Referee report on the manuscript

“Content Analysis of Multi-Annual Time Series of Flood-Related Twitter (X) Data”

submitted to the journal

Natural Hazards and Earth System Sciences

The authors have undertaken a relevant study by leveraging social media data (X, formerly Twitter), to detect topics related to hazards using advanced techniques like SBERT and clustering. The use of pre-trained language models and unsupervised clustering methods demonstrates a sophisticated approach to analyzing the vast and complex data generated on social platforms during disasters. This research contributes valuable insights into how social media can be harnessed for disaster management and impact assessment, and the methodology employed has the potential to improve disaster response strategies. Additionally, the paper is well-structured, enhancing readability and scientific clarity. The figures and tables are clear and effectively represent the results, aiding the overall understanding of the study.

Following are a few of the concerns that require clarification.

1. Data collection from X (formerly Twitter) has become increasingly difficult and costly. In earlier versions of the Twitter API, researchers could easily gather data using queries and relevant keywords without limitations, which contrasts sharply with the current restrictions on the X developer platform. Including a detailed explanation of your data collection strategy, ideally in Section 2, would be highly beneficial. I understand that some data and data collection details are mentioned in section 2.4. However, including how the data was gathered will be useful. It could serve as valuable guidance for researchers working in the social media field, helping them navigate these new challenges.

2. In Table 1, the events E1, E2, E3, E4, and E5 are well-documented. However, I am curious about how the severity or impact of the floods was measured. Was this evaluation based on factors such as the number of casualties, the extent of the affected areas, or other key indicators? Clarifying this would provide a clearer understanding of the flood classification.

3. While the observation that greater topic diversity in tweets may indicate high-impact events is intriguing (mentioned under section 3.1), I have some concerns regarding the potential for ambiguity in the analysis of specific

disaster impact factors, such as casualties or damages. The detection of a higher number of topics may introduce noise and reduce clarity when focusing on key indicators of event severity. In particular, diverse topics may dilute attention on actionable information and make it more difficult to derive clear insights related to specific impact metrics. I suggest that the authors address how they mitigate this potential ambiguity when assessing diverse topics. Clarifying whether any topic weighting or hierarchies are applied could strengthen the argument and ensure that topic diversity does not introduce noise.

4. While the observation that people tend to be more proactive during low-impact flooding is an interesting insight (line 250), I have some concerns about the challenges involved in processing social media data. During disaster events, social media posts often contain typos, grammatical issues, unstructured sentences, and sometimes even lack critical information such as location details. Additionally, many users share images or videos without accompanying text, further complicating the extraction of useful insights. Given these limitations, there is a high chance of missing relevant data if basic pre-processing techniques are used. It would be valuable if the authors could clarify the specific pre-processing methods they applied to address these issues. For example, how were posts with poor grammatical structure handled? Were posts with missing location data discarded or enriched through external methods? Moreover, did the analysis account for non-textual posts, such as images or videos, that might carry significant information? Addressing these points would enhance the robustness of the findings and provide a clearer understanding of how diverse social media data was managed during different phases of the disaster event.

5. The use of SBERT in this study is impressive. However, pre-trained models like SBERT are typically trained on general-purpose datasets. Given that social media data often contains domain-specific terms and jargon, particularly in the context of disaster management, SBERT may not be fully equipped to accurately represent these specialized terms. I recommend that the authors clarify whether they fine-tuned SBERT on disaster-related data or other domain-specific corpora to enhance its performance for this particular application.

6. While the use of SBERT and HDBSCAN for topic detection is appropriate, the paper does not provide any evaluation metrics to assess the performance of the clustering model. Including results from common clustering evaluation metrics, such as Silhouette Score, or others, would give a clearer indication of the quality of the identified clusters. I recommend that the authors include these metrics to validate the effectiveness of their clustering approach. I believe that some details might be available in supplementary materials as mentioned in line 125, but the links to these materials cannot be found in the document.
7. One of the main limitations of using social media data is its reliability and credibility, particularly in disaster management scenarios where the stakes are high. The potential for false information or rumors can undermine the effectiveness of response efforts. Therefore, it would be beneficial if the results of the model were tested against specific disaster events and compared with official records for validation. Such an approach would enhance the credibility of the findings and ensure that the insights derived from social media are both actionable and trustworthy.

Minor typos

1. Line number 60 - Reference is not given within brackets; Konya and Nematzadeh (2024).
2. Line number 98 - I assume the usage of the word “The” came in twice was a mistake.
3. While including in-text citations with multiple references (such as in line 15), the order of reference can be year-wise. This improves readability.