Extracting the 'Why': Flood Cause Annotation and LLM Benchmarking from Multi-Platform Crisis Data

Jane Arleth dela Cruz*
Center for Language and Speech Technology
Center for Language Studies
Radboud University, Nijmegen, Netherlands
janearleth.delacruz@ru.nl

Aleksei Asimov
FloodTags
The Hague, Netherlands
a.asimov@floodtags.com

Jurjen Wagemaker
FloodTags
The Hague, Netherlands
wagemaker@floodtags.com

Iris Hendrickx
Center for Language and Speech Technology
Center for Language Studies
Radboud University, Nijmegen, Netherlands
iris.hendrickx@ru.nl

ABSTRACT

We present a framework for extracting and classifying flood cause information from crisis-related online media data. By harnessing data across multiple online platforms from both social media and online news sources, we can understand causes of floods ranging from environment factors like heavy rainfall to human-related factors like infrastructure failure. This task is operationally critical not only for disaster response but also for public accountability, yet it remains a challenging task for automated systems. Our key contributions are: (1) a comprehensive flood cause annotation protocol that operates at two levels of granularity i.e., document-level and sentence-level, with strict exclusion criteria for distinguishing causal statements from impacts; and (2) initial benchmarking experiments to evaluate the performance of zero-shot large language models (LLMs) on the classification task. Although the underlying corpus—covering 70 distinct flood events from 2022–2025 across News, X (formerly Twitter), YouTube, and Bluesky is not publicly released, we describe the data construction methodology in full to support replication on comparable data. Our initial analysis on a stratified 10-event sample demonstrate that LLMs struggle: they have high recall but suffer from low precision in flood cause classification.

Keywords

flood cause extraction, annotation protocol, causality classification, crisis informatics, large language models (LLM)

INTRODUCTION

During flooding events, online media streams are filled with posts containing flood-related information. While reports of impacts (e.g.,"thousands evacuated," "people trapped") are frequently reported, the causes of these events (e.g., "blocked drainage","dam failure," "overflowing rivers") can be buried in the noise and irrelevant content. Causality extraction is a subtle but critical task for disaster risk management (DRM). Across the DRM cycle, this type of analysis is very useful: in the immediate response phase, to triage localized hazards, and in the recovery and mitigation phases, for downstream accountability attribution and risk analysis.

Information extraction studies for disaster risk management are heavily focused on situational awareness (Fu et al. 2022; Chowdhury et al. 2024), impact classification (Imran, Ziaullah, et al. 2025; McDaniel et al. 2024; Dela Cruz et al. 2025) , and rescue prioritization (Koju et al. 2024). Hence, existing crisis datasets focus broadly

*corresponding author

on classification tasks in 'relevance' or 'informativeness', and in humanitarian information type (Alam et al. 2021; Olteanu et al. 2014; Imran, Mitra, et al. 2016). In contrast, causality extraction which is an essential task for post-event accountability and mitigation planning, remains underexplored, with a few exceptions like (Duong et al. 2025; Peng and Zhu 2025), both utilizing large-language models (LLMs) for feature extraction. Our main focus is causality extraction, distinguishing causal statements from all relevant information from the crisis events. By extracting causal statements, decision-makers can cross-reference social media observations and news reporting, with official data to form a more complete operational picture. This can aid in the creation of comprehensive post-event reports for local governments and global institutions. We contribute a reusable annotation protocol and a set of benchmarking results that together support the development and evaluation of flood cause extraction systems.

Our key contributions are as follows:

- **Comprehensive Flood Annotation Protocol:** a framework designed to distinguish flood causes from flood types and impacts, and operated at two levels of granularity – document-level (flood cause or no flood cause) and sentence-level (flood cause or no flood cause, and cause categorization)
- **Benchmarking LLM Capabilities:** we evaluate the zero-shot performance of commonly used LLMs and demonstrate the dataset serves as a benchmark for LLMs' capabilities in high-stakes scenarios.

We establish initial baselines on commonly used off-the shelf LLMs, GPT 4o-mini (OpenAI 2024), Llama 3.1 8B-Instruct (Llama Team 2024) and Mistral NeMo (Mistral AI Team 2024) in zero-shot settings on a stratified 10-event subset (615 documents; 967 sentences). The results indicate that while models achieve very high recall, precision remains to be poor. This is especially apparent in the sentence-level, suggesting that LLMs frequently include non-cause information like generic flood information i.e., flood types and more commonly impacts.

Although the underlying multi-platform corpus covering 70 flood events from 2022–2025 is not publicly shared, we provide a full description of the data collection and filtering methodology so that researchers can apply the same pipeline to comparable data sources.

RELATED WORK

Crisis and Flood Datasets. Existing crisis informatics datasets focus broadly on relevance detection, informativeness classification, and humanitarian information classification. For example, CrisisBench (Alam et al. 2021) comprises multiple Twitter-based crisis datasets (e.g CrisisLex (Olteanu et al. 2014) and (Imran, Mitra, et al. 2016) into a unified benchmark for classification for informativeness and humanitarian aid information type. Wiegmann et al. (2020) perform a similar Twitter-based crisis dataset consolidation. Flood-specific instances are represented within these multi-hazard datasets, however, labels are focused on the describing informativeness and general impact (e.g. affected people, infrastructure damage) rather than explicit causal statements. As a result, existing crisis datasets emphasize on *'what is happening'* or *'what has happened'*, and not *'why it happened'*. To the best of our knowledge, there is currently no openly documented annotation protocol or benchmark specifically designed for flood cause extraction and categorization across heterogeneous platforms.

Annotation Protocol and Causality Extraction. Information extraction from crisis-related texts requires structured annotation frameworks. Feng et al. (2022)'s review on extraction of geographic information from social media noted that extracting underlying disaster triggers remains challenging. There have been studies developing targeted information extraction frameworks, such as Ye et al. (2024) proposing a disaster state information extraction framework for Chinese texts during typhoons, and Ma et al. (2023) introduce an ontology-based BERT model for automated entity and relation extraction from geological hazard reports. However, these frameworks embed causal information within broader state or hazard labels rather than treating causality as a stand-alone target. Our annotation protocol addresses this gap by defining operationally driven guidelines that identify causality from subsequent impact or other tangentially-related information.

LLMs in Disaster Management. There have been surveys highlighting the applications of LLMs across the disaster management cycle (Xu et al. 2025; Lei et al. 2025). Existing LLM benchmarks in disaster management tend to focus on broad classification or generation tasks, especially focused on producing situational reports or classifying messages into humanitarian categories. LLMs have demonstrated capabilities in extracting actionable information both through few-shot and zero-shot techniques (Xu et al. 2025; McDaniel et al. 2024; Lei et al. 2025). Qian et al. (2024) utilized ChatGPT to classify precipitation-related keywords from social media data into rainfall, flood or other related terms. Koju et al. (2024) demonstrate LLMs' capability in rescue prioritization classification. These studies, therefore, focus on situational awareness and impacts, not on causality extraction. We

attempt to evaluate LLMs on the narrow, high-precision reasoning task of separating the causal statements from the impact-related text. Our work complements the emerging LLM research in disaster management by providing initial zero-shot baselines for the specific task of flood cause extraction, laying the groundwork for future work on more fine-grained causal categorization and summarization.

DATASET CONSTRUCTION

We describe the full construction pipeline of the dataset below so that the methodology is transparent and reproducible on similarly sourced data.

Data Source

We curated documents from four distinct online platforms. Table 1 shows what a document is defined as for each online media platform. All these different platforms offer unique linguistic characteristics. Furthermore, we note that these sources have a difference in credibility of information. For example, social media posts can be subjective, may contain speculation, and unverified causality; however, they offer complementary information to official sources.

Platform	Document	Linguistic Characteristics
News	News Article	formal style, structured, and longer form
X (formerly Twitter)	Individual Post	short text, real-time, unstructured, informal, and user-generated
Bluesky	Individual Post	short text, real-time, unstructured, informal, and user-generated
YouTube	Video Caption	short, descriptive text, transcript of video reports, may include spoken-language features, user-generated

Table 1. Data Source Description. We describe the what each document means for each platform

Data Collection: Events

The flooding event identification was based on a proprietary retrieval algorithm from FloodTags (FloodTags n.d.). Flooding events were selected from a wide representative range of document counts. The key criteria for initial flood event selection are that (1) flood event have English-language documents and (2) flood event have documents with initial flood event cause keywords. These keywords were extracted based on extensive domain-specific experience. This led to the selection 70 flooding events from 2022-2025.

We did initial preprocessing of the documents by replacing web URL links with the <URL> token. We also removed exact duplicates and kept the first document that occurred by timestamp.

Document Filtering

We applied two types of document-filtering methods based on the amount of documents retrieved per flooding event: (1) retrieval and reranking method and (2) simple filtering method. This is described in Figure 1.

Method 1: Retrieval and Reranking (event has > 50 documents)

Step 1: Keyword-Based Retrieval. We apply BM25 retrieval algorithm using predefined keyword lists of flood causes to extract the top 25 most relevant documents per keyword, creating an initial candidate set. We use the PyTerrier (Macdonald and Tonellotto 2020) library.
Step 2: Removing Duplicates. We extract the TF-IDF representations for all retrieved documents using sklearn python library and calculate the pairwise cosine similarity. We remove documents with similarity scores 0.85 to eliminate near-duplicates while preserving unique content.
Step 3: Neural Reranking using MonoT5. We use Monot5 (Nogueira et al. 2020) using the transformers (Wolf et al. 2020) python library to rerank the retrieved documents for each query (keyword), selecting only the top 5 most relevant documents per query to reduce noise and improve precision.
This step was inspired by Pereira et al. 2023's document reranking approach applied for multi-document crisis summarization.
Step 4: Category-Based Selection. We group keywords by predefined category groups. Within each category, rank the top 5 documents, then perform cross-category ranking to select up to 80 documents that represent diverse, high-quality content across all categories.
Step 5: Diversity Sampling. From the pool of non-selected documents, we randomly sample up to 25% of the number of selected relevant documents (maximum of 20 documents). This ensures inclusion of potentially relevant edge cases and maintains dataset diversity.

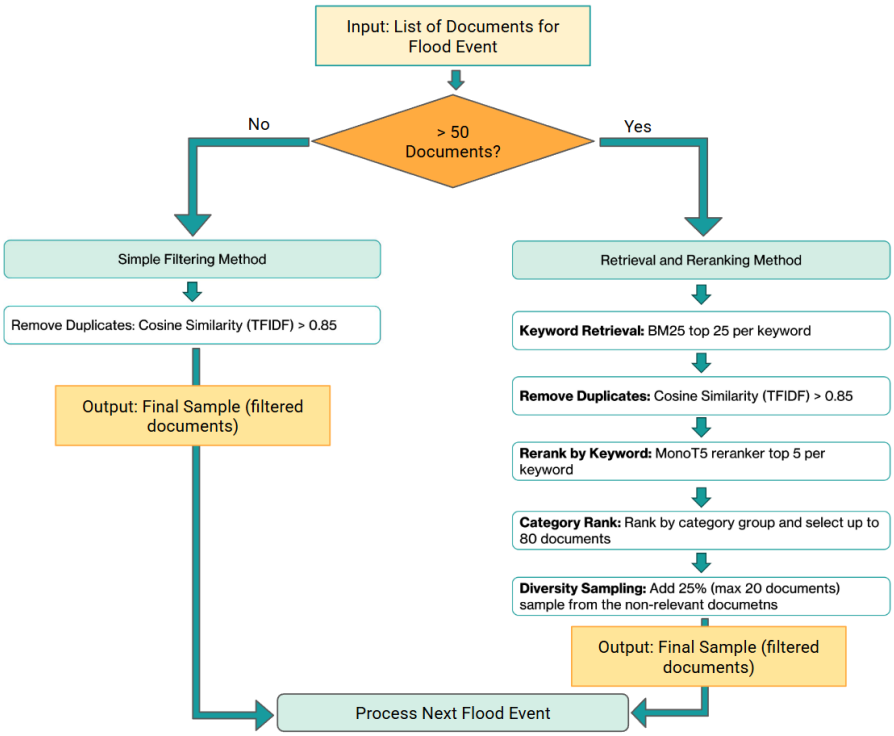

Figure 1. Document Filtering Process: If there are more than 50 documents retrieved for a flooding event, we apply The Retrieval and Reranking Method (Right). If there are fewer than 50, we apply the simple filtering approach (Left).

Method 2: Simple Filtering (event has ≤ 50 documents)

For flooding events with up to 50 documents, we apply only duplicate removal using cosine similarity on TF-IDF representations with a threshold of 0.85, preserving computational efficiency while ensuring data quality.

Sentence-level Filtering

We split our documents into sentences for our sentence-level annotation process. We use the sentence tokenizer from the NLTK (Bird et al. 2009) python library. Since not all of the sentences contain useful information for annotation, we filtered all the sentences that do not contain meaningful linguistic content and are therefore considered unintelligible or non-informative for annotation. We specifically removed sentences that composed entirely of links, metadata markers (hashtags/mentions), only numeric strings, only punctuations, or trivial promotional phrases. This ensures that only sentences containing substantive natural language content are retained for downstream annotation.

ANNOTATION PROTOCOL

The Flood Cause Annotation Protocol is the central contribution of this paper. It defines a standardized procedure for identifying and categorizing information about *why* flooding occurred in a set of documents. As shown in Figure 2, our protocol operates in two levels of granularity and in three stages: (1) document-level annotation (2) sentence-level annotation and (3) cause categorization.

The design is motivated by two complementary perspectives: the operational requirements and the NLP and linguistic annotation standards. First, the protocol reflects the operational requirements identified through direct experience in social media monitoring for disaster risk management. Organizations working across operational management, disaster response, insurance, and climate trend analysis benefit from the insights provided by online multi-platform data during crisis events. They share a common requirement: to go beyond the detection and the impacts of the flood, towards understanding the contributing factors of why it occurred. For crisis managers, for example, knowing whether the flood was caused by heavy rainfall vs. infrastructure failure can directly shape the resource allocation. For insurers, distinguishing long-term developmental causes from environmental causes is essential for risk modelling.

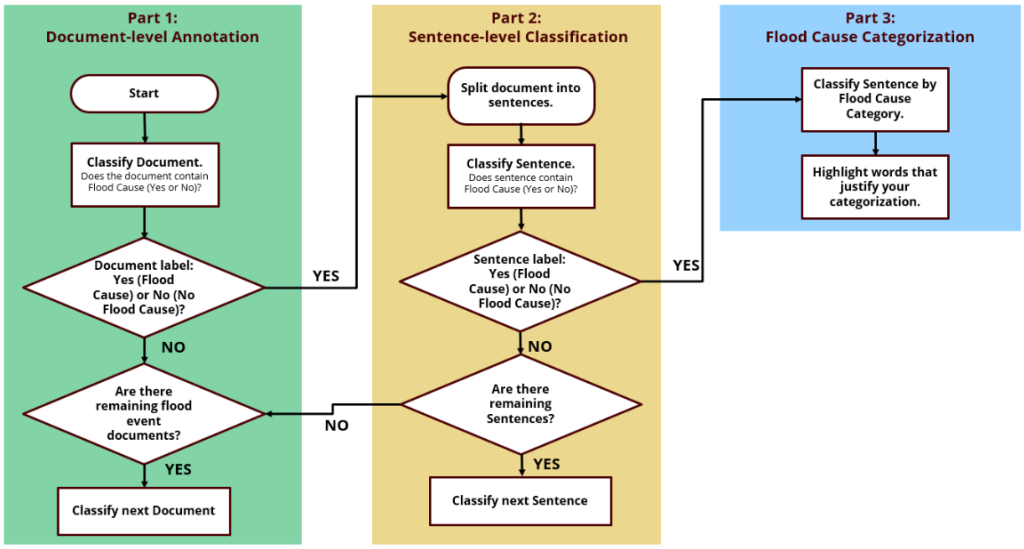

Figure 2. The Flood Causes Annotation Protocol defines a standardized procedure for identifying and categorizing information about why flooding occurred in a set of documents. The protocol is executed in three stages: (1) Document-Level Annotation – Determine whether a document contains information about flood causes. (2) Sentence-Level Annotation – Identify which sentences describe flood causes. (3) Flood Cause Categorization – Assign flood-cause categories and highlight supporting evidence.

Second, the protocol is informed by NLP and linguistic annotations standards. The two-level granularity annotation of relevance is a common practice in information extraction. The exclusion criteria were refined to address ambiguities and nuances in crisis text: for instance, flood-type descriptors such as "flash flood" or "pluvial flood" are frequently co-located with causal language, hence explicit guidance is needed. Furthermore, the inclusion of rationale span annotations aimed to capture data that can be suitable for explainability and interpretability research.

We outsourced our annotation to Isahit (Isahit n.d.), a social enterprise providing ethical data annotation services. The annotation was conducted by seven English-proficient annotators from Isahit. Annotation quality was maintained through an iterative process of annotation protocol refinement, annotator feedback, and repeated validation.

Document-level Annotation

The primary task determines whether a document contains specific information about flood causes for a flood event. A positive label requires the satisfaction of two criteria:

1. The document is about a flooding event. This includes a direct description of flooding, mentions of ongoing or recently occurring flood, or reference to concrete flooding event (even with limited details).

2. The document contains at least one flood cause. A flood cause explains why the flooding occurred (natural or human-related). Causes may be explicit or clearly implied. If at least one from any flood cause categories appears, it is annotated as flood cause.

We defined five mutually exclusive flood cause categories as guide for our annotators. The final categories was constructed using an iterative process that combined expertise across hydrology, social media monitoring, report generation, and human annotation.

- **Rain/Tide/Snow** - mentions of rain, snow, tide, wind, other natural triggers causing the flooding
- **Overflowing Water Bodies** - mentions rivers, creeks, streams, channels, or other reservoirs overflowing
- **Infrastructure Failure** - mentions protections structures such as dams, flood walls, dikes breaking or failing
- **Drainage/Blockage/Maintenance Issues** - mentions poor drainage, blocked systems, or other maintenance issues
- **Long-term developments (that caused flood)** - mentions of long-term developments such as land-use changes that increased the flood risk such as deforestation, urbanization, soil erosion

We allowed for an **Other Cause** category that accounts for any other cause that is not captured by the five categories above.

Furthermore, we defined an exclusion list criteria, which include: descriptions flood impacts (damage, evacuation, road closures), flood types (pluvial flood, flash flood, coastal flood), repairs or clean up (sandbags, barriers, crews working) which happen after the flood, future plans (planned drainage upgrades), general weather information, hypothetical or forecasted flooding. These should not be annotated as containing flood causes.

Sentence-level Annotation

For documents identified as containing a flood cause, we perform sentence-level extraction. Annotators select the specific sentences that explain the cause. Similar flood cause categories (see list above) and exclusion lists from the document-level annotation were provided to the annotators.

Sentences that were annotated as containing a flood clause are classified based on the pre-defined flood cause categories. For each classification, the annotators highlight the specific words or phrases in the sentence that directly support their decision as a rationale. Sentences can have more than one cause category and set of rationales.

We observed discrepancies between sentence-level and document-level annotations, particularly, where no flood cause sentences were detected in a document annotated as having a flood cause. For these instances, we re-annotated the document-level labels based on the sentence-level evidence.

Annotation Quality

We created a gold-standard subset (four flood events) annotated by our domain expert with expertise in crisis informatics and computational linguistics (first author). Annotation quality was evaluated by measuring agreement between our hired annotators and this gold standard. The raw agreement for this sample on the document-level annotation was at 95% F1-score and 94% Accuracy, with Cohen's Kappa(κ) inter-annotator agreement at 0.88. For the sentence-level, the raw agreement was at 85% F1-score and a 93% Accuracy agreement. The Cohen's Kappa(κ) inter-annotator agreement was 0.81. Both granularities show almost perfect agreement, with the sentence-level annotation being more nuanced. In future work, we will report agreement metrics on the multi-label flood cause categorization, as well as rationale variations.

DATASET UTILITY: A BENCHMARK FOR CRISIS REASONING

Although the annotated corpus is not publicly released, its construction methodology and the annotation protocol together define a reproducible benchmark framework. Researchers working with comparable crisis data can apply the protocol and compare results against the benchmarking baselines we establish here.

Causality vs. Impact Disambiguation: A major challenge for current AI models is distinguishing between the source of a disaster and its consequences. Our dataset provides meticulously annotated ground truth to test whether models can filter out "impact" noise (e.g., evacuations, road closures) to identify root causes—a capability essential for automated disaster forensics.

Objective vs. Subjective Causality: Our annotation protocol accounts for only objective statements as cause of the flooding event, subjective statements were annotated as not a flood cause. We acknowledge that disinformation or disinformation detection in online media is a challenging task on its own. There will always be the inherent limitation of real truth-finding unless field work is done. However, this dataset can be utilized to diagnose how LLMs extract what we consider as subjective causes –speculations and blaming statements from social media data.

Zero-Shot Performance Testing: The detailed annotation protocol (separating natural causes like "heavy rain" from human errors like "maintenance issues") enables fine-grained error analysis. The dataset can diagnose specific model blind spots—for instance, determining if a model is biased towards predicting natural causes over infrastructural ones.

INITIAL BENCHMARKING OF LLM ZERO-SHOT PERFORMANCE

We focus our initial evaluation on the initial challenge of causality extraction: identifying causal statements from the general disaster-related text, such as impacts, flood types, and weather forecasts. We evaluate the zero-shot performance of the LLMs in the first two stages of our annotation protocol (1) Document-level classification and (2) Sentence-level extraction. Evaluating the multi-label (3) Flood cause categorization stage is reserved for future work, as achieving high performance on the first 2 stages are necessary prerequisites.

Models

We run our initial benchmarking experiments with three commonly used off-the-shelf LLMs. We used GPT 4o-mini (OpenAI 2024),llama 3.1-8B Instruct (Llama Team 2024), and Mistral NeMo (Mistral AI Team 2024) These models were chosen because they are commonly used by both researchers and the public. We set the temperature settings at 0.0 to make all models deterministic in their prediction. All the other parameters were kept default. Details about the models used can be found in Table 2.

Model	Type	Source (OpenAI/Huggingface)
gpt-4o-mini	closed	gpt-4o-2024-08-06
llama 3.1 - 8B Instruct	open	meta-llama/Llama-3.1-8B-Instruct
Mistral NeMo	open	mistralai/Mistral-NeMo-Base-2407

Table 2. Information about the LLMs evaluated.

Prompts

We constructed our classification prompts with reference to the annotation protocol. We experimented on various prompt iterations and selected the prompt template with the highest rate of output format-following. Table 3 shows the document-level flood cause classification prompt. In future work, we will also look into prompt-tuning with the focus on overall flood cause extraction performance.

You are provided with a document. Classify whether the document contains information about flood causes for a flood event.
Criteria for classification:
Yes:
Classify as "Yes" if both of the following conditions are met:
1. The document is about a flooding event. This includes direct description of flooding, mention of ongoing or recently occurred flood, or reference to concrete flood events.
2. The document contains at least one flood cause. A flood cause explains why the flooding occurred (natural or human-related). Causes may be explicit or clearly implied. If at least one from any flood cause categories appears, annotate as "YES".
Flood cause categories include:
- Rain / Tide / Snow (heavy rain, precipitation, snowmelt, storms, high tide, wind-driven waves),
- Overflowing Water Bodies (overflowing rivers, swollen streams, full canals),
- Infrastructure Failure (dam failure, breached levee, broken embankment, collapsed flood wall),
- Drainage/Blockage/Maintenance Issues (clogged drains, blocked pipes, trash build up, poor maintenance),
- Long-Term Developments (deforestation, urbanization, land-use change, soil erosion)
No:
Classify as "No" if:
No conditions were met above.
The document ONLY contains flood impacts, flood types, repairs or clean ups, future plans, general weather, hypothetical or forecasted flooding.'
Output ONLY one word: "Yes" or "No". Do not provide any explanation or reasoning, only the classification label.

Table 3. Prompt for Document-level Flood Cause Classification.

Results

Document-Level Classification

For the initial benchmarking experiments, we construct a diverse evaluation subset comprising 10 distinct flooding events, consisting of 615 documents. These 10 events were purposively sampled from the full 70-event corpus to ensure geographic diversity and represent a range of distinct flood causes (e.g., extreme rainfall, infrastructure failure). This subset provides a robust, manageable test bed to validate the annotation protocol and establish initial zero-shot LLM baselines before scaling to the full dataset in future work.

We present the document-level classification results in Table 4. GPT 4o-mini has the best performance based on F1-score for all Documents at 85% F1-score. This was followed by Llama 3.1 - 8B at 82% and Mistral NeMo at 76%. We observed that for GPT 4o-mini and Mistral NeMo, there is a gap between recall and precision. This means that although they identify the true positives (documents containing flood cause) but also tend to generate a high number of false positives.

Sentence-level Cause Classification

We evaluate the fine-grained sentence-level cause classification of the LLMs independent of their document-level performance. All three models extract causal sentences exclusively from documents that were already annotated as

Event	# of	GPT 4o-mini			Llama 3.1 - 8B			Mistral NeMo		
	Docs	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
All Events	615	91	79	85	83	82	82	90	66	76
Valencia (Nov 2022)	39	92	92	92	68	100	81	92	74	82
Uganda (Nov 2023)	50	77	80	78	65	94	73	85	71	77
California (Jan 2024)	100	91	67	77	67	72	70	91	57	70
Sumatera (May 2024)	100	99	92	95	98	94	96	99	87	93
UK (Oct 2024)	100	82	55	66	85	64	73	79	43	56
Queensland (Mar 2025)	100	86	83	85	76	85	80	83	77	80
Ivory Coast (Jun 2025)	20	100	100	100	92	100	94	92	86	89
West Nusa Tenggara (Jul 2025)	35	100	94	97	94	94	94	100	59	74
Cote d'Azur (Sep 2025)	30	92	80	86	69	90	78	92	55	69

Table 4. Document-level Performance of LLMs in Flood Cause Classification. Recall, Precision, and F1-score.

containing flood cause (positive documents only). These models must identify the specific causal sentences in the documents, while correctly discarding non-causal sentences (e.g., impacts, flood types, weather forecasts) that occur in the same document. We used the same 10-distinct flooding events.

Table 5 shows the sentence-level cause classification. We observe a more apparent gap between precision and recall in sentence-level granularity. While models are relatively good at capturing sentences containing flood causes (recall scores above 82% across all events, they routinely include generic flood information. For example Llama 3.1 - 8B, shows very high recall at 100% in all events, but the very poor precision scores can indicate that the model is generally biased on the "Yes" classification or that it really struggles to disambiguate the other flood-related information from the causal statements only. Their precision is very much lacking.

Event	# of	GPT 4o-mini			Llama 3.1 - 8B			Mistral NeMo		
	Sentences	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
All Events	967	97	51	67	100	30	46	96	32	48
Valencia (Nov 2022)	71	96	72	82	100	48	65	96	47	64
Uganda (Nov 2023)	97	100	32	42	100	18	31	92	18	30
California (Jan 2024)	108	94	52	67	100	28	43	100	27	43
Sumatera (May 2024)	225	100	47	64	100	25	40	100	26	41
UK (Oct 2024)	152	100	33	50	100	18	30	87	19	31
Queensland (Mar 2025)	113	82	41	55	100	24	39	94	28	43
Ivory Coast (Jun 2025)	45	100	50	67	100	42	60	100	47	64
West Nusa Tenggara (Jul 2025)	62	100	89	94	100	56	72	96	59	73
Cote d'Azur (Sep 2025)	25	100	82	90	100	74	85	93	76	84

Table 5. Sentence-level Performance of LLMs in Flood Cause Classification. Recall, Precision, and F1-score.

CONCLUSION AND FUTURE WORK

We introduced a comprehensive flood cause annotation protocol for extracting and classifying flood cause from multi-platform data. The protocol provides a three-stage pipeline and explicit guidelines for distinguishing the flood causes from flood impacts and flood types.

Our initial benchmarking experiments on a stratified 10-event sample demonstrate that LLMs struggle: they have high recall but low precision in flood cause classification. They tend to include general flood information like impacts. This was observed in both document-level and sentence-level granularities, but significantly more apparent in the sentence-level extraction.

Although we cannot release the data, we provide a full description of the data construction pipeline and the complete annotation protocol to facilitate replication on comparable data. Future work includes:(1) evaluating LLMs in the multi-label cause categorization task (stage 3 of the annotation protocol) (2) exploring few-shot and fine-tuned LLM approaches; and (3) extending the protocol to flood cause summary creation.

ACKNOWLEDGMENTS

This publication is part of the project 'Indeep: Interpreting Deep Learning Models for Text and Sound' with project number NWA.1292.19.399, which is partly financed by the Dutch Research Council (NWO).

REFERENCES

Alam, F., Sajjad, H., Imran, M., and Ofli, F. (May 2021). "CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing". In: *Proceedings of the International AAAI Coanference on Web and Social Media* 15.1, pp. 923–932.

Bird, S., Loper, E., and Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Chowdhury, M. T. A., Datta, S., Sharma, N., and KhudaBukhsh, A. R. (2024). "Infrastructure Ombudsman: Mining Future Failure Concerns from Structural Disaster Response". In: WWW '24. Singapore, Singapore: Association for Computing Machinery, pp. 4664–4673.

Dela Cruz, J. A., Hendrickx, I., and Larson, M. (July 2025). "Improving Large Language Model Confidence Estimates using Extractive Rationales for Classification". In: *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*. Vienna, Austria and virtual meeting: Association for Computational Linguistics, pp. 549–560.

Duong, H. M., Nguyen, L., Levin, E., and Gary, T. (2025). "CaST: Causal Discovery via Spatio-Temporal Graphs in Disaster Tweets". In: *2025 IEEE International Conference on Big Data (BigData)*, pp. 5603–5612.

Feng, Y., Huang, X., and Sester, M. (2022). "Extraction and analysis of natural disaster-related VGI from social media: review, opportunities and challenges". In: *International Journal of Geographical Information Science* 36.7, pp. 1275–1316.

FloodTags (n.d.). *FloodTags: Real-time media monitoring for weather impact events*.

Fu, S., Lyu, H., Wang, Z., Hao, X., and Zhang, C. (2022). "Extracting historical flood locations from news media data by the named entity recognition (NER) model to assess urban flood susceptibility". In: vol. 612, p. 128312.

Imran, M., Mitra, P., and Castillo, C. (May 2016). "Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 1638–1643.

Imran, M., Ziaullah, A. W., Chen, K., and Ofli, F. (2025). "Evaluating Robustness of LLMs on Crisis-Related Microblogs across Events, Information Types, and Linguistic Features". In: *Proceedings of the ACM on Web Conference 2025*. WWW '25. Sydney NSW, Australia: Association for Computing Machinery, pp. 5117–5126.

Isahit (n.d.). *Isahit: Ethical data annotation and impact sourcing*.

Koju, S., Takeuchi, K., Watanabe, A., Hirayama, T., and Nakao, H. (2024). "Estimating Task Priority in Japanese Disaster Chronology Logs". In: *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval*. NLPIR '23. Seoul, Republic of Korea: Association for Computing Machinery, pp. 304–309.

Lei, Z., Dong, Y., Li, W., Ding, R., Wang, Q. R., and Li, J. (July 2025). "Harnessing Large Language Models for Disaster Management: A Survey". In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, pp. 14528–14551.

Llama Team (2024). *The Llama 3 Herd of Models*. arXiv: 2407.21783 [cs.AI].

Ma, K., Tian, M., Tan, Y., et al. (2023). "Ontology-Based BERT Model for Automated Information Extraction from Geological Hazard Reports". In: *Journal of Earth Science* 34, pp. 1390–1405.

Macdonald, C. and Tonellotto, N. (2020). "Declarative Experimentation inInformation Retrieval using PyTerrier". In: *Proceedings of ICTIR 2020*.

McDaniel, E., Scheele, S., and Liu, J. (2024). "Zero-Shot Classification of Crisis Tweets Using Instruction-Finetuned Large Language Models". In: *2024 IEEE International Humanitarian Technologies Conference (IHTC)*, pp. 1–7.

Mistral AI Team (2024). *Mistral NeMo*. https://mistral.ai/news/mistral-nemo.

Nogueira, R., Jiang, Z., Pradeep, R., and Lin, J. (Nov. 2020). "Document Ranking with a Pretrained Sequence-to-Sequence Model". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 708–718.

Olteanu, A., Castillo, C., Diaz, F., and Vieweg, S. (May 2014). "CrisisLex: A Lexicon for Collecting and Filtering Microblogged Communications in Crises". In: *Proceedings of the International AAAI Conference on Web and Social Media* 8.1, pp. 376–385.

OpenAI (2024). *GPT-4o mini: advancing cost-efficient intelligence*.

Peng, H. and Zhu, K. (2025). "FAR-AM: A hybrid attention framework for fire cause classification". In: *PLOS ONE* 20.10, e0333131.

Pereira, J., Fidalgo, R., Lotufo, R., and Nogueira, R. (2023). "Crisis Event Social Media Summarization with GPT-3 and Neural Reranking". In: *Proceedings of the 20th International ISCRAM Conference*, pp. 371–384.

Qian, J., Du, Y., Liang, F., Yi, J., Wang, N., Tu, W., Huang, S., Pei, T., and Ma, T. (2024). "Quantifying Urban Linguistic Diversity Related to Rainfall and Flood across China with Social Media Data". In: *ISPRS International Journal of Geo-Information* 13.3, p. 92.

Wiegmann, M., Kersten, J., Klan, F., Potthast, M., and Stein, B. (2020). "Analysis of detection models for disaster-related tweets". In: *17th Annual International Conference on Information Systems for Crisis Response and Management, ISCRAM 2020*, pp. 872–880.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45.

Xu, F., Ma, J., Li, N., and Cheng, J. C. (2025). "Large language model applications in disaster management: An interdisciplinary review". In: *International Journal of Disaster Risk Reduction* 127, p. 105642.

Ye, P., Zhang, C., Chen, M., and Li, S. (2024). "Typhoon disaster state information extraction for Chinese texts". In: *Nature Scientific reports* 14.1, p. 7925.