SASBE 2025 aims to encourage the international exchange of innovative ideas between researchers from academia and industry. In addition to knowledge dissemination, the conference offers a valuable platform for professional networking, particularly benefiting university professors, graduate students, and postdoctoral researchers.

Research Article

Ontology-Guided Extraction of Infrastructure Damage Information during Floods Using Large Language Models

Zaishang Li[1,2], Dina D'Ayala[1, 3]

[1]Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, United Kingdom
[2]Department of Construction Management, Tsinghua University, Beijing 100084, China
[3]UNESCO Chair in Disaster Risk Reduction and Resilience Engineering, Department of Civil, Environmental and Geomatic Engineering, University College London, London WC1E 6BT, United Kingdom
Correspondence: zaishang.li@ucl.ac.uk

Abstract

Urban floods are becoming increasingly frequent due to intensified extreme rainfalls and ongoing urbanization, posing significant threats to both human life and the built environment in cities worldwide. Timely, high-resolution data on infrastructure damage during such events is essential for effective emergency response and long-term adaptation planning. However, conventional data collection methods often fail to capture large-scale and event-specific impacts. In contrast, social media have emerged as a publicly available source of real-time user-generated observations, which can complement current flood monitoring approaches. This study proposes a novel approach to automatically extract structured infrastructure damage information from social media posts related to flood events. A prompt engineering strategy for Large Language Models (LLMs), guided by a domain-specific ontology of infrastructure assets and damage typologies, is developed to ensure classification consistency and extraction reliability. A case study of the Zhengzhou "7·20" flood event in China illustrates the effectiveness of the proposed method. The findings demonstrate that while the LLM-based approach achieves high precision, its overall performance is constrained by the model's limitations in processing long-form texts. Nevertheless, this research marks a significant step towards leveraging LLMs for real-time disaster response. The proposed method can directly assist urban emergency managers by enabling the rapid collection of social media-based disaster information, thereby complementing conventional approaches. Moreover, it provides a scalable foundation for developing essential infrastructure damage databases, which are vital for advancing research on infrastructure resilience and informing downstream analyses.

Keywords: infrastructure damage; resilience; flood; large language model; information extraction

Highlights

- A novel LLM-based approach for structured information extraction from social media during floods
- A domain-specific ontology is developed to guide LLM for reliable and consistent extractions
- Demonstrated LLM's precision in information extraction while identifying its limitations in handling long-form texts.

1 Introduction

Floods are becoming increasingly frequent due to climate change and ongoing urbanisation, posing significant threats to both human life and the built environment in cities worldwide [1, 2]. According to Centre for Research on the Epidemiology of Disasters (CRED) [3], 142 flood events occurred globally in 2024, affecting 48.8 million people with 5 883 deaths, ranking the first among all the disaster types, and causing USD 32.8 billion economic losses, ranking the second. Infrastructure, such as roads, bridges, railways, power stations, and hospitals, is particularly susceptible to damages or loss of function during floods, which can severely disrupt societal operations. For instance, the 2024 Spanish floods destroyed at least 232 kilometres of road and rail tracks [4]. In Valencia region, 155,000 homes lost the access to electricity and the malfunction of telephone networks and flooded roads further hampered the efforts to reach stricken communities [5].

Rapid information collection on infrastructure damage is crucial for efficient disaster response following a flood event. Existing flood monitoring approaches mainly include manual inspection, sensor-based methods, and remote sensing [6-8]. However, these approaches have their own inherent limitations or challenges in practice. For example, while manual inspections offer flexibility, they are labour-intensive, inefficient, and limited in scope, thus hindering timely responses to large-scale flooding. Hydrological sensors, conversely, are capable of producing high-precision, real-time data, but their utility is constrained by considerable deployment costs and limited spatial coverage. Remote sensing imagery, including both optical and radar types, provides extensive flood monitoring capabilities; however, its effectiveness is compromised by factors such as atmospheric conditions, ground occlusion, and imaging artefacts that may lead to misclassification of inundated areas [8].

Social media, as a form of publicly crowdsourced data, can complement current flood monitoring approaches [8, 9]. During flood events, individuals frequently post real-time observations and experiences about the events on social media platforms. Unlike the physical signals monitored by sensors, social media text reflects people's subjective perception, focusing more on the societal impacts of floods. Despite challenges such as unstructured formats, subjective biases, and the risk of misinformation, social media posts offer valuable insights into overlooked or locally specific flood impacts, and therefore serve as a crucial complement to conventional monitoring systems.

Relevant research interests include hazard assessment, flooding area identification, water depth estimation, and sentiment analysis [8-12], all of which involve extracting impact-related information from social media posts. There are also some explorations looking into infrastructure damage during floods. For instance, Zhang, et al. [13] proposed a taxonomy of community disruption events, such as flood control infrastructures, transportation, housing, and utilities and supplies, to categorise tweets for societal impact assessment. Mihunov, et al. [14] applied latent Dirichlet allocation (LDA) topic modelling to subset the Twitter data related to infrastructure, the outcomes of which were used for further analyses of temporal and spatial patterns. However, existing infrastructure-related studies on social media primarily focus on event categorisation, with limited attention to the detailed characterisation of flood-induced impacts on infrastructure (e.g., the type of infrastructure affected and the nature of the damage). This gap constrains a nuanced understanding of infrastructure resilience to flooding.

Conventional information extraction methods from social media, such as keyword matching or traditional machine learning classifiers, often struggle with the short, noisy, and non-standard nature of microblogs [15, 16]. Large language models (LLMs), which have demonstrated remarkable ability in processing natural language [17], have been employed in structured information extraction from various texts [18, 19]. Similar attempts have also been made in flood contexts, where LLMs were used to extract flooding depth and the number of people trapped [20]. However, to our knowledge, no studies have addressed the extraction of detailed, infrastructure damage-related information from social media during floods using LLMs.

This study aims to fill this critical gap by proposing an LLM-based approach to extract structured infrastructure damage-related information from social media during flood events. A domain-specific ontology of infrastructure assets and damage typologies is developed to guide the information extraction process. The outcomes of this study will enable urban emergency managers to rapidly collect and synthesise disaster information from social media, serving as a valuable complement to conventional approaches. Furthermore, the proposed approach supports the development of flood-related infrastructure damage databases from public sources, which are critically needed for advancing research on infrastructure resilience.

2 Methodology

2.1 Data Preparation

The Zhengzhou "7·20" flood event was selected as the case study for this research. From 18:00 on 18 July to 00:00 on 21 July 2021, the city of Zhengzhou experienced an extreme rainfall event that caused widespread disruption, affecting approximately 14.79 million people and resulting in direct economic losses of 120.06 billion yuan [20].

To build a relevant dataset, Weibo posts were collected using an open-source web-scraping tool "Weibo-search"[1]and a keyword-based search approach. Weibo, a prominent Chinese social media platform, had approximately 570 million monthly active users during the event period[21]. Posts spanning from 14 to 28 July 2021 were initially collected using the search term "Zhengzhou heavy rain". A subsequent refinement step was then employed to filter the retrieved texts for relevance to infrastructure damage. Specifically, a post was selected if it contained both infrastructure-related nouns (e.g., "road", "water supply") and damage-related verbs (e.g., "flood", "collapse"). A further manual refinement was conducted to select posts with clearer descriptions of infrastructure damage events, particularly those that explicitly mentioned the location, infrastructure type, and the nature of the structural or functional damage. This process resulted in a final dataset comprising 138 Weibo texts.

Each of the selected Weibo texts was then manually annotated with structured information, including: (1) temporal attributes: date, time, and time type; (2) locational attributes: city and location; (3) hierarchical infrastructure types: infrastructure types L1-L3; and (4) hierarchical damage types: damage types L1-L2. The structure of these labelled fields is further detailed in Table 1. If a single Weibo text described multiple distinct infrastructure damage events, each event was individually identified and annotated, all linked to the original post. In total, 787 distinct events were annotated, creating a comprehensive ground-truth dataset for validating the proposed information extraction approach.

Table 1. Fields of the labelled Weibo texts

Field	Description
Date	The date when the event happened. The post date would be applied if the event date was not reported
Time	The time when the event happened. The post time would be applied if the event time was not reported.
Time type	Labelled as "reported" if both the date and time of the event were reported, otherwise as "post time"
City	The city where the event happened
Location	The detailed location of the event
Infrastructure type L1	The first level of the infrastructure type
Infrastructure type L2	The second level of the infrastructure type
Infrastructure type L3	The third level of the infrastructure type
Damage type L1	The first level of the damage type
Damage type L2	The second level of the damage type

[1] https://github.com/dataabc/weibo-search

2.2 Ontology Design

To guide the LLM-based structured information extraction, a domain-specific ontology was developed. This ontology comprises two primary components: infrastructure types and damage types.

The infrastructure ontology is hierarchically structured into three levels (L1-L3). Level 1 (L1) refers to the main categories of infrastructure, including transportation systems, power systems, water supply systems, drainage systems, energy systems, telecommunication systems, and building systems. Level 2 (L2) and Level 3 (L3) represent sub-categories. For example, within transportation systems, L2 and L3 could specify roads and traffic lights, respectively. For water supply systems, L2 and L3 could detail water plants and pumps.

The damage ontology is similarly divided into two hierarchical levels (L1 and L2). L1 indicates the basic types of damage, which are classified into three main categories: structural damage, with L2 subsets of "damaged" and "collapsed"; functional damage, with L2 subsets of "partially functional" and "non-functional"; flooding, with L2 subsets specifying a series of water depth ranges.

This comprehensive ontology serves as a predefined schema that is embedded within the LLM's prompt. This enables the model to extract information from the Weibo texts and map it to a structured, consistent format, ensuring the output is both relevant and coherent.

2.3 Information Extraction

The LLM "Qwen-turbo" API[2], developed by Chinese company Alibaba Cloud, was selected in this study given the cost, stability, and support for Chinese language. To ensure the consistent and accurate extraction of structured information from Weibo texts, this study designed a modular prompt framework tailored for the LLM. The prompt comprised six interrelated components. First, the task definition specified the model's role and overarching objective, thereby constraining its focus to event identification and information extraction. Second, a set of core rules established the fundamental principles of extraction, covering event segmentation, filtering, and merging. Third, the ontology designed in Section 2.2 was embedded to provide a structured dictionary for classifying infrastructure and damage types, which served as the backbone of the extraction process. Fourth, field-specific rules outlined additional instructions for information retrieval and prescribe the required output format for the fields "Date", "Time", "Time type", "City", and "Location". Fifth, the output structure enforced a reasoning process prior to the final answer, ensuring that the model demonstrated its intermediate reasoning steps before presenting the structured output. Sixth, representative examples were incorporated to guide the model's responses through demonstration, followed by an explicit execution trigger that initiated the processing of Weibo data. Collectively, this layered prompt design enhanced the reliability, robustness, and transparency of LLM-driven information extraction task.

2.4 Performance Evaluation

Due to the fact that the number of identified events from one Weibo text by the LLM may not match the number of the true events, it is not feasible to evaluate the performance of the information extraction by simply comparing the identified event with the truth one by one. This study employed a matching strategy that paired identified events with their corresponding ground-truth events. Since each event has ten fields, a similarity score was calculated between each identified and true event, by summing up the scores of the fields where the identified and true events shared the same values. For the "Location" field, the fuzz.ratio function[3] was employed to quantify the similarity considering non-standard descriptions of locations. This string similarity measure is based on edit distance and provides a straightforward and computationally efficient way to capture the overlap between two texts.

[2] https://qwen.ai/apiplatform

[3] https://github.com/seatgeek/thefuzz

The score allocated to each field can be found in Table 2. The score for "Location" was set the highest as the location was regarded as the identifier of an event. Fields "Date", "Time", "Time type", and "City" were given small scores as they had less importance in information than infrastructure types and damage types. The matching algorithm determined an optimal one-to-one matching that maximised the total similarity across all pairs, ensuring that each predicted event was paired with at most one true event. Unmatched events are treated as false positives or false negatives.

Table 2. Score for each field

Field	Score
Date	0.25
Time	0.25
Time type	0.25
City	0.25
Location	5.00
Infrastructure type L1	1.00
Infrastructure type L2	1.00
Infrastructure type L3	0.50
Damage type L1	1.00
Damage type L2	0.50

3 Results

The results of the structured information extraction task are presented in Table 3. True Positives (TP) are defined as events that were correctly identified by the LLM. Due to the challenge of correctly extracting information for all fields, an event was classified as a TP if its evaluation score was 8.25 or above. This scoring criterion ensured that (1) the location information was correct and (2) the model did not simultaneously misclassify both the infrastructure and damage types (L1), thereby ensuring the extracted event was sufficiently informative. False Positives (FP) represent events that the LLM incorrectly identified, while False Negatives (FN) represent events that the model failed to identify. The True Negative (TN) set was empty, as all texts in the ground-truth dataset corresponded to actual infrastructure damage events.

Table 3. Statistics of information extraction results

Item	Value
TP	497
FP	210
TN	0
FN	290

To evaluate the performance of the proposed approach, four standard metrics were employed: accuracy, which measures the proportion of all correct classifications; precision, which quantifies the proportion of positive classifications that are genuinely positive; recall, which measures the proportion of actual positive events that were correctly identified; and the F1 score, the harmonic mean of precision and recall. The values of the metrics were calculated as below based on the values from Table 3.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 49.8\%$$

$$Precision = \frac{TP}{TP + FP} = 70.3\%$$

$$Recall = \frac{TP}{TP + FN} = 63.2\%$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} = 66.5\%$$

The results show that the model achieved a relatively high precision of over 70%, indicating that the information it did extract was of moderate to high quality. However, the lower recall (63.2%) and accuracy (49.8%) suggest that the model missed a significant number of events (high FN) and made numerous incorrect classifications (high FP). The low accuracy is particularly notable due to the substantial numbers of both FP and FN.

Furthermore, a more granular analysis was conducted to investigate the model's performance on different infrastructure and damage types (L1), with the recall values shown in Table 4.

The LLM demonstrated varied performance across different categories. The results in Table 4 indicate that the LLM performed exceptionally well in identifying damage to power systems (92.9%) and energy systems (100.0%), and quite well for building systems (82.3%). In contrast, its performance was significantly lower for transportation (58.2%) and water supply systems (61.5%). Notably, the model completely failed to identify any events related to drainage systems (0.0%) since there were only two drainage systems-related events in the ground-truth dataset. Regarding damage types, the model achieved the highest recall for flooding (70.0%), which is understandable given that keywords related to flooding (e.g., "flooding," "waterlogging") are more consistently and explicitly mentioned in social media texts than the diverse and often more ambiguous keywords associated with structural or functional damage.

Table 4. Recall values for different infrastructure and damage types

Data field	Category	Recall value	Average recall value
Infrastructure type L1	Transportation system	58.2%	68.1%
	Power system	92.9%	
	Water supply system	61.5%	
	Drainage system	0.0%	
	Energy system	100.0%	
	Telecommunication system	66.7%	
	Building system	82.3%	
Damage type L1	Structural damage	65.6%	67.2%
	Functional damage	66.1%	
	Flooding	70.0%	

4 Discussion

This study presents a novel LLM-based approach for structured information extraction from social media during flood events. The results, while validating the model's potential, also highlight several key performance challenges.

The initial performance metrics revealed a notable disparity between the model's precision and recall, with a high number of FN. Upon closer investigation, it was found this was primarily due to the LLM's inability to fully process long Weibo posts. Two particularly lengthy posts, each over 1,900 Chinese characters and containing summaries of multiple events, were responsible for a substantial portion of these missed extractions - the LLM was able to correctly identify only approximately 20% of the events within these lengthy texts.

To understand the impact of this limitation, these two long texts were removed from the ground-truth dataset, and the performance metrics were re-calculated, as shown in Table 5. The results demonstrate a dramatic improvement across most metrics - the accuracy, recall, and F1 score increased significantly to 60.0%, 81.9%, and 75.0%, respectively. This finding clearly demonstrates that the LLM's context window limitation is the main bottleneck for performance.

In addition, while the model demonstrated good performance on certain infrastructure types like energy (100% recall) and power systems (92.9% recall), its lower scores on others, such as transportation (58.2% recall), indicate that performance is highly dependent on the nature and consistency of the language used in the posts. This suggests that the model's effectiveness is tied to how explicitly and uniformly the public describes different types of infrastructure damage.

Table 5. Results of information extraction after the removal of two long texts

Item	Value	Change
TP	461	-36
FP	205	-5
TN	0	0
FN	102	-188
Accuracy	60.0%	10.2%
Precision	69.2%	-1.1%
Recall	81.9%	18.7%
F1 Score	75.0%	8.5%

5 Conclusions

In this study, an LLM-based approach was developed to extract structured infrastructure damage information from social media posts during flood events. The methodology involved creating a domain-specific ontology for infrastructure and damage types, which was used to guide the LLM's extraction process. This approach was applied to a real-world dataset of Weibo posts from the Zhengzhou "7·20" flood event. The results demonstrated that while the model achieved a respectable precision of over 70%, its overall performance was significantly impacted by its inability to process long-form texts, a common characteristic of social media summaries.

This study marks a significant step towards leveraging LLMs for real-time disaster response. The proposed method can directly assist urban emergency managers by enabling the rapid collection of social media-based disaster information, thereby complementing conventional approaches. Moreover, it provides a scalable foundation for developing essential infrastructure damage databases, which are vital for advancing research on infrastructure resilience and informing downstream analyses. Moving forward, the following research plans to address the identified limitations by exploring advanced prompt engineering techniques, such as chunking for long texts, and investigating the potential of fine-tuning smaller and specialised LLMs for enhanced performance and robustness. The development of more adaptive ontologies that account for the nuances of human language also remains a key priority.

Acknowledgements

Authors gladly acknowledge the support of the UNESCO Chair in Disaster Risk Reduction and Resilience Engineering, which funds some of their time.

Funding

This work was funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No. EP/Z001978/1) and the National Natural Science Foundation of China (Grant No. 72304162).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of 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

[1] J. Ji, L. Fang, J. Chen, and T. Ding, "A novel framework for urban flood resilience assessment at the urban agglomeration scale," *International Journal of Disaster Risk Reduction,* vol. 108, p. 104519, 2024.

[2] Y. Wang *et al.*, "A novel framework for urban flood risk assessment: multiple perspectives and causal analysis," *Water research,* vol. 256, p. 121591, 2024.

[3] CRED, "2024 Disasters in Numbers," Brussels, Belgium, 2025. [Online]. Available: https://files.emdat.be/reports/2024_EMDAT_report.pdf

[4] T. Medrano. "Spain's catastrophic floods by the numbers: At least 219 dead, 93 missing and billions in damage." https://apnews.com/article/spain-floods-valencia-numbers-21fccd4a9eba4fada0745db0a5a0dd12 (accessed 30 August, 2025).

[5] FRANCE 24. "Spanish rescuers race to save flash flood victims as death toll tops 90." https://www.france24.com/en/europe/20241030-several-bodies-found-as-heavy-rains-flash-floods-slam-spain (accessed 30 August, 2025).

[6] M. I. Zakaria and W. A. Jabbar, "Flood monitoring and warning systems: A brief review," *Journal of Southwest Jiaotong University,* vol. 56, no. 3, pp. 140-156, 2021.

[7] S. A. S. Bukhari *et al.*, "Review of flood monitoring and prevention approaches: a data analytic perspective," *Natural Hazards,* vol. 121, no. 5, pp. 5103-5128, 2025.

[8] H. Hou, L. Shen, J. Jia, and Z. Xu, "An integrated framework for flood disaster information extraction and analysis leveraging social media data: A case study of the Shouguang flood in China," *Science of the total environment,* vol. 949, p. 174948, 2024.

[9] Q. Guo, S. Jiao, Y. Yang, Y. Yu, and Y. Pan, "Assessment of urban flood disaster responses and causal analysis at different temporal scales based on social media data and machine learning algorithms," *International Journal of Disaster Risk Reduction,* vol. 117, p. 105170, 2025.

[10] J. Peng and J. Zhang, "Spatiotemporal assessment of urban flooding hazard using social media: A case study of Zhengzhou '7· 20'," *Environmental Modelling & Software,* vol. 176, p. 106021, 2024.

[11] H. Li, Y. Han, X. Wang, and Z. Li, "Risk perception and resilience assessment of flood disasters based on social media big data," *International Journal of Disaster Risk Reduction,* vol. 101, p. 104249, 2024.

[12] M. W. Boota *et al.*, "How effective is twitter (X) social media data for urban flood management?," *Journal of Hydrology,* vol. 634, p. 131129, 2024.

[13] C. Zhang, W. Yao, Y. Yang, R. Huang, and A. Mostafavi, "Semiautomated social media analytics for sensing societal impacts due to community disruptions during disasters," *Computer‐Aided Civil and Infrastructure Engineering,* vol. 35, no. 12, pp. 1331-1348, 2020.

[14] V. V. Mihunov, N. H. Jafari, K. Wang, N. S. Lam, and D. Govender, "Disaster impacts surveillance from social media with topic modeling and feature extraction: Case of Hurricane Harvey," *International Journal of Disaster Risk Science,* vol. 13, no. 5, pp. 729-742, 2022.

[15] Y. Li, Y. Zhou, X. Hu, Q. Li, and J. Tian, "A method for named entity recognition in social media texts with syntactically enhanced multiscale feature fusion," *Scientific Reports,* vol. 14, no. 1, p. 28216, 2024.

[16] L. Derczynski *et al.*, "Analysis of named entity recognition and linking for tweets," *Information Processing & Management,* vol. 51, no. 2, pp. 32-49, 2015.

[17] H. Naveed *et al.*, "A comprehensive overview of large language models," *ACM Transactions on Intelligent Systems and Technology,* vol. 16, no. 5, pp. 1-72, 2025.

[18] J. Dagdelen *et al.*, "Structured information extraction from scientific text with large language models," *Nature communications,* vol. 15, no. 1, p. 1418, 2024.

[19] Z. Ma, J. E. Santos, G. Lackey, H. Viswanathan, and D. O'Malley, "Information extraction from historical well records using a large language model," *Scientific Reports,* vol. 14, no. 1, p. 31702, 2024.

[20] S. Wang, R. Li, H. Wu, J. Li, and Y. Shen, "Fine-grained flood disaster information extraction incorporating multiple semantic features," *International Journal of Digital Earth,* vol. 18, no. 1, p. 2448221, 2025.

[21] Statista. "Number of monthly active users of Weibo Corporation from 2nd quarter 2019 to 2nd quarter 2025." https://www.statista.com/statistics/795303/china-mau-of-sina-weibo/ (accessed 29 August, 2025).