Mining typhoon victim information based on multi-source data fusion using social media data in China: a case study of the 2019 Super Typhoon Lekima

Kejie Wu, Jidong Wu and Yue Li

1. Introduction

Tropical cyclones are often costly and inflict devastating casualties (Peduzzi et al. 2012; Mori and Takemi 2016). In China, the annual direct economic losses caused by tropical cyclones have been estimated at approximately 28.7 billion CNY (Chinese Yuan, 2006 price levels), accounting for 0.38% of China’s annual GDP (gross...
domestic product) from 1983 to 2006 (Zhang et al. 2009; Ye et al. 2019). Effective emergency management in the response phase can reduce disaster losses caused by tropical cyclones (Alexander 2002; Pielke 2007).

In recent decades, social media has been widely applied in disaster management (Xie and Yang 2018). A social media user functions as a disaster event sensor and conducts sensor readings (Sakaki et al. 2010). Social media have proven to be a new way to monitor natural disasters (Crooks et al. 2013) and help to organize emergency response (Sakaki et al. 2010; Crooks et al. 2013). Based on situational awareness and information sharing (Sakaki et al. 2013), millions of microblogs are posted during disasters that can be used to identify victims and develop disaster relief measures (Watson et al. 2017). In Chinese disaster management, the acquisition of disaster indicators mainly relies on a statistical reporting system. Compared with statistical data, social media data is timelier to obtain, and its effective use will greatly improve the efficiency of emergency management (Zhang 2015).

Novel and additional disaster information is presented by crowdsourcing through social media (Simon et al. 2015), especially during tropical cyclone disaster management. Li et al. (2021) concluded that using social media in cyclone disasters serves three major purposes: emotional expression, situational updates, and disaster-related information exchange. Many studies on social media converge on themes of sentiment analysis and emergency response. For example, Wu and Cui (2018) developed a data processing framework and analyzed Twitter (www.twitter.com) postings using multidimensional analysis containing reverse geocoding, emotion analysis, subject tags, and high-frequency methods to track public responses to Superstorm Sandy. Spruce et al. (2020) applied detailed sentiment analysis on the public’s negative feelings about the storm to measure the impact of storms. Additionally, with sentiment analysis on calculation of the emotional intensity of an event, Ragini et al. (2018) used a data-driven approach for disaster response, while Yuan and Liu (2020) assessed the damage caused by Hurricane Matthew. Sentiment analysis is a common method of natural language processing, which is based on text word analysis to determine the specific sentiment and other markers contained in it (Feldman 2013). The sentiment changes revealed in social media are thought to reflect temporal and spatial mood changes in society (Wu and Cui 2018). The above studies showed that sentiment analysis is a common step when using social media for the disaster management of cyclones.

In addition to sentiment, many studies focused on economic damage and victims post disaster. Kryvasheyeu et al. (2016) counted the spatial and temporal distribution of Twitter and confirmed the correlation between Twitter activity and economic damage. Other investigations used geospatial event detection to identify detailed disaster information such as victims and rescues (Sakaki et al. 2010). Although there has been a significant amount of research on extracting disaster information, we still cannot be blind to the limitations. First, few studies extracted the distribution of victims directly through social media (Sakaki et al. 2010). Second, social media data often contained spatial heterogeneity, and geographic location accuracy was insufficient (Fang et al. 2019). Third, some studies lacked of detailed information on victims or high-precision data on damage to verify the validity of the extracted disaster information (Wu et al. 2021).
With the widespread use of Twitter in disaster management abroad, Sina Weibo (weibo.com), as a Chinese version of Twitter, has attracted more attention in disaster management (Deng et al. 2016). Although sentiment analysis was still used in disaster management in China (Ma et al. 2014), due to the lack of geographic location on Sina Weibo, many studies could not extract information with high spatial resolution. The most accurate information existed only at the district level in Wuhan (Fang et al. 2019).

Super Typhoon Lekima was the ninth tropical cyclone named during the 2019 Pacific typhoon season. At approximately 5:00 a.m. on August 7, it was defined as a typhoon by the China Meteorological Administration (CMA). At approximately 11:00 p.m. the same day, it was upgraded to a super typhoon by CMA. It first landed in Chengnan Town, Zhejiang Province, at approximately 1:45 a.m. on August 10 with a 52 m·s⁻¹ maximum wind speed near the center (Bao et al. 2020). Then, it affected Zhejiang, Shanghai, Jiangsu, and Anhui and made a second landfall in Shandong. At 2:00 p.m. on August 13, it ceased to be considered a cyclone according to the CMA. Lekima killed 57 people, 14 people went missing, 14.024 million people were affected, and 2.097 million people were urgently relocated as of 10:00 a.m. on August 14, 2019 (Wu et al. 2021).

The main purpose of this study is to find suitable methods based on multi-source data fusion to mine accurate typhoon victim locations using microblogs from Sina Weibo. This paper takes the 2019 Super Typhoon Lekima as a case study and obtains the possible victim locations firstly through the distribution of negative sentiment and its spatial clustering results. Then, the extraction results are further optimized by fusing population vulnerability using multi-source data.

2. Data and method

2.1. Data

Four types of data were used in this study. First, the typhoon data contained typhoon path, wind speeds and precipitation data. The path and wind speeds were from the Tropical Cyclone Information Centre (http://tcdata.typhoon.org.cn) (Ying et al. 2014; Lu et al. 2021), while the precipitation data were from China Meteorological Data Network (http://data.cma.cn). The data were recorded from August 7, 2019, to August 13, 2019. The typhoon path was recorded every six hours, while the wind speed was given as a two-minute mean maximum near center (m·s⁻¹) in which the highest wind speed was 62 m·s⁻¹. The precipitation data were collected from 2467 meteorological stations, containing latitude, longitude and 1 h precipitation value.

Second, real victim data were obtained from the Ministry of Emergency Management of China, including administrative villages of victims. We defined the counties with victims as affected counties. These were mainly distributed in Zhejiang and Shandong provinces. Among all affected counties, Pan’an County in Zhejiang Province suffered the most victims—up to 30 persons.

Third, microblog data were collected from Sina Weibo during the period from the beginning of typhoon numbering (August 7, 2019) to the cancellation of numbering (August 13, 2019). Microblog data are the information that the public posts on social
media platforms to express their thoughts. Due to its characteristics of timely situational awareness and information sharing, it can be quickly obtained and reflect disaster information, which is an important data source for disaster information extraction (Deng et al. 2016; Fang et al. 2019). These microblog data were extracted from the commercial data interface of Sina Weibo (open.weibo.com/development/businessdata), and all microblogs were searched with the keyword ‘台风’ (‘typhoon’). We collected a total of 416,243 microblogs. Each microblog included information on user ID, content, and publication time, among which 78,300 of the microblogs contained exact posted latitude and longitude information.

Finally, spatial distribution 1 km grid of China’s DEM, spatial distribution 1 km grid of China’s population in 2015 (Xu 2017b), spatial distribution 1 km grid of China’s GDP in 2015 (Xu 2017a) and spatial distribution 1 km grid of China’s NDVI in August, 2019 were obtained from Resource and Environment Science and Data Center (https://www.resdc.cn/).

2.2. Method

Social media is no longer just a communication tool (Kankanamge et al. 2020). In this research, the strategy shown in Figure 1 is used to mine accurate victim locations and spatial distributions using microblogs. This strategy contains three processes: data collection, data preprocessing, and victim information identification. First, a Sina

![Figure 1. Technical flowchart of this study.](image-url)
commercial data interface was used to collect microblogs, and LDA was applied to filter microblogs. Microblogs with posted locations were used to identify victim information. Second, we conducted sentiment analysis on microblogs containing geographic locations from time and space scales to find the distribution of negative sentiment. Negative sentiment will reflect the location of potential victims, which can provide a reference for subsequent extraction. Finally, we selected precipitation, GDP per capita, slope and NDVI to calculate population vulnerability, and combined population vulnerability to identify victim locations using Ordering Point to Identify the Cluster Structure (OPTICS) (Ankerst et al. 1999) algorithm. To evaluate accuracy, combining real victim information, we defined a point in a county containing victims as a hit and the ratio of these points to all points as the hit rate. The details of the method are presented below.

2.2.1. Classify microblogs
To classify microblogs related to typhoons, the LDA algorithm was adopted to classify the 416,243 collected microblogs by topic. The LDA algorithm is a commonly used algorithm for topic discovery and has been widely applied in natural language processing (Gross and Murthy 2014). Disaster-related microblogs about public opinion could be extracted by topic classification. First, we removed the URL, punctuation, and stop words in microblogs. Second, we generated the dictionary and corpus for each word. The dictionary contained sequence numbers and words, while the corpus contained sequence numbers and word frequencies. Finally, after the topic number was specified in advance, all parameters were put into the LDA model to obtain the result of topic classification. However, there was no way to know in advance of how many topics these microblogs would be divided. Thus, we used perplexity (Blei et al. 2003) and coherence (Mimno et al. 2011) to evaluate the topic number, and the formula is shown below:

$$\text{perplexity} = e^{\frac{\sum \log p(w)}{N}}$$  \hspace{1cm} (1)

$$\text{coherence} = \sum_{n=2}^{N} \sum_{m=1}^{n-m} \log \frac{D_2(w_m, w_n) + 1}{D_1(w_m)}$$  \hspace{1cm} (2)

where $p(w)$ is the probability of each word appearing in microblogs, $N$ is the total microblog count, $D_1(w_m)$ is the document frequency of word $w_m$, and $D_2(w_m, w_n)$ is the co-occurrence document frequency of words $w_m$ and $w_n$.

Perplexity is a measure of how uncertain we are about microblogs’ relevance to a topic. The more topics, the lower the perplexity of the model (Blei et al. 2003). However, when there are many topic numbers, the generated models tend to overfit, so it is not possible to judge topic numbers solely by relying on perplexity. In this case, coherence should be combined to judge final topic numbers. A higher coherence means better interpretation (Mimno et al. 2011). Finally, considering perplexity and coherence, we can find suitable topic numbers and several microblog topics and classify disaster-related topics.
2.2.2. Sentiment analysis method

Sentiment analysis was used to extract the distribution of negative sentiment so as to find potential affected people related to Lekima. Public sentiment is a typical representation of public opinion information. Positive and negative sentiments can reflect the things about which people were most concerned during a typhoon. We used the Chinese sentiment analysis package SnowNLP (Yu et al. 2021; Zhang et al. 2021) in Python to score the sentiment of each microblog. This package contained a sentimental training set derived from user reviews, using naïve Bayes methods to train and predict data. The sentiment score of each microblog was normalized to between 0 and 1. A value closer to 0 meant that the microblog was more negative, while a value closer to 1 meant that the microblog was more positive. Negative sentiment often contains disaster information such as injury and death (Lu et al. 2015). We defined sentiment scores lower than 0.25 (0–1 quartile) as negative sentiment, and scores lower than $10^{-11}$ as extremely negative sentiment. Finally, we measured the change in negative sentiment numbers over time and visualized them on a map to master the general spatial distribution of victims from a macroscopic scale for the subsequent extraction of victim locations for reference.

2.2.3. Calculating population vulnerability

Vulnerability is the susceptibility of assets to the impact of disasters determined by socioeconomic (Cutter and Finch 2008) and physical environment factors (UNISDR (United Nations International Strategy for Disaster Reduction) 2017). Assigning weight to different vulnerability indicators was considered to be possible to construct vulnerability, and this method will also be adopted in this paper (Wu et al. 2019). First, we calculated the accumulated precipitation at each weather station from August 7, 2019 to August 13, 2019, and used ordinary kriging to interpolate it (Jasim and Mustafa 2018) into a grid of 1 km. GDP grid was divided by population grid to get the GDP per capita grid. Slope grid was calculated based on Shuttle Radar Topography Mission (SRTM) in ArcGIS 10.2. Second, precipitation ($P$), GDP per capita ($INCOME$), slope ($SLPOE$) and Normalized Difference Vegetation Index ($NDVI$) were normalized to 0–1 respectively. Finally, after unifying the resolution of the above data to 1 km and normalization, we selected 9 provinces and cities affected by Lekima for calculating the vulnerability index on this basis, including Fujian, Zhejiang, Shanghai, Anhui, Jiangsu, Shandong, Hebei, Liaoning and Jilin, and the population vulnerability index ($Vulnerability$) was calculated according to the following formula:

$$Vulnerability = \frac{1}{4}SLOPE + \frac{1}{4}P + \frac{1}{4}(1 - INCOME) + \frac{1}{4}(1 - NDVI)$$

(3)

Considering that the impact of $INCOME$ and $NDVI$ on the typhoon population vulnerability is inverse, it is necessary to use 1 to reverse the calculation.

2.2.4. Identifying victim locations

Since not all microblogs were related to victims, we pre-define the corpus (Table 1) related to victims to extract the related microblogs. If a microblog contained words
from the corpus, this microblog was considered to be related to victims and can be used to identify victim locations. Based on the microblogs related to victims, spatial clustering algorithm (Lucas 2012; Karo and Huda 2016) was selected to identify victim locations. Before that, the problem of collinearity between microblogs should be solved because similar microblogs may affect spatial clustering. We calculated the cosine distance between each pair of microblogs. When the distance value was greater than 0.7, the two microblogs were defined to be similar, and one of them was deleted at that time.

Spatial clustering groups objects based on spatial proximity (Kisilevich et al. 2010). Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a famous implementation of the density-based spatial clustering algorithm (Ester et al. 1996). However, DBSCAN has a heavy reliance on a predefined threshold: radius and threshold density. It is difficult to select a suitable value for these thresholds a priori (Borah and Bhattacharyya 2004). Thus, an improved algorithm—OPTICS (Ankerst et al. 1999)—was proposed to reduce the sensitivity of a predefined threshold (Bakillah et al. 2015). OPTICS has the same input parameters as DBSCAN but is insensitive to parameters (Ankerst et al. 1999). This algorithm does not explicitly generate data clustering but only sorts the objects in the dataset to obtain an ordered object list. In the OPTICS algorithm, the Euclidean distance is usually used to measure the spatial proximity between two points to calculate the reachability distance. However, in order to explore whether integrating the population vulnerability into the OPTICS algorithm can improve the extraction accuracy, we selected the Euclidean distance as distance1 and the Euclidean distance combining vulnerability as distance2 respectively for spatial clustering. The calculation formulas of distance1 and distance2 are as follows:

\[
\text{distance 1} = \sqrt{(\text{Lng}_1 - \text{Lng}_2)^2 + (\text{Lat}_1 - \text{Lat}_2)^2}
\]  \quad (4)

\[
\text{distance 2} = \left((1 - \text{Vulnerability}_1) + (1 - \text{Vulnerability}_2)\right) \\
\times \sqrt{(\text{Lf}_1 - \text{Lf}_2)^2 + (\text{Lt}_1 - \text{Lt}_2)^2}
\]  \quad (5)

where Lng and Lat are the posted location of one microblog, Vulnerability1 and Vulnerability2 are the value corresponding to population vulnerability grid at points
obtained by Sec. 2.2.3. Since the places with higher vulnerability are more likely to have victims, the places with higher vulnerability are more likely to be clustered, and the calculated distance will be smaller.

After using two distances, we obtained different spatial clustering results of victim locations to compare the effects of data fusion. We call a victim location point in an affected county as hit point, and other victim location points as missed points. Finally, we use the hit rate mentioned above to compare the accuracy with real victim locations.

3. Results

3.1. Topic classification of disaster-related microblogs

All 416,243 selected microblogs were classified into four topics by the LDA algorithm (Figure 2a). These four topics represented the public mood and expectation, disaster
Table 2. Top 10 frequency words in each of 7 topics in Figure 2, and artificially defined themes of topics.

| Topic id | Top 10 frequency words in each topic | Theme                          |
|----------|-------------------------------------|-------------------------------|
| V1–1     | 台风(typhoon), 天(day), 前(before), 说(say), 真的[truly], 想(want to), 走(go), 来临[advent], 过后(after), 天空(sky) | Public mood and expectation   |
| V1–2     | 台风(typhoon), 日(day), 月(month), 利奇马[Leckima], 影响[influence], 工作[work], 救援[rescue], 救灾[disaster relief], 群众[public], 消防[firefighting] | Disaster relief and rescue     |
| V1–3     | 台风(typhoon), 登陆[landfall], 日(day), 级(level), 利奇马[Leckima], 中心[center], 影响[influence], 沿海[coastal], 福建[Fujian], 预计[excepted] | Disaster warning and forecasting|
| V1–4     | 台风(typhoon), 蜕变[metamorphosis], 之战[the battle], 灾害[disaster], 周震南[Zhou Zhennan], 超话[super talk], 丁程鑫[Ding Chengxin], 捐款[Donate], 爱[love], 舞台[stage] | Unrelated-information         |
| V2–1     | 台风(typhoon), 天(day), 过后(after), 说(say), 真的[truly], 走(go), 想(want to), 吃(eat), 太(too), 天气(weather) | Public mood and expectation    |
| V2–2     | 台风(typhoon), 利奇马[Leckima], 影响[influence], 日(day), 月(month), 山东[Shandong], 登陆[landfall], 暴雨[heavy rain], 工作[work], 放布[release] | Disaster warning and forecasting|
| V2–3     | 台风(typhoon), 利奇马[Leckima], 灾害[disaster], 爱灾(hit), 捐款[Donate], 成功[succeed], 救援[rescue], 公益[public welfare], 救灾[disaster relief], 项目[project] | Disaster relief and rescue     |

warning and forecast, disaster relief and rescue information, and other irrelevant information (Table 2). For the entire sample (Figure 2a), when the topic number exceeded 8, the perplexity decreased significantly, indicating that the model had been overfitted. When topic number was 4, the coherence was the highest. Thus, 4 was the best topic number, which is shown in Figure 2(a). As seen from Table 2, Topic V1–1 described the public mood and expectation, containing the most microblogs. In Topic V1–2, a large number of words such as ‘work’ and ‘firefighting’ were related to disaster relief and rescue, while in Topic V1–3, words such as ‘influence’ and ‘forecast’ were all related to typhoon early warning and forecasting. The intersection of Topic V1–2 and Topic V1–3 circles indicated that some microblogs belong to both topics, and the distinction between both topics was not high. During Leckima, disaster response was usually related to disaster early warning and disaster relief and rescue information often appeared simultaneously with early warning and forecast information. Microblog counts in Topic V1–4 were defined as unrelated information to typhoon disasters because they were mainly related to entertainment information, article information and so on.

After deleting microblogs contained in Topic V1–4, the remaining microblogs were considered disaster-related microblogs. These were further classified into three topics: public sentiment and expectation, early warning and forecasting, and disaster relief and rescue (Figure 2b). From the perspective of perplexity, when the topic number exceeded 7, the model was overfitted. Coherence was highest when the topic number was 3, so the optimal topic number was selected as 3. The microblog counts about public sentiment and expectations were also the most numerous (Table 2, Topic V2–1). Although topics V2–2 and V2–3 still referred to early warning and forecasting, and disaster relief and rescue, there was no crossover in Figure 2(b), unlike Figure 2(a).

Overall, after topic classification, we retained 406,634 microblogs as disaster-related microblogs, including 78,068 microblogs with posted latitude and longitude. This provided a foundation for further extracting victim information.
### 3.2. Distribution of negative sentiment

To a certain extent, negative sentiment reflected by microblogs can express the distribution of affected people, including injury and victim distribution (Lu et al. 2015). The most negative sentiment was generated on the day of typhoon landfall, and most of the negative sentiment was distributed in Zhejiang and Shandong provinces (Figure 3). As the typhoon approached, negative sentiment began to increase. It increased significantly on the day before landing (August 9) and reached a maximum on the day of landing (August 10). As the intensity of the typhoon weakened, the negative sentiment number gradually decreased. In Zhejiang Province, the negative sentiment was mainly distributed in the coastal areas where typhoons landed, while for Shandong Province, negative sentiment was mainly distributed in the coastal and central and northern regions.

Extremely negative sentiment implies injury and victim information. We extracted 13 extremely negative microblogs, which conveyed three types of information: (1) water leakage and power failure in the disaster area, (2) traffic disruption caused by bad weather, and (3) notification of injury and victim number during Lekima. However, this victim number was released by statistical agencies and official media with a time lag.

Overall, victim information was contained in negative sentiment distributed in a large space, which could reflect the spatial distribution of victim information from a macro perspective. The results of sentiment analysis could provide a reference for subsequent identification of victim locations.

### 3.3. Distribution of victim locations

#### 3.3.1. Multi-source data

Based on Table 1, 2255 microblogs related to victims from 78,068 total microblogs were selected, accounting for 2.89% of the total microblog counts, and were
spatialized as shown in Figure 4(a). It can be seen from the spatial distribution that the points related to victims were distributed all over the country, mainly in Zhejiang province, Shanghai and Shandong Province. This three provinces and cities were also the main areas affected by the typhoon. The distribution of microblogs suggested that people in the area affected by typhoon will post more microblogs about the victims. The vulnerability of victims in this nine provinces and cities affected by Lekima varied greatly according to Figure 4(b). The vulnerability of coastal Zhejiang, Shanghai and central Shandong was the highest, followed by Hebei and Liaoning, and other provinces’ vulnerability was low. The distribution of victim vulnerability suggested that people in Zhejiang, Shandong and Shanghai were more vulnerable to Lekima (Figure 4b), which may produce more victims. In addition, the distribution of vulnerability was consistent with that of the microblogs related to victims in Figure 4(a), and both will become important basis for identifying the location distribution of victims.
3.3.2. Comparison of identification results

Two identification results had similarity (Figure 5a). 72 hours after typhoon landed, the counties with a severe victim toll were distributed in Zhejiang Province, in which the maximum number was up to 30, and 6 counties in Shandong Province had several victims. From two clustering results, the main clustering points were in Yueqing, Shanghai and the central and northern parts of Shandong. Compared with the sentiment distribution in Figure 3, the identified victim locations were basically distributed in areas with relatively dense negative sentiment. Both missed points can be summarized into three situations: First, roughly distributed around the counties with victims, such as the missed points near Yueqing and central and northern Shandong. Although these points did not coincide with the real affected counties, their geographical locations were close to each other, so they can be used as reference points to explore victim locations. Second, Points that had no geographical connection with the real affected counties, such as the missed points in Shanghai. These points had nothing to do with the real affected counties, and this situation interfered with the results. Third, Points that were geographically related to the real affected counties, such as the missed points in Anji and Hangzhou. Although seven people died in Anji, there were not points. However, there were many points in Hangzhou, a nearby city. This suggested that if the county which had victims had a low level of social activity, then the response to the victims in that county might be geographically linked to a higher level of social activity areas in the immediate vicinity.

Figure 4. The extraction of microblogs related to victims and the distribution map of vulnerability of Lekima. (a) Microblogs related to victims extracted by Table 1 as of 0 o’clock on August 14, 2019; (b) Typhoon population vulnerability distribution map based on slope, precipitation, per capita GDP and NVDI.
Although there existed similarity between two results, whether population vulnerability was considered or not will affect the results of identification. First, after considering population vulnerability, the hit rate was 31.8%, which had an 8.3% increase over Figure 5(a). Second, for hit points, compared with Figure 5(a), after considering population vulnerability, more hit points appeared in Shandong. Two counties with no previous hit points also appeared, and more hit points appeared in counties with original hit points. Third, we found a decrease in the number of missed points in Hangzhou, Shanghai and Qingdao et al. Overall, after considering population vulnerability, the identification accuracy of victim locations was improved both from hit rate and the distribution of hit points in affected counties, indicating that considering multi-source data fusion can improve the accuracy of disaster information extraction.

4. Discussion

In this study, we mined victim information related to Lekima using microblogs. We found that the LDA algorithm can classify microblogs and distinguish typhoon-related microblogs effectively. Then, we found extremely negative sentiments containing injury and victim information. Finally, we identified victim locations 72 hours after landing. Compared with the real victims at county level, the hit rates reached 23.5% and 31.8%. Many studies only considered social media as a single data source to extract disaster information. However, this study combined precipitation, slope,
GDP per capita and NDVI to build population vulnerability, and then integrated social media data to improve the accuracy of disaster information extraction.

4.1. Advantage of using microblogs for victim location extraction in emergency management

During a disaster, the safety of affected people is always the primary concern of decision-makers. Timely discovery of the locations and causes of victims and dispatch of rescue forces can prevent more casualties from occurring. For decision-makers, faster access to the number and distribution of victims can provide a more effective reference for determining the extent of the disaster and carrying out disaster relief efforts. In China, disaster information such as victim number is mainly obtained through statistical reporting, which has problems such as time lags. In some cases, local geological hazards may prevent a team from traveling to a disaster area to assess casualties. The geographic distribution of victims through social media gives a rough picture of how victims are distributed across counties. We need to provide references for disaster emergency response in time, and then allocate rescue forces reasonably.

Another advantage of using social media data to extract victim location is that emergency management information can be obtained at the same time. First, secondary disasters such as landslides, mudslides and storm surges caused by typhoon are closely related to the victims. Among the microblogs related to victims identified in Lekima, many of them contain landslide and debris flow information, which can provide a reference for locating the secondary disasters and planning reasonable traffic diversion routes. Second, although the rescue information was not extracted in this study, we still found that there is a lot of rescue demand information in microblogs related to the victims. This information is put forward by the affected people from their own perspective, such as ‘there are victims and injured people here, and stretchers and food are urgently needed’. Rescue information like this is very important for decision makers. The information related to victims provided by social media is helpful for decision makers to understand the user behaviours of the disaster victims (Ghurye et al. 2016), and then formulate reasonable disaster relief actions. Overall, secondary disaster information and rescue information accompanied in the victim information could be further detected.

4.2. Impacts of victim-related keywords selection on information extraction accuracy

The choice of victim-related keywords is very important for the final information extraction accuracy as shown in Table 3. First, we selected 19 victim-related keywords for typhoon-induced victim information extraction, the hit rate could reach 31.8% as demonstrated above. Second, different victim-related keyword combination will influence hit rates and selected microblog counts. Some keywords had little influence on hit rates, such as ‘出事 (accident)’ and ‘掩埋 (buried)’, the hit rate still reached 31.6% even without considering both. However, some keywords had a great influence on hit rates, such as ‘伤亡 (death & injured)’, ‘死亡 (death)’ and ‘遇难 (dead)’, the hit rate
could reduce to 19.5% if the three keywords were not selected. If using the three keywords as a combination, the hit rate reached 27.3%, still lower than that of 31.8%. Overall, if more suitable victim-related keywords were added, hit rates might be improved. In addition, if more microblogs with specific latitude and longitude could be obtained, it is also possible to improve the accuracy.

4.3. Limitations and proposals for improvement

Limitations in this study need to be addressed. First, we only used data containing latitude and longitude, and the data utilization rate was only 19.2%. Second, we cannot guarantee that the description of victim information in microblogs with precise locations is the real location of users, just as the recognition result of Anji County in Figure 5 is shifted. This spatial heterogeneity leads to deviations between the extracted results and real disaster information (Xiao et al. 2015). Analyzing the location names of Weibo content or using crowdsourcing methods to obtain the real location may solve this problem (Jaiswal et al. 2018; Wald et al. 2020). Finally, the 31.8% hit rate is still limited, and more multisource data could further improve the accuracy. Although we have used slope, this still reminds us that if we can utilize more data like land use and house structure data to identify high-risk areas of mountain torrents and mudslides, then they may be used to modify our results. Moreover, remote sensing data and mobile phone data could provide effective supporting information for determining potential casualty areas. Multisource data fusion could overcome the shortcomings of the data and obtain more accurate results. However, this technology is not sufficiently mature.

5. Conclusion

Using multi-source data fusion, this study provided an effective method to mine victim information from social media data using the case study of the 2019 Super Typhoon Lekima in China. The negative sentiment extracted by sentiment analysis from microblogs can roughly show the spatial distribution of injury and victim. Furthermore, different from other studies, by using OPTICS method that considered vulnerability, victim locations were mined and fusing vulnerability improved extraction accuracy. The real victim information extraction accuracy at the county level reached 31.8% at 72 hours after the typhoon landed.

Overall, by combining vulnerability, this case study proved that microblogs can reflect negative sentiment and can be used to identify victim information. Identification of victim information is not only convenient for decision-makers in collecting disaster information but is also conducive to the efficient implementation of emergency management. Further improvement is needed by using more multi-source data fusion technology to improve the extraction accuracy and mine more disaster information. This information may include traffic paralysis, urban waterlogging or house collapses.
Funding
This research was funded by National Key Research and Development Program of China, grant number 2018YFC1508903, and the National Natural Science Foundation of China, grant numbers 42077437.

ORCID
Jidong Wu http://orcid.org/0000-0001-8208-8373

Data availability statement

| Data type                      | Availability details                                                                                                                                                                                                 |
|-------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Typhoon path and wind speed data | The data that support the findings of this study are openly available in http://tcdata.typhoon.org.cn, Ying et al. 2014; Lu et al. 2021 All                                                                                  |
| Precipitation data            | The data that support the findings of this study are openly available in http://data.cma.cn                                                                                                                                 |
| Victim data                   | The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to their containing information that contains sensitive information Basic, Share upon request                                               |
| Microblog data                | The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to their containing information that could compromise the privacy of research participants Basic, Share upon request                                               |
| DEM, population, GDP and NDVI | The data that support the findings of this study are openly available in https://www.resdc.cn/, Xu 2017a, Xu 2017b All                                                                                                       |

Disclosure statement
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
Alexander DE. 2002. Principles of emergency planning and management. Oxford University Press on Demand.
Ankerst M, Breunig MM, Kriegel H-P, Sander J. 1999. OPTICS: ordering points to identify the clustering structure. SIGMOD Rec. 28(2):49–60. http://dx.doi.org/10.1145/304181.304187.
Bakillah M, Li RY, Liang SHL. 2015. Geo-located community detection in Twitter with enhanced fast-greedy optimization of modularity: the case study of typhoon Haiyan. Int J Geograph Inform Sci. 29(2):258–279. http://dx.doi.org/10.1080/13658816.2014.964247.
Bao X, Wu L, Zhang S, Yuan H, Wang H. 2020. A comparison of convective raindrop size distributions in the eyewall and spiral rainbands of typhoon Lekima (2019). Geophys. Res. Lett. 47(23):e2020GL090729.
Blei DM, Ng AY, Jordan MI. 2003. Latent dirichlet allocation. J Mach Learn Res. 3:993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf?TB_iframe=true&width=370.8&height=658.8.
Borah B, Bhattacharyya DK. 2004. An improved sampling-based DBSCAN for large spatial databases. 2004 IEEE International conference on intelligent sensing and information processing. IEEE. 1: 92–96. http://dx.doi.org/10.1109/ICISIP.2004.1287631.
Crooks A, et al. 2013. #Earthquake: twitter as a distributed sensor system. Trans GIS. 17(1): 124–147.
Cutter SL, Finch C. 2008. Temporal and spatial changes in social vulnerability to natural hazards. Proc Natl Acad Sci U S A. 105(7):2301–2306.
Deng Q, Liu Y, Zhang H, Deng X, Ma Y. 2016. A new crowdsourcing model to assess disaster using microblog data in typhoon Haiyan. Nat Hazards. 84(2):1241–1256.
Ester M, Kriegel HP, Sander J, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD. 96(34):226–231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf?source=post_page.
Fang J, Hu J, Shi X, Zhao L. 2019. Assessing disaster impacts and response using social media data in China: a case study of 2016 Wuhan rainstorm. Int J Disaster Risk Reduct. 34: 275–282. http://dx.doi.org/10.1016/j.ijdrr.2018.11.027.
Feldman R. 2013. Techniques and applications for sentiment analysis. Commun ACM. 56(4): 82–89. http://dx.doi.org/10.1145/2436256.2436274.
Ghurye J, Krings G, Frias-Martinez V. 2016. A framework to model human behavior at large scale during natural disasters. 2016 17th IEEE International Conference on Mobile Data Management (MDM). IEEE, New York, USA. Vol. 1: p. 18–27. http://dx.doi.org/10.1109/MDM.2016.17.
Gross A, Murthy D. 2014. Modeling virtual organizations with Latent Dirichlet allocation: a case for natural language processing. Neural Netw. 58:38–49. http://dx.doi.org/10.1016/j.neu-net.2014.05.008.
Jaiswal K, Noh H, Engler D, et al. 2018. An efficient Bayesian framework for updating PAGER loss estimates. AGU Fall Meeting Abstracts. 2018: NH32B-03.
Jasim HS, Mustafa T. 2018. Comparison of spatial interpolation methods of precipitation and temperature using multiple integration periods. J Indian Soc Remote Sens. 46(7): 1187–1199.
Kankanamge N, Yigitcanlar T, Goonetilleke A, Kamruzzaman M. 2020. Determining disaster severity through social media analysis: testing the methodology with South East Queensland Flood tweets. Int J Disaster Risk Reduct. 42:101360.
Karo IMK, Huda AF. 2016. Spatial clustering for determining rescue shelter of flood disaster in South Bandung using CLARANS Algorithm with Polygon Dissimilarity Function. 2016 12th International Conference on Mathematics, Statistics, and Their Applications (ICMSA). IEEE, p. 70–75.
Kisilevich S, Mansmann F, Nanni M, et al. 2010. Spatio-temporal clustering: a survey. Data mining and knowledge discovery handbook. 2nd ed. New York: Springer Press. p. 855-874. https://www.researchgate.net/publication/265153970.
Kryvasheyeu Y, Chen H, Obradovich N, Moro E, Van Hentenryck P, Fowler J, Cebrian M. 2016. Rapid assessment of disaster damage using social media activity. Sci Adv. 2(3): e1500779.
Li L, Bensi M, Cui Q, Baecher GB, Huang Y. 2021. Social media crowdsourcing for rapid damage assessment following a sudden-onset natural hazard event. Int J Inf Manage. 60:102378.
Lu X, Yu H, Ying M, Zhao B, Zhang S, Lin L, Bai L, Wan R. 2021. Western north pacific tropical cyclone database created by the china meteorological administration. Adv Atmos Sci. 38(4):690–699.
Lu Y, Hu X, Wang F, et al. 2015. Visualizing social media sentiment in disaster scenarios. Proceedings of the 24th international conference on world wide web. p. 1211–1215.
Lucas C. 2012. Multi-criteria modelling and clustering of spatial information. Int J Geograph Inform Sci. 26(10):1897–1915.
Ma Y, Deng Q, Wang X, et al. 2014. Keyword-based semantic analysis of microblog for public opinion study in online collective behaviors. International Conference on Web-Age Information Management, Springer, Cham, p. 44–55.
Mimno D, Wallach H, Talley E, et al. 2011. Optimizing semantic coherence in topic models. Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. p. 262–272. https://aclanthology.org/D11-1024.pdf.
Mori N, Takemi T. 2016. Impact assessment of coastal hazards due to future changes of tropical cyclones in the North Pacific Ocean. Weather Clim Extremes. 11:53–69.

Peduzzi P, Chatenoux B, Dao H, De Bono A, Herold C, Kossin J, Mouton F, Nordbeck O. 2012. Global trends in tropical cyclone risk. Nature Clim Change. 2(4):289–294.

Pielke RA, Jr. 2007. Future economic damage from tropical cyclones: sensitivities to societal and climate changes. Philos Trans A Math Phys Eng Sci. 365(1860):2717–2729.

Ragini JR, Anand PMR, Bhaskar V. 2018. Big data analytics for disaster response and recovery through sentiment analysis. Int J Inf Manage. 42:13–24.

Sakaki T, Matsuo Y, Kurilhara S, et al. 2013. The possibility of social media analysis for disaster management. 2013 IEEE Region 10 Humanitarian Technology Conference, IEEE. p. 238–243.

Sakaki T, Okazaki M, Matsuo Y. 2010. Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on WWW. ACM, Raleigh, NC; New York. p. 851–860.

Simon T, Goldberg A, Adini B. 2015. Socializing in emergencies—a review of the use of social media in emergency situations. Int J Inf Manage. 35(5):609–619.

Spruce M, Arthur R, Williams HTP. 2020. Using social media to measure impacts of named storm events in the United Kingdom and Ireland. Meteorol Appl. 27(1):e1887.

UNISDR (United Nations International Strategy for Disaster Reduction). 2017. UNISDR terminology on disaster risk reduction.

Wald DJ, Seligson HA, Rozelle J, Burns J, Marano K, Jaiswal KS, Hearne M, Bausch D. 2020. A domestic earthquake impact alert protocol based on the combined USGS PAGER and FEMA Hazus loss estimation systems. Earthq Spectra. 36(1):164–182.

Watson H, Finn RL, Wadhwa K. 2017. Organizational and societal impacts of big data in crisis management. J Contingencies Crisis Man. 25(1):15–22.

Wu D, Cui Y. 2018. Disaster early warning and damage assessment analysis using social media data and geo-location information. Decis Support Syst. 111:48–59.

Wu J, He X, Li Y, Shi P, Ye T, Li N. 2019. How earthquake-induced direct economic losses change with earthquake magnitude, asset value, residential building structural type and physical environment: an elasticity perspective. J Environ Manage. 231:321–328.

Ye M, Wu J, Wang C, He X. 2019. Historical and future changes in asset value and GDP in areas exposed to tropical cyclones in China Weather. Wea Climate Soc. 11(2):307–319.

Yu S, Eisenman D, Han Z. 2021. Temporal dynamics of public emotions during the COVID-19 pandemic at the epicenter of the outbreak: sentiment analysis of Weibo posts from Wuhan. J Med Internet Res. 23(3):e27078.

Yuan F, Liu R. 2020. Mining social media data for rapid damage assessment during Hurricane Matthew: feasibility study. J Comput Civ Eng. 34(3):05020001.
Zhang B. 2015. Advance in system and standardization of natural disasters information statistics in China. Catastrophe. 30(3):150–155.

Zhang C, Jiang J, Jin H, Chen T. 2021. The impact of COVID-19 on consumers’ psychological behavior based on data mining for online user comments in the catering industry in China. IJERPH. 18(8):4178.

Zhang Q, Wu L, Liu Q. 2009. Tropical cyclone damages in China 1983–2006. Bull Amer Meteor Soc. 90(4):489–496.