Retrieving Floods without Floodlights: Topic Models as Binary Classifiers for Extreme Climate Events in German News

Brielen Madureira[1,2], Mariana Madruga de Brito[2], Andreas Niekler[1,3]

[1]LeipzigLab - Climate Discourse, Leipzig University, Germany
[2]Helmholtz Centre for Environmental Research - UFZ, Germany
[3] Computational Humanities, Leipzig University, Germany

brielen.madureira@uni-leipzig.de
mariana.brito@ufz.de
aniekler@informatik.uni-leipzig.de

Abstract

In studies of media coverage of extreme climate events, NLP methods have become indispensable for identifying relevant texts in large news databases. Still, enough annotated data to train accurate deep learning-based classifiers from scratch is often not available. Topic Models have the advantage of being both unsupervised and interpretable, but are typically used only for exploratory analysis or data characterisation. In this study, we investigate how to employ Topic Models as binary classifiers for refining the retrieval of relevant news about seven types of extreme climate events in the German media. Our method relies on the *posterior* distributions estimated by Topic Models to select relevant documents, without modifying their training procedure. Using an annotated sample to guide the evaluation, we show that the probabilities assigned to keywords used to query news databases can also be informative for selecting relevant topics and improve sample precision. We compare our results to a fine-tuned text embedding classifier and an open-weight LLM, discussing observed trade-offs, e.g. the LLM's lowest precision. Moreover, we show that results are hazard-dependent, which speaks against considering climate events as a single category in NLP tasks.

Keywords: extreme climate events, German news, topic models, text classification, document retrieval

1. Introduction

Assume we are gathering news about floods events to study collective attention in the media. Simply querying a news database to retrieve documents containing the string *flood* would not only match news reporting on actual floods, but also many false positives. Consider this (obviously constructed) example: "*Soccer fans experienced a flood of emotions witnessing floodlights being turned on as players flooded the field: the game could finally begin after the risk of a flash flood was ruled out.*" This illustrates a central challenge in information retrieval: the term *flood* can have metaphorical senses, be part of compound nouns unrelated to climate or refer to a merely hypothetical hazard. Thus, despite the repeated presence of the term *flood*, this text is rendered unrelated to actual flood events.

Pitfalls like that can emerge at the intersection of environmental and social sciences, such as in text-based climate impact and adaptation research. This field often relies on NLP methods to process texts about climate events and their consequences (Alencar et al., 2024; Nunes Carvalho et al., 2024, *inter alia*). In this context, dictionary-based retrieval is a typical procedure: large databases are queried using a curated list of hazard-related keywords to find potentially relevant documents about e.g. floods, droughts or wildfires (e.g. Sodoge et al., 2023; Li et al., 2025). But as we just saw above, the mere presence of a keyword in a document does not guarantee its relevance. If term presence or frequency are directly used as predictors in quantitative assessments, research validity is impaired.

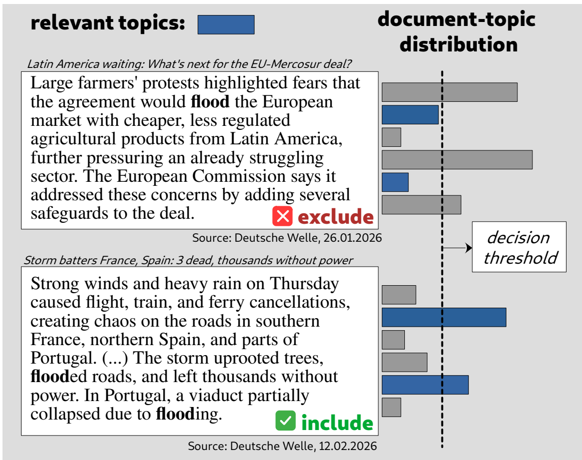

Figure 1: Relevant news articles can be identified based on the saliency of relevant topics in their representation estimated by a topic model.

Although keyword-based matching serves as a reasonable prefiltering step for creating an initial collection of documents with good recall, retrieval results must be further refined using other classi-

Proceedings of the 2nd Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology) @ LREC 2026, pages 115–134
May 12, 2026 ©European Language Resources Association (ELRA), 2026

fication methods to detect true and false positives and improve their precision. This is an open problem recently discussed by Grasso et al. (2024).

In this paper, we investigate the possibility of yielding a binary classification model for identifying relevant documents using the probabilities estimated by unsupervised vanilla Topic Models (TM), as the overview in Figure 1. We assume a situation with a small amount of annotated data that is not enough for training deep learning-based models from scratch but still informative for evaluation.

Our main contributions in this paper are:

i a data analysis of news articles in German annotated with seven types of extreme climate events;

ii the usage of TMs for relevance classification without any needed modification on the training regime and no direct human effort in topic interpretation; and

iii evidence that TMs are, for some hazards, on par with deep learning alternatives, with the advantage of interpretability and a tendency to higher precision.

2. Related Literature

Retrieval of environment-related documents Document retrieval is an ubiquitous step in creating corpora for socio-environmental research. To name a few recent large-scale approaches, Leippold and Varini (2020) implemented a graph-based heuristic on Wikipedia metadata of entries on climate topics, Kong and Purves (2026) relied on climate-related keywords to retrieve news and Cai et al. (2025) used a hazard event database for a targeted query of news articles and refined results using a Large Language Model (LLM). Our work focuses on the step of *refining* an initial sample of documents retrieved via keyword-matching methods.

Topic Models TMs such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorization (NMF) yield distributions of topics in documents in an unsupervised fashion. Many variations exist, e.g. keyword-assisted TMs (Eshima et al., 2024), which exploit keywords to guide clustering and circumvent post hoc topic interpretation, CorEx (Gallagher et al., 2017), which relies on an information-theoretic framework, and Top2Vec (Angelov and Inkpen, 2024), which performs clustering in an embedding space shared by documents and words. TMs often aid characterising corpora in climate and ecology research (Lesnikowski et al., 2019; Stede et al., 2023; Madruga de Brito et al., 2025; Zander et al., 2023; Peura et al., 2025; Beckles and Heidke, 2025; Barz et al., 2025, *inter alia*).

Tuning and assessing the quality of TMs is intricate if performed purely intrinsically (Maier et al., 2021), but our evaluation is enhanced by annotated data that allows known document properties to be compared to the formed topics.

TMs for text classification TMs have been widely used to map texts to classes, e.g. by feeding their outputs as input features for other classifiers (Li et al., 2016b; Anantharaman et al., 2019; Seifollahi et al., 2021). Other works aligned topics to classes, either directly (Sarioglu et al., 2013), by experts (Hingmire et al., 2013) or by configuring the *priors* in a way that induces desired clusters (Miller et al., 2016; Rubin et al., 2012), e.g. by relying on relevant keywords (Chen et al., 2015; Zha and Li, 2019; Li et al., 2016a, 2018). Keyword selection can also derive a lower dimensional set of features for other types of classifier models (Onan et al., 2016). Mcauliffe and Blei (2007) incorporated a response variable into the TM training, to jointly model documents and their classes or scores. While many procedures require adjusting *priors* or the modelling approach, we stick to standard LDA and NMF implementations, which are arguably more accessible for newcomers and researchers from other fields.

Text classification in climate research Climate-related text classification is an established NLP task; in many settings, it remains an unsolved problem even for LLMs, with performance often well below 0.75 F1 in the ClimateEval benchmark (Kurfali et al., 2025). In the study by Li et al. (2024), a fine-tuned encoder achieved an F1 of 0.98 for identifying relevant documents on climate extreme impacts, but only in English and on a small sample of cleaner Wikipedia entries with climate-related keywords *in their titles*. This restriction likely ensured a majority of relevant matches, but resulted in an unknown number of missed cases. When full texts are considered (as we do), there is less room for false negatives while substantially increasing the need for filtering out false positives, especially with imbalanced datasets.

The problem we tackle in this paper is similar to the work by Grasso et al. (2024): corpus construction via keyword-based prefiltering and automatic classification. We differ by focusing on German, handling specific hazards separately and exploring TMs for classification, not only for topic analysis as that work did. Our design builds upon existing work with a novel perspective: we do not change the LDA and NMF internal mechanisms and explore the *posterior* probabilities (or normalised scores) they assign to keywords as a means to automatically partition topics and perform binary classification of news about extreme climate events.

3. Methods

This section formalises the task and explains how topic models are applied for binary classification. Then, it describes the two deep learning strategies used for comparison.

3.1. Task Formalisation

Let D be a set of documents d, each belonging to a binary class $C = \{0, 1\}$, and V be the set of all tokens w that appear in D. Class 1 represents relevant documents. A document classifier is a function $f : D \to C$ that maps documents to classes and can be approximated by various methods.

Furthermore, let $F \subseteq V$ be a set of feature tokens selected from V based on given criteria (e.g. minimum frequency and part-of-speech tags) and $K \subset F$ be a small set of predefined tokens of interest which we name *keywords*. A trained topic model M with n topics T estimates two distributions: p_{feat}, the probability of a feature token in a topic and p_{topic}, the probability of a topic in a document. In other words, a document is represented as a probability distribution over topics and a topic, as a probability distribution over feature tokens.

With M's estimations, we can define a binary relation R between D and T representing whether each document is related to each topic. To use M for binary classification, a partition of topics T with two sets is created, each corresponding to a class in C. The class of a topic must also correctly classify documents related to such a topic.

3.2. Classification with Topic Models

Firstly, a TM is trained on the entire collection of unique documents using selected hyperparameters, following standard procedures (described in Section 5). Then, two further steps are needed: (i) assigning topics to documents and (ii) identifying which topics are to be regarded as relevant.

For (i), we define the relation R as $p_{topic}(t, d) \geq \theta$ with $0 \geq \theta \geq 1$. That means that if the proportion of a topic t in a document d is at least a threshold, we consider that d discusses topic t (as in Figure 1).

For (ii), we propose two ways to partition topics into two classes, relevant and not relevant, avoiding the usual *post hoc* human interpretation in TMs:

- **keyword proximity**: topic t is assigned to the relevant class if $\exists w \in K : p_{feat}(w, t) > \gamma$. In other words, if the topic assigns a high enough probability to at least one keyword, the topic belongs to the partition of the relevant class.

- **top terms**: if there is a keyword among the top k features of a topic (ranked by probability), the topic is assigned to the relevant class.

Deadly 'unprecedented' wildfire rages in southwestern France

More than 2,100 **firefighters** are still battling an intense **wildfire** in the Corbieres hills in a Mediterranean region of France near the border to Spain. They are supported by some 90 **firefighting** planes (...) From Thursday, several dozen soldiers and three army helicopters will help battle the **fire**, which broke out Tuesday afternoon in the village of Ribaute in the Aude region, a rural, wooded area home to wineries. The **blaze** has since swept through 16,000 hectares. Authorities described it as the largest **fire** in nearly 80 years. Source: Deutsche Welle, 26.08.2024

Why Brazil faces a water crisis

Home to the largest river in the world by volume, the Amazon, Brazil holds about 12% of global freshwater reserves. As a result, many Brazilians have long run with the narrative that their water is abundant. But that is changing. (...) Last year, Brazil was plagued by water shortages from the north to the south. Authorities declared a state of water scarcity in five major river basins, and **wildfires** ravaged the Amazon rainforest and Pantanal wetlands after months of drought. Source: Deutsche Welle, 16.04.2025

Figure 2: Wildfires as the main topic (top) or mention (bottom) in news excerpts.

The actual classification of each document is made as follows: if the document is related to at least one of the relevant topics, we consider it to be relevant. Otherwise, it is classified as not relevant.

3.3. Deep Learning Classifiers

The performance of our TM approach is compared to two deep learning alternatives: a fine-tuned text embedding model and an LLM. The first is a binary classifier trained using the SetFit framework (Tunstall et al., 2022) which fine-tunes a pretrained text embedding model with a classification head, aiming at optimising task-specific embeddings based on a set of contrastive examples. The latter prompts an LLM to generate a binary label classifying the document as relevant or not. The implementation details are explained in Section 5.

4. Data

The data for this study derives from an ongoing project on the collective attention to extreme climate events in the German media. Seven types of hazards were selected (cold waves, droughts, floods, heat waves, landslides, storms and wildfires). The wiso-net news aggregation database[1] was queried using a pre-defined list of hazard-related keywords, similar to (Li et al., 2024; Madruga de Brito et al., 2025; Carvalho et al., 2025, see Appendix). The retrieved collection contained 13,771,411 German news articles from around 370 outlets, spanning from 2000 to 2024, split into separate subcollections for each extreme climate event.

We make a distinction between two types of relevant news: `main`, in which the extreme climate event is the main topic, and `mention`, that refer to it *en passant*, of secondary importance among other more prominent topics, as shown in Figure 2. Both

[1] https://www.wiso-net.de/

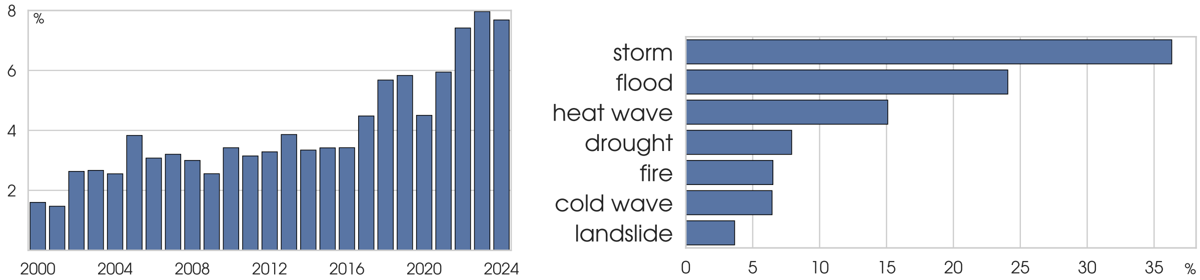

Figure 3: % of documents per year (left) and % of documents per extreme climate event (right).

forms count towards levels of collective attention, but automated identification of the latter is more challenging due to brevity and underspecification.

We are primarily interested in the coverage of *international* events in German news, so basic rule-based procedures were implemented to reduce the amount of local or unrelated news. As the data we are working with was queried via unrestricted keyword matching, many false positives occurred. Of particular relevance to this paper is the filter for what we call "intruder" keywords, i.e. words that derive from valid hazard-related keywords but are unrelated to climate events. For instance, the name *Dürrenmatt*, retrieved by *Dürre* (drought), *Flutlicht* (floodlight), by *Flut* (flood) and *Stürmer* (forward player in football), by *Stürme* (storms). To reduce the number of unrelated documents, we excluded all instances that contained only intruder keywords. Other preprocessing steps for filtering and cleaning the data are summarised in the Appendix.

Final document collection The previous steps resulted in a sample of 2,438,275 documents (17.71% of the originally retrieved instances). They have, on average, 537.19 tokens (std=362.06) and 28.66 sentences (std=20.12). The news database has an inherent temporal bias towards recent years, illustrated in Figure 3 (left). The distribution over types of hazards is also shown in Figure 3 (right).

Gold standard A sample of 3,150 documents was randomly selected while ensuring uniform distribution across hazard types (450 each) and years (18 each per hazard) and no duplicates per hazard. Two annotators classified the news as relevant or not (see annotation instructions in the Appendix) while also judging whether the event of interest was the news' main topic or just a mention. Table 1 shows the percentage of relevant documents identified in the annotated sample. The initial effectiveness of the dictionary-based approach, together with the rule-based preprocessing, depends on the type of extreme climate event: while the landslide portion already reached a precision of almost 0.6, cold and heat waves stayed below 0.2.

	relevant	main topic
cold wave	14.44	4.22
drought	36.00	4.22
flood	43.33	12.44
heat wave	19.78	2.44
landslide	58.67	17.56
storm	27.56	7.33
wildfire	41.56	21.11

Table 1: Percentage of relevant documents for each type of extreme climate event in the gold standard.

Identifying relevant documents is not as straightforward as it may sound. A subset of 100 news was annotated by both annotators independently. The overall agreement proportion in the primary binary decision (relevant or not) was 0.77 ($\kappa = 0.53$), indicating that there are sources of legitimate disagreement in this decision. Apart from potential errors, disagreements may stem from differences in perception of what constitutes an *extreme* and *concrete* natural event. Some cases which may have involved such subjectivity were (translated from German):

- **cold wave**: "...the son, travelled on Monday during the snow chaos from Cologne to Wismar in order stay by his mother."
- **drought**: "the mine was found two weeks ago due to the low water level in the Rhine river."
- **landslide**: "we knew the situation when a country closed the border or a street was blocked for a week due to a landslide or something else."

5. Experiments

The varying estimated proportions of relevant documents for each hazard sample suggest that these phenomena manifest differently not only in their nature but also in their coverage and linguistic features. Therefore, each classification strategy was conducted for each type of extreme climate event separately. The annotated sample was randomly split into a training and a test set with 350 and 100 instances, respectively, for each hazard. The presented results were computed in the test split.[2]

Topic models Documents were preprocessed to extract their features partially based on the procedure by Grasso et al. (2024). We used Spacy's[3] model de_core_news_lg to tokenize, lemmatise and label tokens with their part-of-speech tags. Tokens with less than 3 characters and stopwords were removed, as well as non-alphabetical characters. All tokens were lowercased. The feature selection involved two criteria: the term's document frequency and part-of-speech tag. All keywords were kept as features, even if they did not meet the minimum frequency threshold, to ensure they had a chance to contribute to forming a topic. To avoid the induction of topics based on duplicated news, only one instance of texts with high similarity was included. Gensim's[4] implementation of the LDA and NMF methods was used to train topic models. The number of topics was a hyperparameter. For LDA, the eta and alpha arguments were set to auto. We run various combinations of the three hyperparameters (minimum document frequency, part-of-speech tags and number of topics) and, for each model, we computed results varying the values for k, γ, for top term and keyword proximity, and θ. For each hazard, we selected the best-performing models in the training split. Specific parameters and the final configuration that produced the results are in the Appendix. The code is available at https://codeberg.org/briemadu/tm-as-classifier.

Text embeddings This classifier was trained via the Small-Text (Schröder et al., 2023) wrapper implementation around SetFit (with its default configuration in HuggingFace) and Sentence Transformers (Reimers and Gurevych, 2019). In this method, the classification is performed by a logistic regression component on top of the fine-tuned text embeddings. We opted for the BAAI/bge-m3 text embeddings released by (Chen et al., 2024) due to the model's multilingual capabilities and longer context length (8,192 tokens), since standard Sentence Transformers that typically allow only up to 512 tokens would not suffice for longer news articles. Training was performed with a batch size of 16 instances and a learning rate of 10^{-5}.

LLM Since the purpose of this paper is not to benchmark LLM performance, we chose only one model to serve as a reference. Results were produced by mistralai/ministral-3-14b-reasoning.[5] We selected an open-weight model that could be run locally and keep the data in our own infrastructure.[6] The prompt contained instructions similar to those given to the annotators, including the definition of the hazard and of the labels, the hazard's keywords and the main body of the news article. The exact prompt and values are in the Appendix. We had to programatically parse answers that included spurious prefixes before the actual label.

Evaluation

The models' performance was quantitatively assessed with conventional binary classification metrics: precision, recall and F1 score of the positive class. The test sample's precision and a presumed recall of 1 were used as a baseline to measure how much the classifiers improve retrieval precision without reducing its recall. The evaluation was enriched with a detailed analysis of the TM results.

We present results for three variations of TMs: TM-F1 was run with the configuration that resulted in the highest F1 score (on the training split) in our hyperparameter search; TM-B uses the configuration that balanced precision and recall to be both as high as possible; and TM-P has the configuration with the highest precision while retaining some level of recall. We also compare results to an ensemble strategy that performs classification via majority voting across the outputs of TM-B, fine-tuned text embeddings and LLM classifiers.

6. Results

Aggregated results We first examine results aggregated over the whole test split ($n = 700$), i.e. including all extreme climate events. Table 2 shows precision, recall and F1 score for all classifiers. The rightmost column shows the number of news articles of type main that were correctly identified as relevant. All classifiers succeeded in considerably increasing the low proportion of relevant documents in the keyword-based sample, but TMs and deep

[2]Note that the use of train/test splits depends on the classifier. TMs's unsupervised fitting included all unique documents, since the objective here is not to generalise to unseen data but to optimise for topics that best fit our own documents. Still, only the train split was used to select the best model configuration to avoid overfitting to the test data in this choice. The text embeddings model used the train split for fine-tuning. The LLM was directly prompted with the test data in a zero-shot approach.

[3]https://spacy.io/

[4]https://radimrehurek.com/gensim/

[5]https://huggingface.co/mistralai/Ministral-3-14B-Reasoning-2512

[6]We did not compare results to closed commercial models as they are at odds with open science principles.

	P	R	F1	n main
baseline	0.350	1.000	0.519	58
TM-F1	0.637	0.710	0.672	56
TM-B	**0.710**	0.649	0.678	55
TM-P	**0.808**	0.396	0.532	47
fine-tuned	0.647	0.853	0.736	57
llm	0.583	**0.976**	0.730	58
majority	0.701	0.890	**0.784**	58

Table 2: Aggregated results: binary precision, recall and F1 score of all classifiers in the test split and the number of news of type `main` correctly identified.

learning strategies behaved differently in how precision and recall were balanced. While the LLM had almost maximum recall with a substantial margin over other models, its precision was the lowest. TM-P had the highest precision but at the cost of low recall. TM-B achieved the second highest precision with a more reasonable recall. The majority voting approach led to the highest F1 score. If we focus on the identification of news of type `main`, all classifiers (apart from TM-P) performed very well, identifying at least 55 out of the 58 instances.

Results by hazard Aggregated results can mask variations in performance for each underlying hazard. Table 3 summarises results by hazard type, in line with the fact that models were trained separately. We can see that metrics varied greatly depending on the phenomenon: the lowest best F1 score of 0.59 occurred for heat wave whereas the highest best of 0.92 was observed for landslide. Majority voting achieved the best F1 scores for five hazards and the fine-tuned text embeddings for the other two. The LLM consistently held the highest recall in all hazards. TM-B had the best precision in the three most imbalanced (cold waves, heat waves and storms).

Discussion In aggregated results, TM performance was indeed lower than that of deep learning strategies, but the moderate reduction of only around 0.06 in F1 score still provides a much-desired gain in interpretability: we can explain exactly why each document was classified as relevant. The deep learning strategies tended to incur more false positives whereas TMs could reduce the proportion of unrelated documents while causing more false negatives. Models with higher precision but low recall, like TM-P, can still be useful when sample precision is a priority, since a sample with low recall may still be representative and of enough size in large datasets. High precision helps reduce the impact of unrelated documents in downstream anal-

		P	R	F1
cold wave	baseline	0.170	1.000	0.291
	TM-F1	0.471	0.471	0.471
	TM-B	**0.583**	0.412	0.483
	TM-P	0.500	0.059	0.105
	fine-tuned	0.297	0.647	0.407
	llm	0.455	**0.882**	0.600
	majority	0.542	0.765	**0.634**
drought	baseline	0.440	1.000	0.611
	TM-F1	0.517	0.682	0.588
	TM-B	0.622	0.523	0.568
	TM-P	**0.938**	0.341	0.500
	fine-tuned	0.686	0.795	0.737
	llm	0.525	**0.955**	0.677
	majority	0.692	0.818	**0.750**
flood	baseline	0.360	1.000	0.529
	TM-F1	0.605	0.639	0.622
	TM-B	0.595	0.611	0.603
	TM-P	**0.750**	0.167	0.273
	fine-tuned	0.737	0.778	0.757
	llm	0.600	**1.000**	0.750
	majority	0.738	0.861	**0.795**
heat wave	baseline	0.200	1.000	0.333
	TM-F1	0.423	0.550	0.478
	TM-B	**0.600**	0.450	0.514
	TM-P	**0.600**	0.150	0.240
	fine-tuned	0.439	0.900	**0.590**
	llm	0.322	**0.950**	0.481
	majority	0.429	0.900	0.581
landslide	baseline	0.580	1.000	0.734
	TM-F1	0.785	0.879	0.829
	TM-B	0.778	0.845	0.810
	TM-P	0.816	0.690	0.748
	fine-tuned	**0.877**	**0.983**	**0.927**
	llm	0.826	**0.983**	0.898
	majority	0.826	**0.983**	0.898
storm	baseline	0.270	1.000	0.425
	TM-F1	0.680	0.630	0.654
	TM-B	**0.800**	0.593	0.681
	TM-P	0.625	0.185	0.286
	fine-tuned	0.558	0.889	0.686
	llm	0.614	**1.000**	0.761
	majority	0.714	0.926	**0.806**
wildfire	baseline	0.430	1.000	0.601
	TM-F1	0.773	0.791	0.782
	TM-B	0.825	0.767	0.795
	TM-P	**0.844**	0.628	0.720
	fine-tuned	0.750	0.837	0.791
	llm	0.662	**1.000**	0.796
	majority	0.809	0.884	**0.844**

Table 3: Detailed results: binary precision, recall and F1 score of all classifiers in the test split for each hazard type.

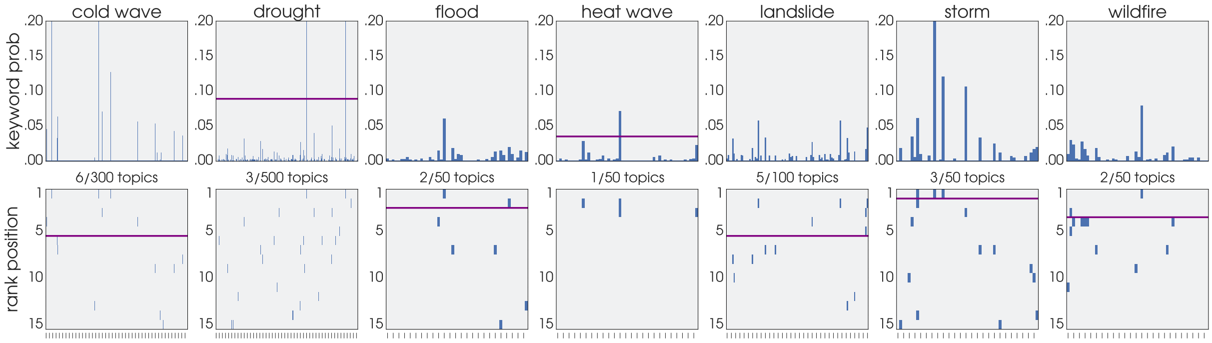

Figure 4: High-level overview of how relevant topics were identified by TM-B. In the first row, bars represent the maximum probability each topic (x-axis) assigns to keywords. The second row highlights in dark the positions in the rank that contain keywords in each topic (x-axis). The purple horizontal bars represent the optimal γ and k, respectively, in our experiments (see Appendix for exact numerical values).

yses. For situations in which news of type `main` are more important than `mention`, the two best TMs and deep learning classifiers worked in like manner, not missing the vast majority of instances.

The majority voting approach seemed to draw out advantages from each model, achieving the best F1 score. Still, employing three computationally costly models for this task is hardly justifiable in practice, given the modest overall increase in the aggregated F1 score compared to single models.

Classification of news turned out to be hazard-dependant. There was no one-size-fits-all best solution across all hazards. The fine-tuned text embeddings balanced precision and recall well in general, but in three hazards (cold waves, storms and wildfires) the F1 score of the TM approaches was on par or better than it, which is a very interesting finding given that TMs are unsupervised and do not rely on the currently prevailing deep learning paradigm. Landslides and wildfires were the easiest to identify with all metrics above 0.75 (except for the LLM's precision for wildfire) in all models apart from TM-P. Cold and heat waves were the most challenging with suboptimal results even for the majority voting method.

Note, however, that comparisons between models should be done with caution, as these experimental estimates by hazard type were computed from samples of only 100 documents each. Rare events become very sensitive to individual predictions in such a small sample. For instance, cold waves contain only 17 relevant documents on which to measure precision and recall, so that a single swapped prediction by a model would already cause a 5.8% increase or decrease in recall.

7. Analysis

In this section, we explore TM's interpretability by providing more details on the TM-B models' behaviour. In our non-exhaustive hyperparameter

DROUGHT

dürre	mitte	trockenheit
notstand	niedrigwasser	waldbrandgefahr
ernteausfall	tag	feuchtigkeit
vieh	fisch	kanton
regenzeit	wasser	brandgefahr
ausmaß	stoff	ernteausfall
zentrum	fischsterben	nässe
versicherung	sand	stress
wasserreservoir	kreis	notstand
helmholtz	jugendliche	leiter

WILDFIRE

feuer	buschfeuer
brand	australisch
waldbränd	koala
flamme	bundesstaat
waldbrand	kontinent
hektar	buschbränd
feuerwehr	australier
region	tier
groß	villa
kontrolle	ostküste

Figure 5: Top 10 terms in each of the topics considered as relevant for drought and wildfire.

search, LDA achieved the best performance for six hazard types, while NMF was superior only for drought. The optimal thresholds θ for assigning topics to documents were between 0.028 and 0.076. Figure 4 illustrates how relevant topics were selected. The top terms decision method achieved the highest performance across five hazards with k values ranging from 1 to 5. Keyword proximity was superior only for drought and heat wave using γ =0.09 and 0.036, respectively. The number of selected topics for each hazard varied from 1 to 6.

Here we focus on wildfires and droughts as they had the smallest and largest differences in F1 score, respectively, in relation to deep learning strategies. For wildfires, topics were considered relevant if a hazard-related keyword was among their top 3 most probable terms. That resulted in 2 out of 50 topics being considered as relevant. For drought, keyword proximity selected 3 out of 500 topics as relevant. The top 10 lemmas representing these topics are shown in Figure 5. Figure 6 illustrates the effect of θ for wildfire's leftmost relevant topic: how well (not) relevant documents are classified based on the θ parameter for the rightmost relevant topic in fire: documents with topic probability above θ are classified as relevant, with a few wrong predictions.

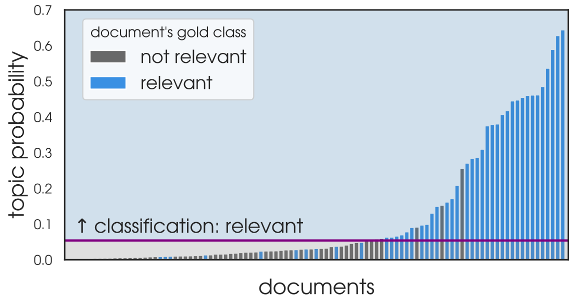

Figure 6: Example of the θ threshold for assigning a relevant fire topic to documents. Above the purple horizontal line, documents are classified as relevant, with a few mistakes with respect to the gold standard.

Wildfires Out of the 100 test instances, 83 were correctly classified. We inspected the 17 errors. The 7 false positives are texts that do refer to fires, but they are either not wildfires (e.g. fires in houses or industrial plants) or discuss technologies to combat wildfires. One of the documents describes a wildfire warning that also refers to an extinguished fire, which may have been missed by the annotators. Among the 10 false negatives, most contain mentions to wildfires occurring in discussion about other main topics (financial markets, heat waves, conferences) in documents that are a concatenation of various news articles (a problem we inherited from the original database were not able to fully solve automatically). In such cases, the fire-related topic may not have been salient enough to be assigned to the document. Although two topics were selected as relevant, only the left one was responsible for making all predictions on the test set. The second topic appeared in only one text, indicating that, although potentially relevant, it may have specialised too much during training. All documents of type `main` were correctly identified.

Drought For this hazard, 65 documents were correctly classified, with 14 false positives and 21 false negatives. Many false positives treat drought as a broad phenomenon rather than a concrete event, for example, when discussing drought-tolerant plants, vegetation stress, or climate change. Such cases are difficult to distinguish using TMs and may also reflect ambiguities in the annotation. False negatives show no clear patterns. Since this model relied on 500 topics, the drought concept may have been diffuse across multiple topics: in Figure 4, it is evident that drought keywords appear in various topics not selected as relevant. The only `main` document that was incorrectly classified has only one keyword, a compound noun (*Dürregebieten*), which was not included as a feature. Including all words *containing* keywords as features could have prevented this but it introduces additional noise from low-frequency terms that form topics.

8. General Discussion

This work was primarily motivated by the lack of a comprehensive global database of extreme climate disasters. Existing disaster databases, for instance the EM-DAT (Delforge et al., 2025), are shaped by reporting practices and inclusion thresholds (e.g. at least 10 fatalities), which have been widely discussed for their biased coverage toward large, well-documented events and wealthier regions, systematically under-representing some regions and hazards (Jones et al., 2023). Our method is designed to support bottom-up data-driven analyses by bypassing the inherent incompleteness and structural biases of top-down lists of worldwide extreme climate events (Gall et al., 2009). Our procedure permits the inclusion of news about events that did not meet the arbitrary inclusion criteria of disaster databases.

Rather than claiming the superiority of one model type for news classification, we have provided evidence that the results are hazard-dependent. This is an important finding for climate impact and adaptation research: the way different hazards are reported in the news varies, so solutions that treat all climate-related hazards as a single category (i.e. disasters in general) risk masking important performance variations, leading subsequent conclusions to be biased towards those that are easier to identify. In this context, an advantage of our approach is that we explicitly consider hazards separately, enabling more reliable downstream analyses.

The exact reasons for such differences require further investigation. First, each hazard is inherently distinct in the abruptness of its onset, its duration, its frequency, and its perceived severity. Then, media coverage can differ depending on socio-economic and geo-political factors. Finally, there is linguistic and discourse-related variation. For instance, while some keywords are very specific to climate events, others are polysemic and appear in multi-word expressions. The interdependencies among these layers are worth studying. Some events are hard to pinpoint even for humans, which can impact gold standards. Treating extreme climate events as a monolithic concept is thus not advisable in NLP tasks. Besides, since multi-hazard events occur in reality, another promising way forward is to analyse how they also co-occur in news.

We aimed to reduce human input in TM interpretation by selecting thresholds automatically and minimizing hyperparameter choices. Further work can investigate whether manual selection of keywords and topics can improve results. Our preliminary experiments with CorEx and Top2Vec yielded comparable results, so we prioritised the more traditional LDA/NMF methods in this study. However, other TM variations can be further investigated, including tuning priors to promote clearer keyword-related topics. The fine-tuned text embeddings achieved some of the highest F1 scores using only 350 documents and can potentially be further improved with active learning (Schröder and Niekler, 2020).

LLMs are being uncritically employed for many NLP tasks. We have shown that even a model with 14b parameters was not sufficiently precise. Our results add to the evidence that LLMs require careful evaluation as any other model. If LLM-based approaches are to be used, TMs can still be helpful in shrinking the amount of unrelated documents (e.g. by excluding those that have high probability for totally unrelated topics), thereby reducing the considerable environmental and financial costs of using LLMs.

9. Conclusion

We have presented a comparative analysis of three binary classifiers for refining collections of news articles on extreme climate events retrieved via keyword-based approaches. Although the LLM and the fine-tuned text embeddings had a higher F1 score in general, the drop in comparison to TMs was 0.148 on the worst case (drought) but also only 0.001 on the best case (wildfire). This is remarkable given TMs' unsupervised training and the simplicity of the keyword-guided topic selection process. Depending on the use case, this difference may be acceptable given other advantages, such as higher precision. Besides, the reason for deep learning-based predictions are beyond human comprehension, whereas decisions based on TMs are fully transparent and explainable.

Limitations

The rule-based filtering may have excluded relevant documents, although it was a price worth paying to reduce the immense volume of unrelated news and to keep the task computationally tractable. Although we are seeking to identify extreme climate events, other types of disasters (.e.g. urban fires and industrial accidents that cause dam collapse) could not yet be fully distinguished by our methods.

The test samples for each extreme climate event contain only 100 documents each, which may obscure variance in the estimates. More definitive claims about differences in models' behaviour require cross-validation and, ideally, a larger sample. The performance of the classifiers is bounded by the quality of the annotation. Despite best efforts, ambiguity is not always easy to resolve and arbitrary decisions can impact models' training and evaluation.

We presented results for varying TM set-ups as we opted for selecting the best-performing configurations. Still, keeping it constant would facilitate the direct comparison across hazards. The hyperparameter search for TM considered only a few dozen combinations of the number of topics, POS-tags and minimum document frequency. This can potentially be further refined for each hazard separately.

We did not perform extensive prompt engineering for the LLM, as these models are supposed to parse natural language instructions well; still, given their unpredictable nature, minor changes to the prompt might have led to different outcomes. Larger models may yield better results, but our focus here was on lower-scale, local solutions.

Acknowledgements

We thank Marc Keuschnigg for his contribution in conceptualising the research project that motivates this paper, as well as Maike Reichel and Julius Hehenkamp for their help in annotating the data. We also thank the anonymous reviewers for their valuable feedback.

10. Bibliographical References

Pedro Henrique Lima Alencar, Jan Sodoge, Eva Nora Paton, and Mariana Madruga De Brito. 2024. Flash droughts and their impacts—using newspaper articles to assess the perceived consequences of rapidly emerging droughts. *Environmental Research Letters*, 19(7):074048.

Aditya Anantharaman, Arpit Jadiya, Chandana Tulasi Sai Siri, Bharath NVS Adikar, and Biju Mohan. 2019. Performance evaluation of topic modeling algorithms for text classification. In *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 704–708.

Dimo Angelov and Diana Inkpen. 2024. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, Florida, USA. Association for Computational Linguistics.

Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. Analyzing the online communication of environmental movement organizations: NLP approaches to topics, sentiment, and emotions.

In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 68–76, Tallinn, Estonia. University of Tartu Library.

Valentina Tretti Beckles and Adrian Vergara Heidke. 2025. Thematic categorization on pineapple production in Costa Rica: An exploratory analysis through topic modeling. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 44–55, Tallinn, Estonia. University of Tartu Library.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Erica Cai, Xi Chen, Reagan Grey Keeney, Ethan Zuckerman, Brendan O'Connor, and Przemyslaw A. Grabowicz. 2025. Identifying and investigating global news coverage of critical events such as disasters and terrorist attacks. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):307–323.

Tais Maria Nunes Carvalho, Andreas Niekler, Christian Kuhlicke, Jakob Zscheischler, and Mariana Madruga de Brito. 2025. Global synthesis of peer-reviewed articles reveals blind spots in climate impacts research. Preprint, available at Research Square.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. 2015. Dataless text classification with descriptive lda. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Damien Delforge, Valentin Wathelet, Regina Below, Cinzia Lanfredi Sofia, Margo Tonnelier, Joris A.F. van Loenhout, and Niko Speybroeck. 2025. Em-dat: the emergency events database. *International Journal of Disaster Risk Reduction*, 124:105509.

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2024. Keyword-assisted topic models. *American Journal of Political Science*, 68(2):730–750.

Melanie Gall, Kevin A. Borden, and Susan L. Cutter. 2009. When do losses count?: Six fallacies of natural hazards loss data. *Bulletin of the American Meteorological Society*, 90(6):799 – 810.

Ryan J. Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. 2017. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Francesca Grasso, Ronny Patz, and Manfred Stede. 2024. NYTAC-CC: A climate change subcorpus of New York Times articles. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 403–409, Pisa, Italy. CEUR Workshop Proceedings.

Swapnil Hingmire, Sandeep Chougule, Girish K. Palshikar, and Sutanu Chakraborti. 2013. Document classification by topic labeling. SIGIR '13, page 877–880, New York, NY, USA. Association for Computing Machinery.

Rebecca Louise Jones, Aditi Kharb, and Sandy Tubeuf. 2023. The untold story of missing data in disaster research: a systematic review of the empirical literature utilising the emergency events database (em-dat). *Environmental Research Letters*, 18(10):103006.

Inhye Kong and Ross S. Purves. 2026. Analyzing geographic bias of newspaper articles reporting global climate disasters. *Annals of the American Association of Geographers*, 116(2):270–288.

Murathan Kurfali, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. 2025. ClimateEval: A comprehensive benchmark for NLP tasks related to climate change. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 194–207, Vienna, Austria. Association for Computational Linguistics.

Markus Leippold and Francesco Saverio Varini. 2020. Climatext: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Alexandra Lesnikowski, Ella Belfer, Emma Rodman, Julie Smith, Robbert Biesbroek, John D. Wilkerson, James D. Ford, and Lea Berrang-Ford. 2019. Frontiers in data analytics for adaptation research: Topic modeling. *WIREs Climate Change*, 10(3):e576.

Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. 2016a. Effective document labeling with very few seed words: A topic model approach. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 85–94, New York, NY, USA. Association for Computing Machinery.

N. Li, W. Thiery, S. Zahra, M. Madruga de Brito, K. Worou, M. Kurfalı, S. Lampe, P. Muñoz,

C. Flynn, C. Trigoso, J. Nivre, J. Zscheischler, and G. Messori. 2025. Wikimpacts 1.0: A new global climate impact database based on automated information extraction from wikipedia. *EGUsphere*, 2025:1–43.

Ni Li, Shorouq Zahra, Mariana Brito, Clare Flynn, Olof Görnerup, Koffi Worou, Murathan Kurfali, Chanjuan Meng, Wim Thiery, Jakob Zscheischler, Gabriele Messori, and Joakim Nivre. 2024. Using LLMs to build a database of climate extreme impacts. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 93–110, Bangkok, Thailand. Association for Computational Linguistics.

Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. 2018. Dataless text classification: A topic modeling approach with document manifold. CIKM '18, page 973–982, New York, NY, USA. Association for Computing Machinery.

Zhenzhong Li, Wenqian Shang, and Menghan Yan. 2016b. News text classification model based on topic model. In *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pages 1–5.

Mariana Madruga de Brito, Jan Sodoge, Heidi Kreibich, and Christian Kuhlicke. 2024. Comprehensive assessment of flood socioeconomic impacts through text-mining. *Water Resources Research*, 61(1).

Mariana Madruga de Brito, Jan Sodoge, Heidi Kreibich, and Christian Kuhlicke. 2025. Comprehensive assessment of flood socioeconomic impacts through text-mining. *Water Resources Research*, 61(1):e2024WR037813.

Daniel Maier, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch, Gerhard Heyer, Ueli Reber, Thomas Häussler, et al. 2021. Applying lda topic modeling in communication research: Toward a valid and reliable methodology. In *Computational methods for communication science*, pages 13–38. Routledge.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Timothy Miller, Dmitriy Dligach, and Guergana Savova. 2016. Unsupervised document classification with informed topic models. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 83–91, Berlin, Germany. Association for Computational Linguistics.

Taís Maria Nunes Carvalho, Francisco De Assis De Souza Filho, and Mariana Madruga De Brito. 2024. Unveiling water allocation dynamics: a text analysis of 25 years of stakeholder meetings. *Environmental Research Letters*, 19(4):044066.

Aytuğ Onan, Serdar Korukoğlu, and Hasan Bulut. 2016. Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, 57:232–247.

Telma Peura, Attila Krizsán, Salla-Riikka Kuusalu, and Veronika Laippala. 2025. Perspectives on forests and forestry in Finnish online discussions - a topic modeling approach to suomi24. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 10–15, Tallinn, Estonia. University of Tartu Library.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. 2012. Statistical topic models for multi-label document classification. *Machine learning*, 88(1):157–208.

Efsun Sarioglu, Kabir Yadav, and Hyeong-Ah Choi. 2013. Topic modeling based classification of clinical reports. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 67–73, Sofia, Bulgaria. Association for Computational Linguistics.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Schröder and Andreas Niekler. 2020. A survey of active learning for text classification using deep neural networks. ArXiv preprint: 2008.07267.

Sattar Seifollahi, Massimo Piccardi, and Alireza Jolfaei. 2021. An embedding-based topic model for document classification. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(3).

Jan Sodoge, Christian Kuhlicke, and Mariana Madruga De Brito. 2023. Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, 41:100574.

Manfred Stede, Yannic Bracke, Luka Borec, Neele Charlotte Kinkel, and Maria Skeppstedt. 2023. Framing climate change in nature and science editorials: applications of supervised and unsupervised text categorization. *Journal of Computational Social Science*, 6(2):485–513.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. ArXiv preprint: 2209.11055.

Kerstin K Zander, Hunter S Baggen, and Stephen T Garnett. 2023. Topic modelling the mobility response to heat and drought. *Climatic Change*, 176(4):42.

Daochen Zha and Chenliang Li. 2019. Multi-label dataless text classification with topic modeling. *Knowledge and Information Systems*, 61(1):137–160.

A. Appendix

Further details about preprocessing and filtering

The lists of German keywords used for each hazard are shown in Figures 8, 9, and 10. Although *derecho* was included as a keyword initially, texts containing only this keyword were removed in a preprocessing step. The inclusion of *Regenfälle* (rainfalls) as a keyword for flood resulted in the inclusion of some texts that may not be about floods.

Here, we provide a summary of the preprocessing and filtering steps applied to the original document collection. Exact implementation details are documented in the preprocessing code which is available upon request.

We removed exactly duplicated instances, i.e. those pairs or groups of documents for which *all* metadata values were exactly the same. Documents with the same text but published by different outlets or on different dates were kept as they count separately towards media attention.

The regex pattern `'<.*?>'` was used to remove reminiscent html content. To split (at least part of) the documents that have been concatenated as a single instance, despite being composed of several different pieces of news, we used another regex pattern with frequent news agencies abbreviations (e.g. dpa and afp) that often appeared in parentheses in between such concatenations.

Approximately duplicated texts were identified using the MinHash algorithm to estimate Jaccard similarity, with a threshold of 0.8 slightly more conservative than the empirical choice in (Madruga de Brito et al., 2024). This was not used to exclude any document, but helped ensure that annotators did not annotate the same text (from different news sources) for the same hazard and that TMs were not trained on similar texts that would form spurious clusters.

Spacy's German model `de_core_news_lg` was used to parse each text and retrieve tokens and sentence counts.

We also applied filters to reduce the number of unrelated documents and local news. The inclusion criteria were as follows:

- The document contains at least one keyword related to its assigned hazard. Although this was an imposed criterion for the database query, after splitting concatenated documents, were a few cases of texts that no longer contained keywords.
- The document contains at least one keyword of its assigned hazard which is *not* an intruder.
- The document's outlet reportedly belongs to the German press.
- In case of exact duplicates (regarding all fields), only one instance was kept.
- The document's ressort does not contain the word `lokal`, since we are only interested in international extreme climate events.
- The number of tokens is at least 30 and no more than 1,700. The thresholds were selected based on empirical observations of the distribution's histogram and by taking into account an initial batch of annotated documents.
- The document contains at least one of the following: a country name, a nationality (as an adjective or a noun) or a city name.
- The first token of the document is not the name of a German city followed by a full stop.
- The proportion of non-alphabetical characters is less than 0.11. The threshold was selected based on empirical observations of the distributions, also considering an initial batch of annotated documents.

Further details about the annotation

Figure 7 shows the instructions given to the two annotators. They also identified the sentences that refer to each type of hazard and the country where it occurred. These variables will be used in future studies.

Further details about the classifiers

LLM Figures 8, 9, and 10 show the values used to fill in the hazard-dependent slots in the prompt for the LLM, which is shown in Figure 11. Definitions were translated from the EM-DAT' glossary[7], except for storm, not defined by EM-DAT, for which we used Wikipedia[8].

Topic Models The number of iterations and passes were fixed at 400 and 20, respectively. The random seed was set to 123. Table 4 shows the number of documents used to train the TMs for each extreme climate event, i.e. the unique texts in the collection. Table 5 shows the selected hyperparameters for the topic model configuration of each extreme climate event. In Table 6 we show all topics selected as relevant for TM-B characterised by their top 10 terms with the highest probability.

cold	91,140
drought	96,441
flood	334,607
heat	197,800
landslide	45,039
storm	488,068
fire	78,426

Table 4: Number of documents used to train the topic models for each hazard.

[7] https://doc.emdat.be/docs/data-structure-and-content/glossary/

[8] https://en.wikipedia.org/wiki/Storm

Annotation Task: Identifying news about natural hazards

You are seeing a collection of news published in the German media. We need to identify documents that are discussing **concrete, specific, natural** hazards events **abroad**. Each document has been pre-assigned to a hazard that it is possibly discussing. There are seven hazards: flood, storm, heat wave, cold wave, fire, drought and landslide/mass movement. Your task is to annotate each document with three types of information:

Decision 1: Main document label

The first step is to identify whether the document's topic is or not about the specific event it was already assigned. You will see the document and the possible category. For example, let's assume that a document is supposed to be about a flood. Pick one of the four options:

- `unrelated`: if the document is not related to any concrete natural flood event, select the option `no, the document is not about any concrete natural flood event`.
- `main`: if the main purpose of the document is to report about one or more concrete natural flood events, select the option `yes, it's the document's main topic`.
- `mention`: if there are one or more sentences mentioning concrete natural flood events, but they are not the main topic of the document (or, in other words, if it does not seem that the main intent of the journalist when writing this document was to report on those specific events), select the option `yes, but only a mention`.
- `mention-concatenated`: unfortunately, some documents in our database are made of a list of concatenated pieces of news that should have been different instances. We applied some preprocessing steps to prevent it. Still, if you identify a case where a flood is discussed in a document that is composed of various documents on different topics, select `yes, but among other concatenated documents` instead.

The word *flood* will be replaced by the corresponding hazard in each document.

Details

- We are only interested in news about specific events, i.e. events that have taken place at some location and some point in time (droughts may be harder in this regard, but still, we want documents about concrete droughts that have happened or were happening at the moment of publication).
- We are only interested in **natural** events. In particular for fires: wildfires are relevant, but e.g. a fire in an industrial plant caused by an explosion is not.
- Documents discussing impacts of natural hazards in general, the possibility or prediction of a hazard that has not yet happened, scientific studies about hazards, metaphorical uses, etc are not relevant.
- Mentions that are too generic are also not relevant for us, for instance *"Climate change has increased the occurrence of floods in Brazil"*. Such cases can be classified as `unrelated`.

Decision 2: Location

If you select any of the yes-options for the first question, please also select the country where the event you identified happened (even if it is Germany). You can select more than one country if needed.

Decision 3: Type of natural hazard in text spans

Please use the hazard labels to highlight all concrete natural hazard being mentioned/reported in the text. We are interested in the following seven hazards (use `other` for any other types):

- `drought`, `flood`, `storm`, `heatwave`, `coldwave`, `mass movement`, `wildfire`, `other`

Figure 7: Instructions given to the annotators.

		model	decision method	θ	γ or k	min. doc freq	tags	topics
TM-F1	landslide	lda	top terms	0.058	5.000	100	noun, verb, adj	100
	fire	lda	top terms	0.040	3.000	50	noun, verb, adj	50
	flood	lda	top terms	0.060	2.000	1000	noun, verb, propn	50
	storm	lda	top terms	0.024	2.000	10000	noun, verb, adj	50
	drought	nmf	keyword proximity	0.016	0.198	500	noun	500
	heat	lda	keyword proximity	0.024	0.108	100	noun, verb, propn	100
	cold	lda	top terms	0.024	5.000	500	noun, verb, adj	300
TM-B	landslide	lda	top terms	0.076	5.000	100	noun, verb, adj	100
	fire	lda	top terms	0.054	3.000	50	noun, verb, adj	50
	flood	lda	top terms	0.062	2.000	1000	noun, verb, propn	50
	storm	lda	top terms	0.030	1.000	10000	noun, verb, adj	50
	drought	nmf	keyword proximity	0.028	0.090	500	noun	500
	heat	lda	keyword proximity	0.052	0.036	100	noun, verb, adj	50
	cold	lda	top terms	0.028	5.000	500	noun, verb, adj	300
TM-P	landslide	lda	keyword proximity	0.064	0.054	50	noun, verb, adj	100
	fire	lda	top terms	0.064	1.000	500	noun, propn	100
	flood	lda	top terms	0.054	1.000	1000	noun, verb, propn	50
	storm	lda	top terms	0.120	5.000	10000	noun, verb, adj	50
	drought	nmf	top terms	0.034	1.000	500	noun, verb, adj	500
	heat	lda	top terms	0.148	2.000	500	noun, verb, propn	50
	cold	lda	keyword proximity	0.146	0.036	5000	noun, verb, adj	50

Table 5: Hyperparameters of all topic models that produced the presented results.

```
landslide:

    hazard: Erdrutsch

    keywords: Erdrutsch, Felssturz, Felsstürz, Schlammlawine, Massenbewegung,
        Hangrutsch, Hangbewegung, Rutschung, Bodenrutsch, Hangabrutschung, Murgang
        , Gerölllawine, Rutschhang, Rutschhäng, Rutschgefahr, Felslawine, Mure

    hazard_event: Erdrutschereignisse

    definition: Jede Art von mäßiger bis schneller Bodenbewegung, einschließlich
        Lahare, Schlammlawinen und Murgänge (unter trockenen/nassen Bedingungen).
        Ein Erdrutsch ist eine durch die Schwerkraft gesteuerte Bewegung von Erde
        oder Gestein, deren Geschwindigkeit in der Regel zwischen langsam und
        schnell liegt, jedoch nicht sehr langsam ist. Er kann oberflächlich oder
        tief sein, aber das Material muss eine Masse bilden, die einen Teil des
        Hangs oder den Hang selbst ausmacht. Die Bewegung muss nach unten und nach
         außen mit einer freien Fläche erfolgen. ODER Jede Art von Abwärtsbewegung
         von Erdmaterialien unter hydrologisch trockenen Bedingungen. ODER Arten
        von Massenbewegungen, die auftreten, wenn starker Regen oder schnelle
        Schnee-/Eisschmelze große Mengen an Vegetation, Schlamm oder Gestein unter
         dem Einfluss der Schwerkraft einen Hang hinunterbefördern.

fire:

    hazard: Waldbrand

    keywords: Flächenbrand, Flächenbränd, Waldbrand, Waldbränd, Wildfeuer,
        Landschaftsbrand, Landschaftsbränd, Buschfeuer, Vegetationsbrand,
        Vegetationsbränd, Naturbrand, Naturbränd, Großbrand, Großbränd, Forstbrand
        , Forstbränd, Heidebrand, Heidebränd

    hazard_event: Waldbrandeereignisse

    definition: Jede unkontrollierte und nicht vorgeschriebene Verbrennung oder
        das Abbrennen von Pflanzen in einer natürlichen Umgebung wie Wald,
        Grasland, Buschland oder Tundra, die natürliche Brennstoffe verbraucht und
         sich aufgrund von Umweltbedingungen (z. B. Wind oder Topografie)
        ausbreitet. Waldbrände können durch Blitzeinschläge oder menschliches
        Handeln ausgelöst werden.
```

Figure 8: Keywords and values used in the prompts for each hazard (1/3).

```
cold:

    hazard: Kältewelle

    keywords: Kältewelle, Kälteeinbruch, Kältestress, extreme Kälte, extremer Kä
        lte, extremen Kälte, Frost, strenger Winter, Wintereinbruch, Wintereinbrü
        ch, Kälteperiode, Kälterekord, arktische Kälte, arktischer Kälte,
        arktischen Kälte, Kältewarnung, Eisregen, Eiseskälte, Schneechaos

    hazard_event: Kältewellenereignisse

    definition: Eine Periode mit ungewöhnlich kaltem Wetter. In der Regel dauert
        eine Kältewelle zwei oder mehr Tage und kann durch starke Winde noch verst
        ärkt werden. Die genauen Temperaturkriterien für eine Kältewelle können je
         nach Standort variieren.

heat:

        hazard: Hitze

        keywords: Hitze, extreme Temperaturen, extremen Temperaturen,
            Temperaturrekord, Tropennacht, Tropennächt, überhitzung, Rekordhitze,
            Hitzetag

        hazard_event: Hitzewellenereignisse

        definition: Eine Periode mit ungewöhnlich heißem und/oder ungewöhnlich
            feuchtem Wetter. In der Regel dauert eine Hitzewelle zwei oder mehr
            Tage. Die genauen Temperaturkriterien für eine Hitzewelle können je
            nach Standort variieren.

drought:

    hazard: Dürre

    keywords: Dürre, Rekorddürre, Trockenperiode, Trockenheit, Wasserknappheit,
        Niedrigwasser, Wassermangel, Niederschlagsmangel, Niederschlagsdefizit,
        Bodenfeuchte-Defizit, Bodenfeuchteverlust

    hazard_event: Dürreereignisse

    definition: Ein längerer Zeitraum mit ungewöhnlich geringen Niederschlägen,
        der zu einer Wasserknappheit für Menschen, Tiere und Pflanzen führt. Dü
        rren unterscheiden sich von den meisten anderen Gefahren dadurch, dass sie
         sich langsam, manchmal sogar über Jahre hinweg, entwickeln und ihr Beginn
         in der Regel schwer zu erkennen ist. Dürren sind nicht nur ein
        physikalisches Phänomen, da ihre Auswirkungen durch menschliche Aktivitä
        ten und den Wasserbedarf noch verstärkt werden können. Dürren werden daher
         oft sowohl konzeptionell als auch operativ definiert. Operative
        Definitionen von Dürre, d. h. der Grad der Niederschlagsverringerung, der
        eine Dürre ausmacht, variieren je nach Ort, Klima und Umweltbereich.
```

Figure 9: Keywords and values used in the prompts for each hazard (2/3).

```
flood:

    hazard: Hochwasser

    keywords: Überschwemmung, Flut, Hochwasser, Überflutung, Flusshochwasser,
        Regenfälle, Sturzflut, Gletscherseeausbruch, Gletscherseeausbrüch,
        Gletschersee-Ausbruchsflut, Jahrhunderthochwasser

    hazard_event: Hochwasserereignisse

    definition: Ein allgemeiner Begriff für das Überlaufen von Wasser aus einem
        Flussbett auf normalerweise trockenes Land in der Aue (Flussüberschwemmung
        ), überdurchschnittlich hohe Wasserstände entlang der Küste (Küstenü
        berschwemmung) und in Seen oder Stauseen sowie Wasseransammlungen an oder
        in der Nähe des Ortes, an dem der Regen gefallen ist (Sturzfluten).

storm:

    hazard: Sturm

    keywords: Sturm, Stürm, Unwetter, Orkan, Blizzard, Derecho, Hagel, Zyklon,
        Gewitter, Tornado, Mikroburst, Hurrican, Hurrikan, Taifun, Blizzard

    hazard_event: Sturmereignis

    definition: Ein Sturm ist jeder gestörte Zustand der natürlichen Umwelt oder
        der Atmosphäre eines astronomischen Körpers. Er kann durch erhebliche Stö
        rungen der normalen Bedingungen gekennzeichnet sein, wie z. B. starker
        Wind, Tornados, Hagel, Donner und Blitz (Gewitter), starke Niederschläge (
        Schneesturm, Regensturm), starker Eisregen (Eissturm), starke Winde (
        tropischer Wirbelsturm, Sturm), Wind, der bestimmte Substanzen durch die
        Atmosphäre transportiert, wie z. B. bei einem Staubsturm, sowie andere
        Formen von Unwettern.
```

Figure 10: Keywords and values used in the prompts for each hazard (3/3).

```
Du bist ein Experte für die Klassifikation von Nachrichtenartikel bezüglich der
    Existenz von Referenzen auf $hazard und extreme $hazard_event.

Definition von $hazard: $definition

Synonyme für $hazard: $keywords.

Die Nachrichtenartikel müssen mit einem dieser Labels klassifiziert werden:

– Label 1: Das Dokument behandelt $hazard, extreme $hazard_event oder damit
    verbundene Auswirkungen.
– Label 0: Das Dokument hat KEINE Verbindung zu $hazard oder extremen $hazard_
    event.

Es sind nur konkrete, spezifische Naturereignisse in der realen Welt relevant.
    Artikel, die sich lediglich mit der Möglichkeit eines Ereignisses befassen,
    metaphorische Verwendungen, allgemeine Diskussionen über die Art der Gefahr
    oder Ereignisse sind nicht relevant.

Analysiere den Inhalt des Dokuments Satz für Satz sehr sorgfaltig und vergib das
     Label 1 auch wenn nur ein Satz im Dokument relevant ist.

Entscheide, welches Label für diesen Nachrichtenartikel das richtige ist, und
    beginne deine Antwort entsprechend mit 0 oder 1.

Klassifiziere den folgenden Nachrichtenartikel:

$text
```

Figure 11: Prompt used for the LLM experiment.

		1	2	3	4	5	6	7	8	9	10
landslide	98	see	bergbau	erdrutsch	wasser	rutschung	tagebau	gefahr	siedlung	erklären	bereich
	99	tal	felssturz	fels	stein	gestein	meter	kubikmeter	berg	stürzen	groß
	80	tote	erdrutsch	zahl	leiche	bergen	opfer	begraben	vermissen	vermisst	verschütten
	22	mensch	erdrutsch	leben	überschwemmung	haus	schwer	behörde	heftig	sterben	region
	60	haus	bewohner	unglück	erdrutsch	gebäude	wohnung	wohnen	früh	bürgermeisterin	ursache
wildfire	1	feuer	brand	waldbränd	flamme	waldbrand	hektar	feuerwehr	region	groß	kontrolle
	26	buschfeuer	australisch	koala	bundesstaat	kontinent	buschbränd	australier	tier	villa	ostküste
flood	43	mensch	überschwemmung	haus	leben	region	dpa	behörde	stadt	land	angabe
	20	hochwasser	sachsen	elbe	dresden	polen	pegel	meter	donau	brandenburg	tschechien
storm	16	unwetter	wasser	hochwasser	überschwemmung	heftig	region	schwer	regen	betreffen	schaden
	13	gewitter	wetter	regen	blitz	absagen	regnen	mark	himmel	heftig	schlecht
	7	hurrikan	sturm	bundesstaat	treffen	kilometer	wirbelsturm	land	schaden	windgeschwindigkeit	stunde
drought	457	dürre	notstand	ernteausfall	vieh	regenzeit	ausmaß	zentrum	versicherung	wasserreservoir	helmholtz
	246	mitte	niedrigwasser	tag	fisch	wasser	stoff	fischsterben	sand	kreis	jugendliche
	319	trockenheit	waldbrandgefahr	feuchtigkeit	kanton	brandgefahr	ernteausfall	nässe	stress	notstand	leiter
heat wave	22	grad	temperatur	hitze	celsius	tag	hitzewelle	sommer	liegen	wetter	mensch
cold wave	1	mensch	sterben	leben	kältewelle	erfrieren	behörde	obdachlose	tote	zahl	dutzend
	194	schneefall	heftig	stark	schneechaos	teil	fallen	sperren	schneemasse	verkehr	blockieren
	137	eisregen	verspätung	reisend	glatt	mittag	behindern	glätte	vereist	schiene	störung
	12	winter	kalt	mild	wehen	östlich	wintermonat	stark	atlantik	flachland	luft
	111	kälte	warm	wärme	eisig	frieren	decke	kleidung	thermometer	anziehen	klirrend
	119	wetter	regen	kälteeinbruch	regnen	wetterlage	kalender	bauernregel	wetterstation	eisheilig	schlecht

Table 6: Top 10 terms of each topic considered as relevant in the TM-B experiments. The second column contains an arbitrary topic ID.