AI-Enabled Social Media Data." IEEE Access **8**: 103395-103404.

Wood, E. and M. Sanders, et al. (2021). "The practical use of social vulnerability indicators in disaster management." International Journal of Disaster Risk Reduction **63**: 102464.

Xu, N. M. W.(2017). "A Residual Merged Neutral Network for Multimodal Sentiment Analysis.".

Xu, Z.(2020). "How emergency managers engage Twitter users during disasters." Online Information Review **44**(4): 933-950.

Xue, F. and R. Hong, et al. (2020). "Knowledge-Based Topic Model for Multi-Modal Social Event Analysis." IEEE Transactions on Multimedia **22**(8): 2098-2110.

Yu, M. H. Q. Q.(2019). "Deep learning for real-time social media text classification for situation awareness – using Hurricanes Sandy, Harvey, and Irma as case studies.".

Yu, Y. L. H. M.(2016). "Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks.".