Response letter: referee comment 1

We would like to sincerely thank the reviewer for the detailed, constructive, and insightful review of our manuscript. We greatly appreciate the reviewer's careful assessment of both the methodological design and the resulting datasets. The reviewer's comments have been useful to help us to clarify terminology, improve the storyline of the manuscript, improve methodological transparency, and better formulate the limitations, assumptions, and applicability of the HazMiner framework. In particular, the reviewer's insights regarding the use of natural language processing, and evaluation strategies resulted in substantial revisions that we believe significantly improved the clarity, robustness, and scientific value of the manuscript.

Below, we address each comment in detail. Our answers to comments in italic bold.
Comments by the Anonymous Referee are in normal style.

General Comments

This paper addresses a pressing research gap: the creation of reliable and validated (global) text-based datasets for natural hazard impact and adaptation research. The authors propose and implement an automated pipeline with 14 steps to cluster news paragraphs into temporally and geographically thematic clusters for three types of hazards: floods, flash floods and landslides. The work demonstrates a remarkable effort in coordinating all processing steps for large scale multilingual text data and analyzing the resulting structures. Importantly, known limitations and biases are transparently and extensively discussed in Section 4. In that regard, it is a valuable contribution with good potential to pave the way for relevant further research in various fields. Still, a few drawbacks must be pointed out and require improvement or at least a more detailed discussion of consequences and possible solutions.

Specific Comments

1. There seems to be a misuse of the term LLMs in the text. As of now, "LLM" is being mostly used to refer to generative AI models, i.e. decoders for text completion. Although Transformer-based encoders like BERT can also be called large language models when they are trained for an underlying language modelling task, it seems that the pre-trained deep learning models used in this work (those listed in Table A1) are standard token classifiers, text classifiers or machine translation models and not general purpose LLMs (in the now popular prompt-reply sense). I strongly recommend using the right technical terms (e.g. pre-trained sequence labeling models, deep learning classifiers, etc). It would make the paper even stronger if

dedicated NLP models are used for each subtask instead of merely "prompting" its way into the results with a chain of generative AI outputs. If de facto LLMs (the instruction-tuned chatbots) were indeed used, then all prompts should be documented in the appendix.

We agree with the reviewers' clarification of the correct terminology. We therefore used the terminology proposed by the reviewer.

We would like to point out that HazMiner does not rely on prompt-based chatbots to process text but on language models with a transformer-based architecture, such as BERT encoder models, that were fine-tuned for a specific task. We therefore removed any reference to the models used in HazMiner as large language models, to avoid confusion with prompt-based chatbots.

2. One of the most important aspects of a dataset paper is data provenance. Section 2.1 discusses text collection but there is no information on where exactly the data is coming from. What are the used URLs? What types of news outlets were included? Section 4.4.1 refers to Common Crawl and GDELT, but one cannot surely infer what news database was used in HazMiner.

From the comment of the reviewer, we understand that our manuscript did not clearly state the application of GDELT on the HazMiner methodology. The manuscript first introduces the HazMiner methodology in Section 2.1 until Section 2.4. Afterwards, this method is applied on news articles extracted from GDELT, as described in Section 2.6.

To improve readability, we clarified our full workflow at the end of the first paragraph of the methodology in L126-L127: "In the following sections, we further describe the main steps of the HazMiner methodology (Sect. 2.1 to 2.4), its evaluation (Sect. 2.5) and its application to GDELT (Sect. 2.6)"

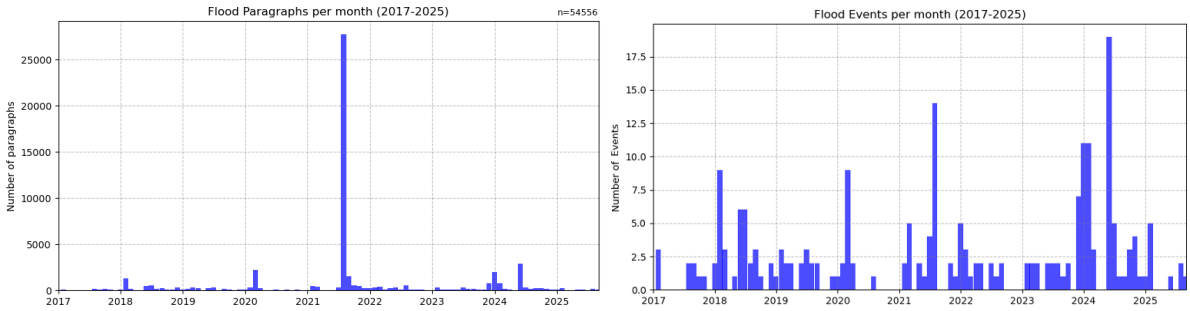
The use of GDELT resulted in large set of URLs. The URLs were consulted from the GDELT dataset based on a query containing a list of keywords (Table 1 of the manuscript). All URLs were made available in the article dataset on Zenodo. We clarified in the text where the reader can consult all the used URLs (L313-L314): "The list of URLs can be consulted in the final dataset (see in Sect. Code and data availability)".

3. The authors opted for working with paragraphs instead of texts and discuss some of the advantages like input length limitation of many encoders. But only in Section 4.3 the paper explains that the news text units were not preserved. If each paragraph is being counter as a separate observation, then the number of most events will be inflated, as arguably many news texts contain multiple paragraphs about the event

of interest. We have no guarantees that the larger number of identified events in comparison to existing databases is not merely due to the potentially exponentially duplicated pieces of information from each text. The text argues that "sharing information across paragraphs within the same article without an in-depth knowledge of the article's content comes with a risk", but the risk of not doing it is as critical. It is also common for the same news text from an agency press to be replicated by several news outlets, which further increase the number of duplicated information.

HazMiner presents two different datasets, these are the paragraph and event datasets. The paragraph dataset contains all the processed text at the paragraph level. We agree with the reviewer that the paragraph dataset is prone to duplicates, and each paragraph should not be considered as an individual event but rather an observation of an event. We elaborately explained this in L635-L641.

The event dataset are clustered paragraphs that occur in the same region over the same period. We believe that clustering the paragraphs accounts for most of the duplicate paragraphs, as they would be assigned the same location and timing and therefore end up in the same event. We would like to share an additional figure. We plotted the timeseries for paragraphs and events located in Belgium and Germany. As we can see, the floods that occurred on July 14th-15th represent around 32 000 paragraphs, receiving much more media attention compared to other events, with multiple duplicated paragraphs. Nevertheless, this imbalance in paragraphs is not represented in the event distribution.



4. The pipeline comprises a few arbitrary decisions not well justified in the paper and the presented evaluation of each subtask is limited. Timing and two types of impacts, for instance, were not evaluated at all. Even if the timing extraction algorithm is rule-based, it is also prone to error as any other model and must be evaluated. The performance of the paragraph and binary classifiers is suboptimal. The majority negative class is mostly correctly identified, but precision of the positive class remains low, which causes the final sample to contain many unrelated instances. This problem is intensified in a pipeline as errors are propagated from one step to the next.

We evaluated the time extraction algorithm of HazMiner by comparing it to the Global Flood Monitor (GFM) and the EM-DAT datasets. We tried to find spatiotemporal matches, the smallest spatiotemporal difference between the reference datasets and HazMiner, between the reference datasets and HazMiner. For each match the difference between the HazMiner time and reference time is calculated. At the event level both start and end of a reference event is compared to a HazMiner event. At the paragraphs level the start and end time was evaluated against the timing of the paragraph. The results can be consulted in Table 4 of the manuscript.

We agree that the precision for the classifiers is low. The precision for all evaluated classifiers is now reported in the appendix in Table B6. Nevertheless, we do not believe that errors are propagated over the algorithm. False positives generated by earlier models are mostly filtered out in later models. We explained this in the following paragraph manuscript (L392-L399): "The classification of hazard-related text, and the flood classification steps are characterised by a low precision. Nevertheless, the propagation of this low precision over HazMiner is limited due to sequential filtering; many false positives are filtered out at a later stage. For example, from the 750 paragraphs evaluated for the hazard-related text classification, 46 were falsely classified as hazard-related text (Fig. 3A). Of these 46 false positives, only six were subsequently classified as either flood, landslide or flash flood by the zero-shot hazard classification step. The other 40 false positive paragraphs were removed from the HazMiner dataset by the zero-shot hazard classification step. Similarly, 248 paragraphs, evaluated for zero-shot hazard classification, were falsely classified as floods (Fig. 3B). In the subsequent event-detection step, 110 out of these 248 false positives were correctly identified as non-events and excluded from the final HazMiner dataset."

5. Not all groups of news refer to ongoing events. Some events are so severe that newspapers continue to report on them after months or years (for instance, when a new judicial decision comes up). It seems that these cases were not handled and are always being counted as new events.

We want to thank the reviewer for this insightful comment. We added an additional filter to the HazMiner algorithm that filters out non-events.

We first annotated 200 paragraphs for them describing a hazard event or not. Non-event included paragraphs that issued a warning, contained general knowledge on geo-hydrological hazards or did not explicitly mention the hazard event but only its trigger or cause. For example, heavy rains were classified as non-event when they did not specifically mentioned floods.

Using the annotated data, we fine-tuned a sentence transformer model (sentence-transformers/all-MiniLM-L12-v2) using SetFit. During finetuning the model received 16 event paragraphs and 16 non-event paragraphs. After 2 epochs the model achieved an F1 score of 0.73, a recall of 0.73 and a precision of 0.73. Increasing the number of

epochs would lead to overfitting of the model. The confusion matrix of the fine-tuned model can be found in the appendix (Figure B1). We did not include this into the evaluation section, as we fine-tuned the model ourselves, compared to other models which we downloaded directly from the HuggingFace Hub.

We tested a total of five different models, in the table below you can find the F1-score, recall and precision for each tested model.

Model	*F1-score*	*Recall*	*Precision*
sentence-transformers/all-MiniLM-L12-v2	*0.73*	*0.73*	*0.73*
sentence-transformers/paraphrase-mpnet-base-v2	*0.70*	*0.70*	*0.70*
sentence-transformers/all-MiniLM-L6-v2	*0.71*	*0.71*	*0.71*
sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2	*0.64*	*0.64*	*0.65*
sentence-transformers/all-roberta-large-v1	*0.66*	*0.66*	*0.68*

We added one paragraph to Section 2.2 describing the non-event filter (L168-L175): "Lastly, non-events are filtered out through few-shot classification (Li Fei-Fei et al., 2006). Non-events are paragraphs reporting early warnings or describing only the trigger of a hazard without the hazard event occurring. Like zero-shot classification, few-shot classification assigns classes to text based on the model's embedding space. But few-shot classification, compared to zero-shot classification, is additionally fine-tuned on a limited number of annotated texts. For this, we fine-tuned the "sentence-transformers/all-MiniLM-L12-v2" sentence transformer model (Tunstall et al., 2022; Wang et al., 2020) on 16 positive and 16 negative annotated paragraphs through two epochs using Sentence Transformer Fine-tuning (SetFit – Tunstall et al., 2022). The new fine-tuned model, "ValkB/all-MiniLM-L12-v2_HazMiner_nonevents", filters out non-events with an F1-score of 0.73. The confusion matrix can be consulted in the appendix (Fig. B1)."

The non-event filter significantly reduced the number of flood events from 21,411 events to 15,313 events. The number of landslides and flash floods events, for which generally are less warning in the news, only reduced minorly. Landslide events reduced from 7,659 events to 6,836 events and flash floods events from 3,606 events to 2,258 events.

6. GDELT is referenced many times in the text. What makes the HazMiner dataset different from GDELT? A section dedicated to comparing the advantages and disadvantages of existing datasets and HazMiner with an overview illustration or table would be useful. Related to this point, the paper is lacking a dedicated related

work section. Currently, some related work is discussed in the first section, and some in the limitations and biases section.

GDELT is a dataset that tries to collect global events from news, and is not limited to geo-hydrological hazard events. Therefore, their definition of events is very broad. We only used GDELT to build our text corpus. The added value of HazMiner compared to existing global datasets in geo-hydrological hazards is explained in Sect. 4.1 and 4.2.

7. For impact extraction, there is no clear information on how the impact from different paragraphs and sources was aggregated. It is common for news outlets to report preliminary estimates that are updated that as the event unfolds. How was this handled?

Section 2.3.3 informs the reader on the impact extraction at the paragraph level. This method tries to extract numbers from text based on a Q&A algorithm. Section 2.4, more specifically, L267-L271 informs the reader on impact aggregation at the event level: "The events' impact is estimated by the median of the paragraphs' impact, for each impact category, namely the number of fatalities, homeless people, injured people, affected people, missing people, and evacuated people. Before calculating the median, two filters are applied: (1) removing outliers based on 1.5 times the interquartile range, and (2) removing impact numbers that are reported fewer than five times".

8. The Global South bias in existing datasets is mentioned as a drawback many times in the text. Still, line 315 mentions that EM-DAT concentrates on this area. Besides, the bias was not alleviated: HazMiner still contains many more observations about the Global North, as mentioned in Section 3.2.

We agree with the reviewer that there are still biases present in the HazMiner dataset, especially towards the Global South, as discussed in Section 4.4.

Nevertheless, the goal of HazMiner is not to reduce biases at the global scale. HazMiner was developed to increase the amount of information on hazards, especially illustrating the potential and limitations of using text as a source for hazard documentation.

We adapted our terminology from the Global South to the country-level income classification of the World Bank, hereby the Global South is redefined as lower-middle-income and low-income classes. Furthermore, we added Table 2 and Table 3 to the manuscript. Table 2 illustrates the number of HazMiner events and its relative variation in event numbers to EM-DAT, GFM and GLC per hazard class and per income class. While Table 3 shows the percentage of extreme events in HazMiner and EM-DAT, and their difference in share.

We further clarified this in the discussion as follows (L570-L575): "Our study indicates that these inclusion criteria of EM-DAT result in significant data gaps in low- and lower-middle-income countries. HazMiner records a substantially larger number of less severe events that are absent from EM-DAT (Tables 2 and B2) thereby providing a more comprehensive documentation of geo-hydrological hazards in the most vulnerable regions. In high-income countries, HazMiner and EM-DAT have a similar share in events with more than 10 fatalities, one of EM-DAT's inclusion criteria (Table 3); meaning that low-fatality events in high-income countries are also documented in EM-DAT due to the other two thresholds."

Clarifications / corrections needed with some suggestions

Abstract

- l1: Since the number of pages is not critically limited, the abbreviation GH could be avoided throughout the whole text. The full version "geo-hydrological hazards" does not take that much extra space and really facilitates reading.

We agree with the reviewer that this facilitates the reading process. We adapted the manuscript with the full version.

- l15: What is data-scared?

We thank the reviewer for noticing this spelling mistake. It was supposed to be "data-scarce". This typo was corrected in the reviewed version.

- l18: The 58 languages should be listed in the appendix.

We agree with the reviewers' suggestion, the full list of the languages was added to the manuscript as appendix A.

- l20: It does not really advance the text mining field itself as no new techniques or findings related to the text mining methods themselves are presented.

We agree with the comment of the reviewer, we did not develop any new methods in the field of text mining. We did develop a new method that advances the accessibility and use of text mined natural hazard data. We clarified this in our text through the following line (L22-L23): "This new versatile multilingual method and its dataset advance the accessibility and use of text mined natural hazard data."

Introduction

- Almost 2.5 pages is quite long for an introduction. It mentions many related works that deserve a dedicated related work section. I suggest splitting this part into two sections, one

with a compact overview of the paper and the other for the list of related work, ideally extended to account for the various tasks in the pipeline.

We want to thank the reviewer for the suggestion. We restructured the introduction of the manuscript so that it fits better with its story line. We did not add a section describing related work as this would increase the length of the manuscript.

- l63: Text mining methods are not perfect and cannot always extract the right information. It is important to phrase it in a way that makes it clear that these techniques do not always work.

We adapted the introduction so that the different advantages and disadvantages of different text mining methods are explained. These include the following lines (L88-L95): "Despite these rapid developments in computational linguistics, important research gaps remain in natural hazard text mining. One of them is the underuse of encoder-based models, such as BERT, compared to prompt-based text processing with LLMs such as GPT. The use of text generating LLMs through prompt engineering has gained popularity because they do not require annotated data. They perform NLP tasks, such as text classification, NER, or Q&A, without task-specific fine-tuning (Mayo et al., 2026). However, these non-text generating NLP tasks, do not necessarily require text generating LLMs and can also be performed more efficiently by much smaller fine-tuned encoder models, such as BERT (Rogers et al., 2020). BERT, compared to LLMs, can be run locally on commonly available computers without sharing copyrighted text data through Application Programming Interfaces (APIs)."

- l65: And disadvantages.

We thank the reviewer for the suggestion. We added this to the sentence.

- l74: LLMs do not "understand" texts. I recommend avoiding verbs related to human cognition to refer to how LLMs process language.

We thank the reviewer for the clarification. We removed the word "understand" from the sentence.

- l75-78: LLMs are mentioned before GPT and BERT, which makes it sound as if these two appeared after the popularization of LLMs. I suggest reversing the order of these sentences. First, refer to the advent of Transformer-based architectures like GPT and BERT, then to the popularization of LLMs as a tool for many NLP tasks.

We agree with the reviewer. We adapted the manuscript accordingly, through the following lines (L81-L86): "These steps can be performed using various NLP tasks, including text classification, named entity recognition (NER), Question and Answer (Q&A), text generation or translation (Avcıoğlu et al., 2025; Fu et al., 2022; Rogers et al., 2023). The performance of these tasks significantly improved with the introduction of transformer-based language models (Vaswani et al., 2017), such as BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) or with the introduction of large language models (LLMs), such as GPT (Generative Pre-trained Transformer) (Radford and Narasimhan, 2018)."

- l82: The "While only a few..." needs to be connected to another sentence.

We thank the reviewer for this comment. While revising and restructuring the manuscript, this line was removed and no longer appears in the revised version.

- l94: It is ambiguous whether "the dataset" refers to GDELT or your dataset.

We understand that this is not clear to the reader. This line referred to the HazMiner dataset, introducing its structure of event and paragraph dataset. After reviewing the manuscript, we removed this line as we found that introducing the structure of the HazMiner dataset was more appropriate in the methodology.

- l96: The "Followed by..." needs to be connected to another sentence.

We thank the reviewer for this comment. While revising and restructuring the manuscript, this line was removed and no longer appears in the revised version.

Methodology

- It is a matter of personal taste, but since Figure 1 already shows an overview of the pipeline, it would be better if the description of each subtask was already accompanied by its own evaluation. It is hard to read and move on without knowing how well each model performs at each step. Then Section 3 could focus only on describing the resulting dataset.

We thank the reviewer for the suggestion; we did not adapt the manuscript to their suggestion. We wanted to make a clear distinction between the method, what we developed, and the results. For example, the new non-event filter was the only language model that was fine-tuned by us, therefore we referred the evaluation metrics of this model in the methodology.

Nevertheless, the other language models were not developed by us. Therefore, we see their evaluation as part of the result of our manuscript.

- l109: What does it mean that HazMiner was "specifically" configured to extract GH events, isn't it its main purpose?

HazMiner is designed to be applied on natural hazards and not only on geo-hydrological hazards. This was briefly mentioned in the introduction in L105-L110 as: "Given these research and knowledge limitations, we develop HazMiner, a versatile multilingual text mining method designed for application from the regional to the global scale. It leverages the full extent of textual sources to extract natural hazard event-related information such as location, timing and impact (Sect. 2). Within this research, we use HazMiner to generate a global dataset on geo-hydrological hazards by applying it to the Global Dataset on Events, Language and Tone (GDELT – The GDELT Project, 2025), with the specific aim to improve their documentation, especially for regions that are typically under-reported such as low- and lower-middle-income countries (Sect. 3)".

We now understand that this is not clear for a reader to understand these nuances. In the first paragraph of the methodology, we further clarified (L121-L124) by adding: "Whereas HazMiner is designed to document natural hazards from text sources, in this work, we specifically configured it to extract geo-hydrological hazard events."

- Figure 2: "hazardous text" sounds like the text itself is hazardous.

We changed this into hazard-related text on the figures and in the text of the manuscript.

- Figure 2: An arrow connecting the output of a step to the input of the other would be useful to avoid the reliance on color, which may not be visible for color-blind readers.

We understand the concern of the reviewer. All figures were designed to be inclusive towards colourblind persons. We believe that adding arrows to the figure might unnecessarily increase its complexity.

- 2.1 should be named "text selection" or "document selection", as it involves steps that are not about extraction.

With text extraction we meant the parsing of the html's of the corresponding URLs. The text selection step is handled through the GDELT, not by HazMiner. The GDELT API requires a list of keywords for which it will return all relevant URLs within the GDELT dataset. Thereby it is the users his task to select the text corpus as HazMiner does not provide any web crawling algorithm.

- l115: Subtasks rather than suprocesses.

We want to thank the reviewer for this suggestion. We changed sub-processes into subtasks

- l116: Scrapping news is usually not allowed by media outlet websites. Was it performed in accordance with the terms of use?

This is addressed in the disclaimer of the manuscript. We only automatically extracted data from publicly available websites. The HazMiner dataset was created through lawful text and data mining (TDM) in accordance with Article 3 of Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market.

- l119: The simple heuristic to use backspace to identify paragraphs can be problematic, as there is no guarantee that all websites follow this rule and this information may be lost if the raw texts were already altered by news aggregators. This step should thus also be evaluated.

We understand that there is a diversity formats to illustrate a paragraph. When working with large number of data sources it is difficult to adapt to every format. For the html parsing we used the newspapers3k python package which returns a "\n\n" for each html paragraph annotation <p>.

However, we do understand that not all websites follow this annotation. We did an additional analysis on 178,025 news articles in the HazMiner dataset. From these 174,137 articles successfully split into 3,544,364 paragraphs.

- l121: It is not really more efficient because it makes the number of instances to process much larger and removes important context needed for disambiguation.

We agree with the reviewer. We removed this passage from the manuscript.

- l122: It's not true that current LLMs are optimized for shorter inputs (unless it's not actually LLMs but text encoders, see comment above). Not fewer parameters but fewer computations.

We agree with the reviewer that current LLMs can process much longer text. To avoid confusion with the above reviewer's comment, we removed this passage from the manuscript.

- l150: I don't know how these specific embeddings were optimized. But generally, embeddings are trained to capture contextual relations, it is expected that the representations of all these concepts will be close in the embedding space.

We acknowledge that zero-shot classification is not the best method to classify the paragraphs into hazard types, as discussed in L707-L711. Nevertheless, no language

models supervised to classify floods, landslide and flash floods were available on HuggingFace. Developing these would be out of the scope of this research.

- l161: Publication date is not very reliable as the timing of the hazard because many important events resurface in the news much later, and some are delayed.

We understand the reviewers concern that publication date does not always reflect the timing of the hazard itself. Limiting to only time mentions in the text would significantly reduce the amount of available information for event creation.

The non-event filter, described in specific comment 5, should remove these resurfacing news events. Furthermore, we included an uncertainty calculation at the event level, see L271-L272: "Lastly, for each event an uncertainty of the abovementioned parameters, in location, time and impact, is reported through a bootstrap analysis with 100 cycles". Resurfacing news within an event should result in large uncertainties.

- l170: It's unlikely that a rule-based approach really solves this problem. There are many news about studies that refer to events as a general category, or about events that occur in movies or books, which are not always in future tense.

As described in specific comment 5, we added a filter to remove paragraphs that do not describe a hazard event.

- l178: This assigns coordinates to all locations mentioned in the text, but how is the event location detected among them? Normally, there is a third step.

HazMiner does not include a step which relates the extracted place names to the event. Large language models would be able to extract this context, but due to computational limits it is not possible to apply open-source large language models on 6 million news articles.

- l180: Hallucinations is a common (although not very suitable) term for wrong content generated by text completers / chatbots. It seems that, here, it refers to normal translation errors instead.

We agree with the reviewer, we removed any reference to chatbot hallucinations in this context.

- l196: Weighted average may likely end up in a location that is not even in the text, maybe even in a different country.

We agree with the reviewer that computing a weighted average from all locations in a paragraph reduces the location accuracy. As stated in L202, each paragraph also

contains a bounding box, which is the smallest rectangle surrounding all locations of the paragraph. If the location of the paragraph would move to a different country, the bounding box of the paragraph would still include the original country. Furthermore, we included filters that avoid shifting of a paragraph location to an unrelated location due to place name ambiguities (see L205).

- l220: The sentence "As hazard events..." is not finished.

We changed the sentence into (L227-L231): "Specifically, news articles have a spatially strong hierarchical structure, due to the uneven global distribution of media coverage on hazard events (Berlemann and Thomas, 2019; Walter et al., 2016), as hazard events in high-income countries receive more media attention compared to those in low and lower-middle-income countries (Eisensee and Stromberg, 2007; Murthy and Longwell, 2013; Porter and Evans, 2020; Solman and Henderson, 2019)"

- l262: It is unclear what "until the number of false positives remained constant" means here.

We agree with the reviewer that this might confuse the reader. We mean that the false positive rate remains constant with increasing number of validated paragraphs. We clarified this in our text (L281-L282): "They are manually annotated until the false-positive rate converged and remained constant with increasing annotation sample size."

- l290: The fact that the timing algorithm does not rely on neural network-based models does not mean it works well. It's a rule-based approach that will not always work.

As mentioned in our answer to specific comment 4, we evaluated the time extraction algorithm of HazMiner by comparing it to the Global Flood Monitor (GFM) and the EM-DAT dataset. This is done by finding spatiotemporal matches between the reference datasets and HazMiner. For each match the difference between the HazMiner time and reference time is calculated. The results can be found in Table 4.

Results

- For classification, traditional metrics like binary precision, recall and F1 scores should be reported in a Table. Only recall (true positive rate) was reported but, in such tasks, high precision is as (or even more) important, as "inventing" events that never happened can be a more serious mistake than missing real events depending on the use case.

We agree with the reviewer's comment. We added table B6 to the appendix containing F1 score, precision, recall and false positive rate for the evaluated processing steps.

- l345: Precision is quite low, though (only around 0.6 based on the confusion matrix). So the performance is mostly coming from identifying the vast majority of unrelated instances.

We acknowledge that the precision of the model is low, we reported the precision from our evaluation in table B6. The following metrics were reported on HuggingFace (https://huggingface.co/DunnBC22/distilbert-base-uncased-Tweet_About_Disaster_Or_Not):

- ***Loss: 0.2557***
- ***Accuracy: 0.9138***
- ***F1: 0.7752***
- ***Recall: 0.8204***
- ***Precision: 0.7348***

There is a need for more fine-tuned text classification models with a higher precision. Especially (fine-tuned) multilingual encoder models remain a significant gap. Training or fine-tuning these models was out of scope of our research.

- l364: How exactly are the errors computed? Are the bounding boxes considered or only the centroid?

We did this by calculating the difference between the point reported in the Global Landslide Catalog and the HazMiner point (not bounding box) using the Haversine function. This is mentioned in L296-L298

- l380: Very often, news report on preliminary number that keep being updated as the event unfolds. How was this handled? If multiple paragraphs from different articles are reporting about the same event, how are their estimates harmonized?

We acknowledge that impact reporting from media evolves over time. As mentioned in our answer to specific comment number 7, the paragraphs impact of one event is aggregated based on the median impact number reported. During the development of HazMiner, we ran different test for comparing different aggregation strategies; these include: the most frequently reported number, the maximum number, and the median number.

As illustrated by the below figure, the difference between the reported impact of EM-DAT and HazMiner was similar for each aggregator strategy. The "maximum" seemed to be the worse match, while the median seems the best match. Therefore, we choose for the median as aggregator for the impact.

Furthermore, we believe that the corresponding paragraphs of an event will help to understand the evolving nature of impact reporting. As for each event you can study the temporal distribution of the reported impact at the paragraph level. Whereas

traditional datasets report one impact number, HazMiner allows for a detailed temporal analysis of impact reporting per event.

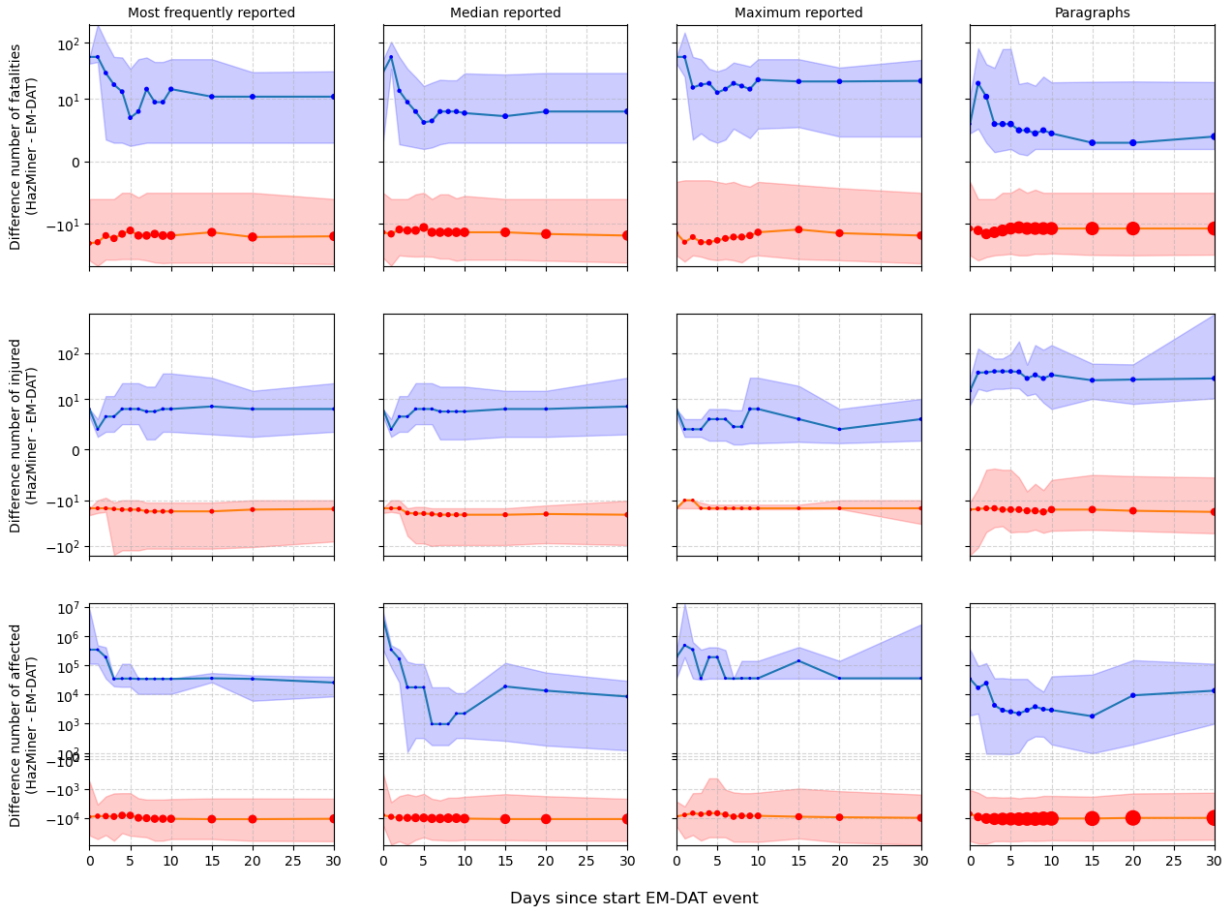


Figure 1: For each aggregation strategy (event level) and the paragraph level the difference between the impact (fatalities, injured and affected) number reported by EM-DAT and the one reported by HazMiner for different days since the start of the EM-DAT event. Blue represents positive differences, where HazMiner reports a larger number of impact, and red represent negative differences, where HazMiner reports fewer numbers. Lines represent the median difference while coloured zones represent the data between the 25-quantile and 75-quantile. Larger points represent more datapoints being matched.

- l393: The mean duration is not of the GH itself but of its news coverage?

We agree with the reviewer's clarification. We adapted this in the text by stating that a HazMiner geo-hydrological event that reflects news reporting (L453-L454): "At the global scale, a geo-hydrological hazard event extracted from news reporting is characterised by a mean duration of two days and an average size, i.e., the bounding box containing all paragraphs per event, of 1,929 km² and median of 0.009 km²."

- l468: "one death but fewer than 10 fatalities" sounds like a death is not a fatality. Maybe rephrase.

We rephrased this into (L527): “at least one death but fewer than ten.”

Discussion

- l509: with the goal

We thank the reviewer for this comment. While revising and restructuring the manuscript, this line was removed and no longer appears in the revised version.

- l520: The sentence "Mostly because..." should be connected to another sentence.

We thank the reviewer for this comment. While revising and restructuring the manuscript, this line was removed and no longer appears in the revised version.

- l524: The term "only" is not correct here as it can include events with less deaths if they affect enough people.

We agree that our statement was incomplete. We therefore changed it into (L568-L570): “In the case of EM-DAT, the dataset includes disasters with more than 10 fatalities or at least 100 affected people or events with a call for international assistance or an official emergency declaration (Delforge et al., 2025)”

- l545: The sentence "As studies..." should be connected to another sentence.

To improve the readability of the text we will adapt it to (L587-589): “At the regional scale, the potential of text mining for hazard documentation is also important, as studies focusing on regional text often deliver the most detailed information in large quantities (Calvello and Pecoraro, 2018; Domínguez Cuesta et al., 1999; Franceschini et al., 2022; Valenzuela et al., 2017).”

- l574: country's

We thank the reviewer for noticing this spelling mistake, it will be corrected in the reviewed version.

- l643: It's not the language that is non-alphabetical, it's the writing system (see https://en.wikipedia.org/wiki/List_of_writing_systems).

We agree with the reviewer on this distinction. We adapted the manuscript referencing Vietnamese as an alphabetic writing system (L681-L684): “For example, Vietnam, which mainly supports Vietnamese - an alphabet-based writing system - for

its documentation, contains remarkably more events compared to other Southeast Asian countries (Fig. 7 and 8).”

- l671: The sentence "Such as news..." should be connected to another sentence.

We thank the reviewer for this comment. While revising and restructuring the manuscript, this line was removed and no longer appears in the revised version.

- l674: into the (or rather "as input to the"?)

We would like to thank the reviewer for noticing this spelling mistake, it was corrected in the reviewed version of the manuscript.

- l716: Switching models is likely not that easy, and each may use different labels, output format, programming libraries or require a different input structure.

HazMiner uses the HuggingFace pipeline to load and use the language models. To our experience, fine-tuned models could easily be exchanged when loaded through the pipeline, if the models are fine-tuned for similar tasks.

- l725: Subconsciously is not a suitable term since model/study are the agents in this sentence.

Thank the reviewer for noticing this grammatical mistake, we changed “subconsciously or not” into “explicitly or implicitly”.

Response letter: referee comment 2

We would like to thank the reviewer for the careful and critical evaluation of our manuscript. We appreciate the reviewer's detailed assessment of the HazMiner methodology, and the emphasis placed on validation and robustness. The reviewers' comments helped us to better clarify the scope of HazMiner, to refine statements comparing HazMiner with existing datasets, and to strengthen the discussion of methodological limitations, uncertainty, and data quality.

Below, we address each comment in detail. Our answers to comments in italic bold.
Comments by the Anonymous Referee are in normal style

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.

We want to thank the reviewer for his critical view on the methodology. We tried to address their comments (1) including an evaluation of the time extraction algorithm, (2) adding a filter which removes non-events and (3) reporting uncertainties of event characteristics based on a bootstrap analysis.

First, we evaluated the time extraction algorithm of HazMiner by comparing it to the Global Flood Monitor (GFM) and the EM-DAT dataset. We tried to find spatiotemporal matches, the smallest spatiotemporal difference between the reference datasets and HazMiner, between the reference datasets and HazMiner. For each match, the difference between the HazMiner time and reference time is calculated. At the event level both start and end of a reference event is compared to a HazMiner event. At the paragraphs level the start and end time was evaluated against the timing of the paragraph. The results can be consulted in Table 4 of the manuscript.

Second, we added an additional filter to the HazMiner algorithm that filters out non-events. We first annotated 200 paragraphs for them describing a hazard event or not. Non-events included paragraphs that issued a warning, contained general knowledge on geo-hydrological hazards or did not explicitly mention the hazard event but only its trigger or cause. For example, heavy rains were classified as non-event when they did not specifically mentioned floods. Using the annotated data, we fine-tuned a sentence transformer model (sentence-transformers/all-MiniLM-L12-v2) using SetFit. During finetuning the model received 16 event paragraphs and 16 non-event paragraphs. After 2 epochs the model achieved an F1 score of 0.73, a recall of 0.73 and a precision of 0.73. Increasing the number of epochs would lead to overfitting of the model. The confusion matrix of the fine-tuned model can be found in the appendix (Figure B1).

Third, we included an uncertainty calculation at the event level, see L271-L272: "Lastly, for each event an uncertainty of the abovementioned parameters, in location, time and impact, is reported through a bootstrap analysis with 100 cycles".

We agree that the precision for the classifiers is low. The precision for all evaluated classifiers is now reported in the appendix in Table B6. Nevertheless, we do not believe that errors are propagated over the algorithm. We clarified this by adding the following paragraph to the manuscript (L392-L399): "The classification of hazard-related text, and the flood classification steps are characterised by a low precision. Nevertheless, the propagation of this low precision over HazMiner is limited due to sequential filtering; many false positives are filtered out at a later stage. For example, from the 750 paragraphs evaluated for the hazard-related text classification, 46 were falsely classified as hazard-related text (Fig. 3A). Of these 46 false positives, only six were subsequently classified as either flood, landslide or flash flood by the zero-shot hazard classification step. The other 40 false positive paragraphs were removed from the HazMiner dataset by the zero-shot hazard classification step. Similarly, 248 paragraphs, evaluated for zero-shot hazard classification, were falsely classified as floods (Fig. 3B). In the subsequent event-detection step, 110 out of these 248 false positives were correctly identified as non-events and excluded from the final HazMiner dataset."

Lastly, we would like to clarify some decisions made.

- ***Through the comparison of HazMiner with EM-DAT we wanted to illustrate that much more information on hazard events is available in online news articles and can be combined with automatic text mining approaches. Furthermore, we did not state that the comparison was impossible. We merely wanted to draw the readers' attention to the decisions we imposed in the comparison of impact data (L301-307).***

- ***Translation was performed by the "Helsinki-NLP/opus-mt-mul-en" containing around 77 million parameters. This language model is relatively small compared to large language models such as the GPT, DeepSeek or QWEN models, which contain several hundreds of millions of parameters, hence with fewer translation errors. We do believe that running those models is not feasible for the limited infrastructure available in our research team. The translation model used in HazMiner took six months of processing alone. Furthermore, to minimalize translation errors, we used a multilingual language model for location extraction, as they avoid translation and therefore the propagation of translation errors into the geoparsing algorithm. Yet this was not possible for other steps in HazMiner, as there is a need for multilingual models fine-tuned for text classification and information extraction. Training or fine-tuning these models was out of scope of our research.***

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

We agree with the reviewer that a sensitivity analysis would better inform the reader. Therefore, we performed a sensitivity analysis (L242-L245): "The sensitivity of three parameters—xi-value, minimum number of paragraphs, and maximum epsilon—was evaluated with respect to the number of events, the average number of paragraphs per event, and the average duration. Among these parameters, the minimum number of paragraphs showed the largest influence on the results, whereas the xi-value and maximum epsilon had only a minor effect (Fig. B2)"

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

We want to thank the reviewer for his comment. We added the following citation: Gill, J. C. and Malamud, B. D.: Reviewing and visualizing the interactions of natural hazards: Interactions of Natural Hazards, Rev. Geophys., 52, 680–722, https://doi.org/10.1002/2013RG000445, 2014.

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

We cited the following manuscripts:

Avcıoğlu, A., Demir, O., and Görüm, T.: An automated approach for developing geohazard inventories using news: integrating natural language processing (NLP), machine learning, and mapping, Nat. Hazards Earth Syst. Sci., 25, 2421–2435, https://doi.org/10.5194/nhess-25-2421-2025, 2025.

Eisensee, T. and Stromberg, D.: News Droughts, News Floods, and U. S. Disaster Relief, The Quarterly Journal of Economics, 122, 693–728, https://doi.org/10.1162/qjec.122.2.693, 2007.

Madruga De Brito, M., Sodoge, J., Kreibich, H., and Kuhlicke, C.: Comprehensive Assessment of Flood Socioeconomic Impacts Through Text-Mining, Water Resources Research, 61, e2024WR037813, https://doi.org/10.1029/2024WR037813, 2025.

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.

We thank the reviewer for the clarification on this terminology. We changed the terminology from real-world to in-situ observations to make a clear distinction to remote sensing observations.

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.

We want to thank the reviewer for noticing that our manuscript did not clearly state where a reader can find the URLs used to generate the HazMiner dataset. URLs were consulted from the GDELT dataset, based on a structured query (Table 1 of the manuscript). This resulted in large set of URLs. All URLs were made available in the article dataset on Zenodo. We clarified in the text where the reader can consult all the used URLs, by adding (L313-L314): **"The list of URLs can be consulted in the final dataset (see in Sect. Code and data availability)".**

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.

These two instances were found during validation. Misclassifications are a limitation of zero-shot classification, using a different zero-shot language model with a different word embedding space would result in similar instances. We specifically chose zero-shot classification as this allows HazMiner to be used to detect other hazards than geo-hydrological hazards without the need of supervised classification. The zero-shot classification model could be replaced by a language models supervised to classify floods, landslide and flash floods. Nevertheless, no such model was available on HuggingFace. Developing these would be out of the scope of my research.

By adding the non-event filter to the HazMiner method we further reduced the effect paragraphs describing only heavy rainfall being misclassified as floods, as explained in our answer to the major comment.

Figure 2: The double legend on the maps is misleading

We do understand that the legend might be difficult to interpret at the first sight. We changed it to a single legend label: "Share of HazMiner floods" and as illustration referenced tick 0 as "Only EM-DAT floods" and tick 100 as "Only HazMiner floods".