Social Media Text Mining and Flood Disaster Analysis of Small Towns in Southern Shaanxi Qinba Mountain Area Based on Deep Learning

Xin Zhao [1], ZuoBin Wu [2]

[1 2] Department of Urban and Rural Planning, School of Architecture, Xi'an University of Architecture and Technology, China

ABSTRACT

The advent of the dataization era has made social media a new trend and tool for analyzing and managing flood risk. This paper aims to use the BERT-BiLSTM-CRF method to analyze the information on social media such as Weibo, and extract the content related to the flood of small towns in the southern Shaanxi Qinba mountain area. Firstly, we use python crawler to crawl the text data on social media such as Weibo, and then preprocess the data, including removing stop words, punctuation marks, emoticons, etc. Secondly, we use the BERT-BiLSTM-CRF method to perform named entity recognition on the text data, identify entities such as place names, person names, organization names, etc., and annotate them in BIO format. Then, we use methods such as geodetector to geocode the identified place name entities, obtain their latitude and longitude coordinates, and match them with the flood data of small towns in the southern Shaanxi Qinba mountain area, and analyze their flood distribution and flood relationship. Finally, we use flood risk analysis methods, such as flood frequency analysis, flood depth analysis, flood loss analysis, etc., to evaluate the similarity and difference of Weibo information and urban flood risk, and explore the impact and reflection of Weibo information on urban flood risk. In addition, using the high-precision risk analysis method of social media analysis, the flood risk and danger of small towns in the southern Shaanxi Qinba mountain area are evaluated, and the potential impact of flood disaster on the spatial characteristics of small towns is analyzed. The innovation of this paper lies in the combination of deep learning methods and flood analysis methods, extracting the information related to small town floods from large-scale social media data, providing new data sources and analysis methods for towns flood prevention and disaster reduction, and considering the impact factors of flood disaster, providing reference for small town planning and development.

Keywords: Flood Disaster Analysis, climate change, Flood resilience strategies, Small Town Planning, Social Media Text Mining

1. Introduction

In recent decades, climate change has led to an increased frequency of natural disasters, posing severe threats to urban areas and resulting in significant social and economic impacts(Chang et al., 2021). Under the influence of global warming and urbanization, extreme rainfall events and the subsequent urban flooding have become increasingly common. The frequency and intensity of urban flooding in major densely populated cities worldwide are expected to rise substantially. Statistics indicate that floods account for over 30% of the total economic losses caused by natural disasters globally(Shamsuddin, 2020). Floods are among the most destructive disasters on Earth, posing a substantial threat to large populations. Due to

meteorological factors and observational limitations, traditional data sources on flood disasters are increasingly inadequate for addressing the growing frequency of such events(Chen et al., 2020).

To minimize the impact of disasters and enhance recovery, urban resilience is crucial for cities to withstand and recover from external disruptions. Urban resilience, as an advanced approach to addressing natural hazards, has become a key concept and is highly emphasized in scientific and policy discussions(Shi et al., 2021). Due to the interplay of various factors such as cultural adaptation, social, and economic development, the concept of urban resilience is complex and multifaceted, making the creation of universally applicable metrics challenging. As illustrated in the table, research on urban resilience reveals our progressively refined understanding of the concept. Holling introduced the concept of ecosystem resilience, distinguishing between engineering resilience and ecological resilience, and then applied the concept of ecosystem resilience to human social systems, proposing evolutionary resilience(Sharifi, 2019). Engineering resilience emphasizes the system's return to its original state after a disturbance, while ecological resilience prioritizes system balance. On the other hand, evolutionary resilience prioritizes learning and innovation to adapt to new environments(Liu et al., 2020). Although there is no unified definition of urban resilience in the current literature, a relatively consistent understanding exists. Urban resilience refers to the capacity of urban systems to maintain their functionality in the face of natural or unintended disasters, encompassing social, economic, ecological, and infrastructural functions. Various stakeholders implement a series of measures through learning and experience to address and adapt to current and future challenges.

Social media have already established people's connection with other people living both in and beyond the boundaries of urban areas through real-time interaction, crowdsourcing, and accessibility, which strengthens the relationships among social groups. It is a well-known fact that Weibo helps in the retrieval and dissemination of information during urban research on disaster awareness and response. Meanwhile, human behavior related to real-time disaster could be dug out, hence of critical importance under the background of social perspectives of urban resilience, compared with those from the traditional information sources. With the wide application of social media data, social media, with enormous strengths in the dissemination of information, are increasingly one important channel for disaster communication and also an important source for rapid and comprehensive disaster assessment. The use of social media data in flood risk analysis contributes a new method for disaster management. In this way, it can complement traditional flood risk assessment through real-time data and provide a basis for urban planning and disaster response. Through the BERT-BiLSTM-CRF model, flood-related information from Weibo is extracted and geocoded to allow the identification and analysis of flood events in small towns in the southern Shaanxi mountain area of Qinba, and the work shows that the combination of deep learning with social media analytics has the potential to significantly advance flood forecasting and response planning in small towns.

2. Methodology

As shown in Figure 1, This process flow chart shows extraction and analysis information in a Flooding Scenario on Weibo using the BERT-BiLSTM-CRF Model. In data collection, a crawler is used to collect text data resources on Weibo. There follows a step of text preprocessing to exclude from it the so-called "stop-words," punctuation, and other redundant information. Here, the BERT-BiLSTM-CRF method applies named entity recognition, and annotation of place names, personal names, and the names of organizations in BIO format was applied. Then, the place names are geocoded to get its longitude and latitude coordinates by the geodetector, matching with the small town-related data of floods in the southern Shaanxi Qinba mountain area. The flooding analysis in this case is the step to examine the spreading and relationships of flooding. The appraisal examines the commonalities and the disparities between the Weibo information and urban flood risk. Equally, it reflects on the implications for flood risk and hazard in small towns.

Figure 1: Weibo Flood Information Extraction and Analysis Process

2.1 Study area and data

The study area is the Qinba Mountain region of southern Shaanxi, which has a large number of rivers, including the Han River, running through southern Shaanxi. The geographical location and meteorological conditions make the Qinba Mountain region in southern Shaanxi highly prone to heavy rainfall and flooding. In July and August 2020, this region had a huge amount of rainfall, most areas were submerged in water, inducing floods. Therefore, this paper attempts to extract location and water level information related to the flood disaster in social media information. The present study was based on the spatial and temporal characteristics of the selected data after extraction, to explore the process of urban inundation in relation to exposure information. Posts published on the Sina Weibo platform from July to August 2020 had been retrieved and, as shown in this table, key fields comprise the user ID, time stamp, content, and check-in position. The direct scraping of web data removed the sources of sensitive attributes before data collection, but the fields of user nicknames and IDs had not been collected in advance in order to meet the requirements of user privacy protection. We will only collect data captured from the positions of posts on the platform using Python web scraping techniques.

2.2 Acquisition of social media data using web crawlers

The number of users of the Sina Weibo platform is the largest among social media platforms in China. The first-quarter financial report for 2023 indicates that the company had 593 million monthly active users as of March 2023, having increased by approximately 11 million compared to the same period a year earlier. Of this, 95% came from mobile monthly active users. In March 2023, the average number of daily active users was 255 million, an increase of about 3 million over the same period last year. With the so many users, this provided a strong support for this work(Wang et al., 2020).

Information on Weibo will have the characteristic of timely information dissemination. Once the public opinion event bursts, there will be an instant discussion among tens of millions of users, which may cause the hot topic to rise to the top of the headline in real-time hot search, updated once every minute. Python-based web crawlers are used for extraction of social media data on Weibo, and the target information contains user ID, account category, tweet creation time, spatial coordinates, and content information, including text, images, and videos. The timestamp of the tweet is used to classify the check-in data into several time intervals, and the spatial coordinates can be used to determine the geographical location where each tweet has been published. The text content of each tweet is examined in order to find whether the tweet is related to the research topic. In the first place, this keyword search can be an effective filter in translating unstructured data using the specific words and phrases. The approach allows for tuning to fit specific needs and objectives, which means the author can create results that can be reproduced under different sets of circumstances. Yet, if the keywords are nonsensical or sufficiently generic, then the search results can be imprecise and in some cases reflect inaccuracies, given the current need for appropriate keywords to be selected and most appropriate to be refined.

In this study, the keywords used for data crawling include location-specific terms such as "Lueyang County," "Nanzheng District," "Xunyang City," "Pingli County," "Ningqiang County," "Hanyin County," "Ziyang County," "Yang County," "Zhen'an County," "Hantai District," "Langao County," "Xixiang County," "Zhenba County," "Liuba County," "Chenggu County," "Foping County," "Shangzhou District," "Zhenping County," "Baihe County," "Shangnan County," "Danfeng County," "Zhashui County," "Luonan County," "Ningshan County," "Shiquan County," "Shanyang County," "Hanbin District," and "Mian County." Additional keywords include "waterlogging," "flooding," "water inundation," "water immersion," "water rise," "water disaster," "internal flooding," "flood disaster," and "flooding disaster."

2.3 Extracting address entities in Sina Weibo texts

2.3.1 BERT-BiLSTM-CRF model

Extracting location names from text falls under the domain of Named Entity Recognition (NER) in Natural Language Processing (NLP)(Zhang et al., 2020). NER involves identifying structured information, specifically entities with distinct meanings, from unstructured text. Compared to English NER, Chinese NER is more challenging due to issues like segmentation ambiguities and polysemy in the Chinese language. Techniques for implementing NER include

rule-based methods, unsupervised learning methods, feature-based supervised learning methods, and deep learning methods, with the latter being the most widely applied.

Recently, deep learning has achieved significant breakthroughs in the NLP field. When applied to NER, it can learn complex hidden representations without the need for elaborate feature engineering and extensive domain knowledge(Devlin et al., 2018a). For example, Mai et al. used a coupled model of BERT, Bidirectional Long Short-Term Memory (BiLSTM), and Conditional Random Fields (CRF) to identify Chinese place names, referred to as the BERT-BiLSTM-CRF model(Devlin et al., 2018b). BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained model derived from the Transformer architecture, used for various NLP tasks. It operates in two main steps: pre-training and fine-tuning. Compared to traditional neural network language models, BERT has achieved state-of-the-art results in several NLP tasks, including text classification, text similarity, intelligent question answering, text annotation, and named entity recognition(Cui et al., 2021). The BiLSTM model consists of forward and backward LSTMs, allowing it to learn the structure of sequences in both directions, effectively addressing the challenge of distinguishing different sentence sequences. The input to the BiLSTM model is the word embeddings obtained from the BERT model, and the output is a score probability matrix for each label corresponding to each word. Additionally, the LSTM units comprise memory cells and three gates—input, forget, and output gates—that control their inputs and outputs. The CRF model simulates Chinese grammar through conditional probabilities, constraining the model's output to conform to grammatical rules and ensuring the accuracy of the output.

2.3.2 Extracting Process

After conducting tokenization and encoding of the text data, it is passed through a pre-trained tokenizer. The rbt6 is a pre-trained Chinese model from hfl that has been fine-tuned through training on a large Chinese corpus and shows excellent performance in Chinese text tokenization tasks. The main goal of the tokenizer is to change the raw text into a format that a model can understand, using the model language knowledge learned from a huge corpus to improve the text data's tokenization accuracy and model input normalization. A custom dataset class was defined to load and process the text data. A class with a data loading function to load and preprocess text data is able to load data from the local disk and filter data to ensure quality and high efficiency in model training. In particular, it thins out sentences with more than 512 tokens, since unnecessarily long sentences might waste computational resources and degrade the model's performance. The dataset class was written to implement the basic idx and getlen methods, which work with data loaders and, in this way, significantly increase flexibility and efficiency in data processing.

A data collation function was written to organize batch data into formats to be input into the model. Mainly, the code participated in: encoding of the input text in a way that may be further transformed into the form of tensor to perform further processing in the model; padding of the labels to assure that the length of the label is the same as the input text; forming the batch data, which will later participate in the model training and evaluation process, so that raw data is transformed into directly processable format to assure the uniformity and completeness of

data input. After preprocessing, a pre-trained BERT model as a feature extractor is used. Since the BERT model has been trained on such a humongous corpus, it generates super quality text representation. The pre-trained model loads and makes a person fully use the huge amount of linguistic representations learned from the big corpus and enhances a lot of performance for the downstream task.

A Bidirectional Long Short-Term Memory (BiLSTM) layer on top of the BERT module is then added to extract sequence features. In a BiLSTM layer, due to the characteristic of bidirectional information transfer, realizing the retention of dependencies that lie on both sides of a word corpus, so the feature representation is more abundant. A Conditional Random Field (CRF) is added to the top of the BiLSTM layer in a sequence labeling task. The major benefit of the CRF layer is that it essentially models the dependencies between the labels. Consequently, it is improving the accuracy of the label prediction. The realization of a training function, to include the following steps: First, the definition of the optimizer and loss function. Second, data are loaded in batches. Features are extracted and then input into the model to obtain the feature vector, which will be used for forward computation to obtain the prediction result. Third, calculate the loss between the prediction result and the true label and carry out back propagation to update the model parameters; finally, save the trained model parameters for subsequent test and prediction. The cross-entropy loss function, in the process of training, is adopted to compute the difference between the predicted results and the true labels. The gradient descent method is adopted to optimize model parameters, thus enhancing the model performance step by step.

3. Results

In this article, we have effectively applied the BERT-BiLSTM-CRF model for processing data on the social platform Weibo to extract information related to floods in small towns in the Qinba Mountain area of southern Shaanxi. We collected social platform text data using Python web crawlers and preprocesses this data, which includes the deletion of stop words, punctuation marks, emojis, and other non-relevant parts. The named entity recognition of the text data was also carried out to obtain place names, personal names, and organization names as entities and were annotated according to the BIO format rules. Then, the identified place name entities were geocoded for the extraction of their corresponding latitude and longitude. In Figure 1, it can conveniently be understood that flood points that were extracted from Weibo data were so successfully mapped to obtain the distribution of floods in the form of visualization.

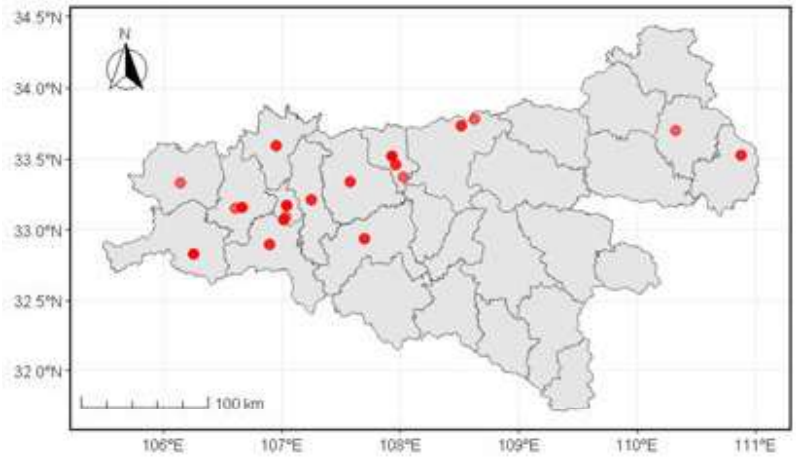

Figure 3: Flood Distribution Map of Towns in the Qinba Mountain Area of Shaanxi

We performed spatial analysis and visualization of flood points in small towns of the Qinba Mountain area in southern Shaanxi. Based on kernel density estimation on the flood data in 2021, 2022, and 2023, we presented the spatial distribution characteristics and trends of the flood points. We successfully located the flood points in these three years and extracted the flood information from Weibo. The flood information was converted into coordinate points and marked on the map. The density distribution of the flood points is represented using blue lines and filled polygons. We compared the distribution of the flood points in different years. We conducted the analysis of the spatiotemporal changes in flood risk, which provided important information for flood risk management. Figure 4 is a comprehensive map where all the kernel density analyses are combined. The red points represent filtered flood points, and the blue lines and polygons represent the kernel density estimates of flood points. This way, we are not only able to see the position of each flood point but also to understand the distribution density of the flood points in the whole area. These analyses promote our understanding of the distribution of flood points in small towns of the Qinba Mountain area in southern Shaanxi and provide a scientific basis for future flood prevention and urban

planning.

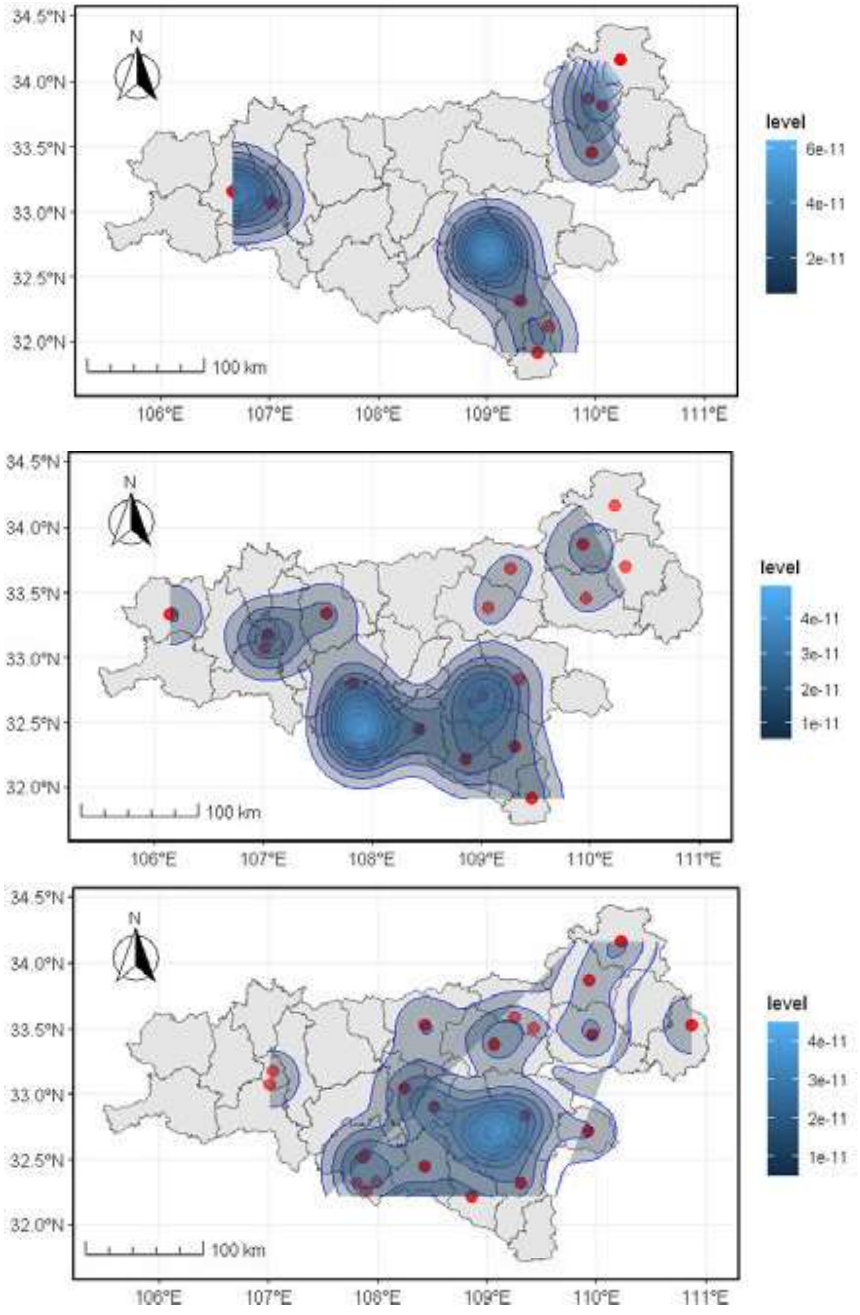
34.5°N
34.0°N
33.5°N
33.0°N
32.5°N
32.0°N
N
100 km
106°E
107°E
108°E
109°E
110°E
111°E
level
6e-11
4e-11
2e-11
level
4e-11
3e-11
2e-11
1e-11
level
4e-11
3e-11
2e-11
1e-11

Figure 4: Comprehensive Flood Point Distribution and Kernel Density Map in Small Towns of the Qinba Mountain Area, Southern Shaanxi (2021-2023)

We made a more detailed spatial analysis of flood points in small towns of the Qinba Mountain area in south Shaanxi. By integrating multi-year flood data, we not only counted the number of flood points in each geographic location, but also visually presented the flood point density on the map with gradient colors. We managed to read the flood point data within the three years of 2021, 2022, and 2023 and converted them into sf objects. We counted the number of flood points of each shapefile and summarized them to compare each year of data. As to the amount of flood points, we used gradient colors from yellow to deep red to present the geographical area, which is the density of a number of flood points, and contribute to an intuitive risk assessment. We made a full-map visualization with all the shapefiles merged in which the distribution of the flood points could be read from the fill colors. Figure 5 provides a comprehensive visualization of the flood point density in small towns within the Qinba Mountain area of southern Shaanxi from 2021 to 2023. Using gradient colors ranging from yellow to deep red, the map highlights the varying density of flood points across different geographical areas. Each color represents the density level, with yellow indicating lower density and deep red indicating higher density.

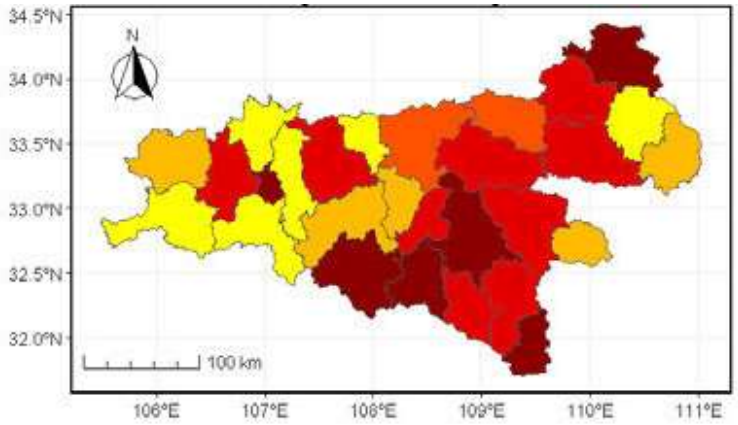

Figure 5: Gradient Map of Flood Point Density in Small Towns of the Qinba Mountain Area, Southern Shaanxi (2021-2023)

This is more reliable for the monitoring of the risk of floods in the small towns in the Qinba Mountain region of southern Shaanxi, and it can offer compelling data arguments for the flood prevention measure and the urban planning provided in the later sections.

4. Conclusion

This research has pioneered a novel way of urban planning and disaster management by combining the analysis of social media information with advanced deep-learning methodologies to evaluate flood risk in small towns of the Qinba Mountain region in southern

Shaanxi. We have effectively extracted and processed the flood information through Weibo with the BERT-BiLSTM-CRF model and thereby obtained a comprehensive understanding of the spatial distribution and features of flood events in this region. Our approach was to gather large-scale social media data using Python web scrapers and then pre-process the data carefully by filtering out irrelevant data. The BERT-BiLSTM-CRF model was used to extract and categorize named entities of place names, which were then geocoded to extract their actual latitude and longitude. With the help of this, the spread of floods in small towns over a few years could be visualized and interpreted to extract meaningful insights into the temporal and spatial patterns behind the flood risks.

The results showed that the use of deep learning and its combination with social media analytics was possible to realize real-time flood risk assessment and urban planning. The spatial analysis and visualization of the flood points played an essential role in managing flood risk and urban resilience planning. The combination with the multi-year flood data enabled us to assess the population density and distribution of the flood points, thereby providing a scientific basis for the formulation of effective flood prevention measures and urban planning policy. This research also demonstrates the necessity to incorporate social media data in the disaster management system, as it provides a comprehensive and real-time source of data that supplements the traditional approaches. The incorporation of advanced computational methods and multi-objective optimization algorithms in urban planning could help to improve the accuracy and reliability of flood risk analysis as well as enable urban development, which is more robust and sustainable.

References

Chang, Heejun, Pallathadka, Arun, Sauer, Jason, Grimm, Nancy B., Zimmerman, Rae, Cheng, Chingwen, Iwaniec, David M., Kim, Yeowon, Lloyd, Robert and McPhearson, Timon (2021), Assessment of Urban Flood Vulnerability Using the Social-Ecological-Technological Systems Framework in Six US Cities. *Sustainable Cities and Society*, 68: 102786.

Chen, Changkun, Xu, Lili, Zhao, Dongyue, Xu, Tong and Lei, Peng (2020), A New Model for Describing the Urban Resilience Considering Adaptability, Resistance and Recovery. *Safety science*, 128: 104756.

Cui, Yiming, Che, Wanxiang, Liu, Ting, Qin, Bing and Yang, Ziqing (2021), Pre-Training with Whole Word Masking for Chinese Bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina (2018a), Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton and Toutanova, Kristina (2018b), Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Liu, Wei and Song, Zhaoyang (2020), Review of Studies on the Resilience of Urban Critical Infrastructure Networks. *Reliability Engineering & System Safety*, 193: 106617.

Shamsuddin, Shomon (2020), Resilience Resistance: The Challenges and Implications of Urban Resilience Implementation. *Cities*, 103: 102763.

Sharifi, Ayyoob (2019), Urban Form Resilience: A Meso-Scale Analysis. *Cities*, 93: 238–252.

Shi, Yijun, Zhai, Guofang, Xu, Lihua, Zhou, Shutian, Lu, Yuwen, Liu, Hongbo and Huang, Wei (2021), Assessment Methods of Urban System Resilience: From the Perspective of Complex Adaptive System Theory. *Cities*, 112: 103141.

Wang, Bo, Loo, Becky PY, Zhen, Feng and Xi, Guangliang (2020), Urban Resilience from the Lens of Social Media Data: Responses to Urban Flooding in Nanjing, China. *Cities*, 106: 102884.

Zhang, Hongwei, Ren, Fu, Li, Huiting, Yang, Renfei, Zhang, Shuai and Du, Qingyun (2020), Recognition Method of New Address Elements in Chinese Address Matching Based on Deep Learning. *ISPRS international journal of geo-information*, 9(12): 745.