Review of egusphere-2026-722

Unravelling information on impactful geo-hydrological hazard events with HazMiner, a multilingual text mining method developed through a global scale coverage application

The paper presents a new method “HazMiner”, which contributes to the field of automated hazard documentation. The authors present a multilingual, paragraph-based approach to automatically analyse newspaper articles from online sources worldwide and point out the use of theit approach for improving data coverage in the Global South. Texts are downloaded from the GDELT API, then translated if they are not in English. Subsequently the paragraphs are filtered and with a fine-tuned version of dilstilBERT and classified into three categories, flood, landslides and flash floods with a fine-tuned DeBERTa model. In a third step the authors extract impact, time and location from the paragraph and geocode the found instances of a potential hazard. These instances are then clustered with OPTICS and DBSCAN and transferred into an events database. Altough many limitations are addressed I have doubts that the method is useful in the current state since the authors do not present a robust validation of their findings.

Major comment

The authors repeatedly state that HazMiner "outperforms" or "significantly improved" existing datasets compared to EM-DAT and the Global Landslide Catalog (GLC). However, this claim is based solely on the *quantity* of detected events — not on their accuracy. The authors describe, that they compared their events with EM-DAT, but also state that the comparison is not directly possible because of the country scale and difference in timing. Why did the authors choose this comparison is not understandable. Further, in line 287, the authors state that there is no reference dataset with sufficiently detailed information for validation, but they could combine different datasets to validate different aspects. For example the HazMiner data could be compared to the timing of the Global Flood Monitor data. The authors claim that there is a method that detects more events does not outperform another unless it can be shown that those additional detections are correct. The possibility that many of the additional events are false positives is never adequately ruled out. The described evaluation is purely internal and cannot replace a validation with verified data in an area where we know exactly how many floods and landslides happened. This method can only be robust if it is properly validated and accuracies are reported. Given that events are constructed from clusters of paragraphs, even a moderate false positive rate at the paragraph level could result in spurious events, particularly in regions with sparse reporting.

Even the internal evaluation and the limitations of the model that the authors mention directly are not very promising. For example In Line 167 the authors state that not all paragraphs relate to an actual hazard event due to articles on early warning. However, there are other scenarios in which an article may not report an event directly, but relate to an event in the past, but still be formulated in present tense. Further, the translation error that the authors state in 385 needs to be addressed. This method cannot claim to effectively work in many languages, if it consistently shows such grave translation errors.

Minor Comments

The clustering parameters (xi = 0.1, epsilon = 0.5°, minimum paragraph thresholds) are selected based on visual interpretation of the data. A more quantitative justification or sensitivity analysis of these parameters would strengthen the analysis

L24: Citation needed for the claimed similarity in geo-environmental drivers.

L25 :Citation needed for the claim that events usually make the news. This sentence seems very anecdotal to me.

L55: “real-world” observation is a strange term to refer to measurements. One could argue that the remote sensing system are just as much part of our real world as a sensor on the ground.

L117: Which URLs were used? This is a critical information, especially because the results later analyze the geographical distribution of paragraphs, so this result probably directly reflect the choice of URLs.

L150-155: There may be many more instances of the same issue especially in different languages. It is very concerning methodologically, that you found these two instances of misuse of words in newspaper articles, without clarifying further how big the effect of this is on your model estimates.

Figure 2: The double legend on the maps is misleading