

Natural hazard database from Internet publications: text mining with a large language model

Anna DERKACHEVA, Maria SAKIRKINA, Gleb KRAEV, Tatiana ANISKINA

HSE University, Moscow, Russia – 2026

Natural hazard; database; large language model; text mining

Abstract

Comprehensive data on natural hazards and their consequences are crucial for effective risk assessment, adaptation planning, and emergency response. However, many countries face challenges with fragmented, inconsistent, and inaccessible data, particularly regarding local-scale events. To address this data gap in Russia, we developed an end-to-end processing pipeline that scrapes news from various online sources, including national newspapers, regional emergency office websites, and social media, and extracts hazard-related information into a structured database. The pipeline employs ChatGPT 4o mini for its cost effectiveness, quality performance and the ready-to-use mode. It identifies 14 information nuggets, including hazard type, location, dates, impact on infrastructure and people, and response measures. To enhance the result reliability, the outputs of a generative model are controlled by non-generative approaches of text processing. The resulting data are spatially located using OpenStreetMap service. In a test run, we scraped over 8.3 million primary news articles on 2018-2024, resulting in nearly 49,000 hazard-relevant texts that were related to 21,000 unique events across 18 hazard types. Evaluation against expert-labeled texts indicates recall and precision rates from 0.70 to 0.99 depending on the information rubric, with no instances of model hallucination. The combination of a ready-to-use large language model for text mining and the world-wide open-source geodata for geolocation supports the easily transfer of the pipeline to other regions and languages.

Cite as: Derkacheva A., Sakirkina, M., Kraev, G., & Aniskina, T. (2026). *Natural hazard database from Internet publications: Text mining with a large language model.* HSE University. Moscow, Russia.

1. Introduction

Collecting and organizing data on natural hazards (NH) is fundamental to study, prioritize and forecast them. These data support research, practical tasks, policy making, and evaluation of the risk reduction efforts (UNDRR 2021; UNDRR 2025; UNDRR 2015; The World Bank 2016; Mazhin et al. 2021).

NH databases (NHDBs) are typically maintained by government agencies (Shamin et al. 2019) or independent parties, like scientific institutes (e.g. EM-DAT by CRED), insurance companies (NatCatSERVICE by MunichRe, Sigma by SwissRe), and individuals (Petrova 2009). Long-term databases often rely on internal government reports (Zuzak et al. 2022; Ministry of the Russian Federation for Civil Defence 2021), ground monitoring networks (Shamin et al. 2019; U.S. Geological Survey 2025) or newspapers (Guzzetti et al. 1994). Since many stakeholders fit NHDBs to build internal work processes, database structures and inclusion rules vary (Wirtz et al. 2014; Mazhin et al. 2021; Kron et al. 2012). Common practices include limiting the hazard list (Froude & Petley 2018; Brakenridge et al. 2009), setting thresholds for hazard intensity (U.S. Geological Survey 2025; Shamin et al. 2019) or the resulting damage (EM-DAT, NatCatSERVICE). The wider the geographical coverage of NHDB, the higher thresholds are used, thus making the international NHDBs less sensitive to local events. Same time, the data systematically produced by domestic institutions could remain in restricted access due to legal regulations or commercial intentions (Sudnitsyna & Shikhov 2024). As a result, finding actual, structured, and harmonized NH data about the non-catastrophic and local events may remain difficult. This became especially highlight in recent years since UNDRR proposed to uses such databases to assess progress in risk reduction (UNDRR 2021; UNDRR 2015).

To address the lack of open data, scientists and humanitarians increasingly compile databases using publicly available media. With the rise of Internet, digital publications and social media have become an important data source. Despite the spatio-temporal (Carrara et al. 2003; Kron et al. 2012; Raška et al. 2014) and contextual (Melnikova 2019) biases of disaster representation, this sources are used to collect historical data on occurrence, infrastructure damage, and loss of life (e.g. Petrova 2009; Froude and Petley 2018) or to support crisis management of an ongoing disaster (e.g. Earle et al. 2010; de Bruijn et al. 2019). For the latter task, live-chatting platforms like Twitter are more useful (Imran et al. 2015; Atefeh & Khreich 2015). Traditional news communications with more detailed texts are more appropriate for structured databases (Raška et al. 2014; Kron et al. 2012; Imran et al. 2013). In many countries, including Russia, such databases are mainly compiled via manual news mining (Chernokulsky et al. 2020; Sudnitsyna & Shikhov 2024; Shnyparkov & Gryaznova n.d.; Petrova 2009; Shikhov & Bykov 2014), focusing mostly on large-covering hydrological and meteorological hazards rather than localized phenomena (Raška et al. 2014).

In this Russia-based study, we aimed to create a damage-oriented NHDB from news texts that covers a broad range of phenomena, includes events of any intensity if they worry people, enables native geolocation of the events on a map, records socio-economic impact and responses, and is fast

and easy to update. Despite the widely expected better quality of manual text mining in such an ambiguous area (Kron et al. 2012; Froude & Petley 2018), it is time- and cost- intensive. To handle large volumes of publications, we designed an automated processing pipeline using a generative language model for text mining. The pipeline collects news across the Internet, extracts the information nuggets, and populates a database. The work was done on Russian-language content, but the model choice and pipeline structure allow easy and low-cost adaptation for other regions and languages.

2. Scope of the creating pipeline

Thirty natural phenomena were predefined for recognition (Supplementary Table S1). The list includes meteorological, hydrological, shallow geohazards, and environmental degradation (UNDRR 2020). The list is based on a checklist recommended for risk assessment in Russian climate adaptation policy (Ministry of Economic Development of the Russian Federation 2021). To prevent the misclassification due to mislabeling by non-specialists (Chernokulsky et al. 2020), we grouped the close terms, e.g., “hurricane”, “tornado”, and “storm” are classified as “strong wind”.

There are a number of issues and limitations common for the news-based databases of NH – spatial and social bias, information dispersion across several news, homonyms in geographical names, etc. (Kron et al. 2012). Before developing the automated pipeline, we manually compiled a database prototype to examine the potential influence of such issues on the LLM-driven approach and to design the corresponding solutions. Several experts labeled a hundred of news items reflected the variety of NH types, geography, text complexity, publication timing, etc. The key outcoming decisions are highlighted in the Methodology and Discussion when applicable.

3. Methodology

This section describes the automated processing pipeline to generate a NHDB from the digital text news (Fig. 1). Its general structure resembles similar efforts (Yzaguirre et al. 2015; Enikeeva et al. 2016; de Bruijn et al. 2019; van den Homberg et al. 2022). First, the news items are downloaded from the selected sources. The corps of texts is filtered to keep communications on natural hazards only. Then, the required information nuggets are extracted to form NHDB entry. Each entry is geocoded and mapped. Finally, the news-based entries, reporting about the same event, are grouped. The entire programming code is done with Python.

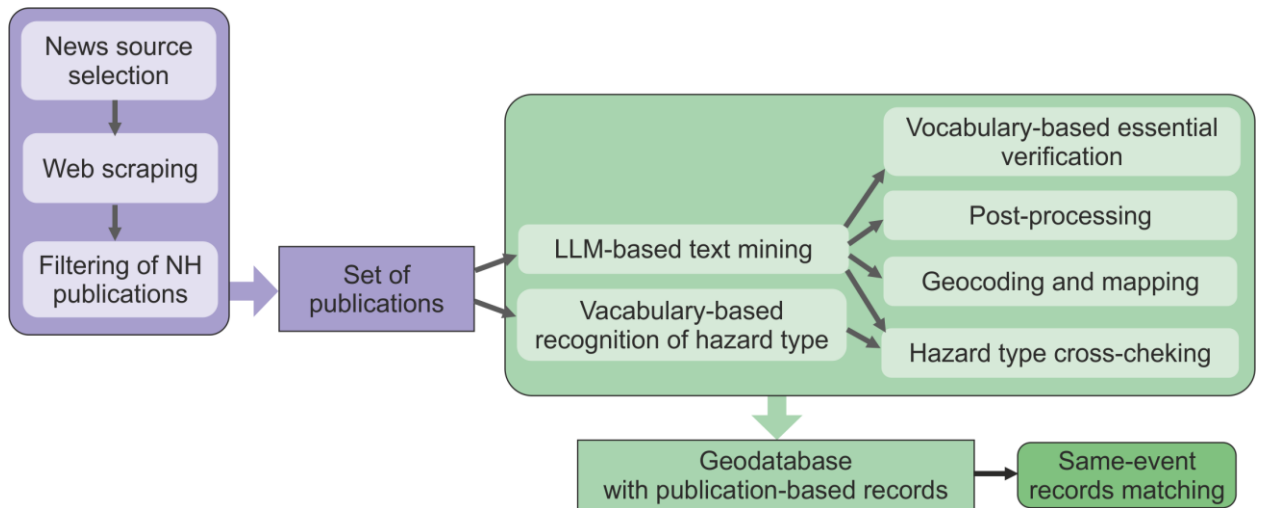


Fig. 1 Overview of an automated pipeline producing a damage-oriented natural hazards database from text news.

3.1 Source data acquisition

The search and downloading of relevant news items in an automated mode, rather than manual, allows a significant increase in the speed of those processes and thereby to process a larger number of information sources and publications. We organize this work as follows. First, we manually identify the sources (internet sites) and the subsections that most probably contain the relevant news. Using the web-scraping techniques, all news items available in a target site subsection are downloaded for a selected period of time. Finally, the downloaded mass of news is filtered to get only NH-related items. To avoid data privacy limitations, we used only freely available online publications.

3.1.1 Web-scraping

Web-scraping is a process of content extracting from web pages. We defined the following criteria for the potential news sources: well-structured HTML code preserved over years, clear URL elements of navigation, stable archive with the past content. Three source types identified for a project are national newspapers, the *Vkontakte* social network, and regional EMERCOM branch sites. Such a choice is discussed further in Section 4.1. The downloaded content per news item includes source name, title, main body of text, publication date, and link.

3.1.2 Filtering of downloaded news items

At the web-scraping step all texts in a given site section are downloaded regardless their topic and only a small fraction is related to NH. To save the language model resources at the text mining step, the news are filtered first for relevance to the topic. We opted on regular expression (RE) approach, which proved efficient results for such a task (van den Homberg et al. 2022; Taylor et al. 2015; Yzaguirre et al. 2015; Froude & Petley 2018; Enikeeva et al. 2016) with low resource costs: it identifies patterns of character sequences in text, including flexible word forms. The lists of desirable and restricted patterns

were developed by experts based on manual news mining, domain expertise, and the tests on a hundred of news.

Testing on 300 random news articles from various sources yielded a recall of 0.92 and precision of 0.84, prioritizing clean sampling over quantity and matching results comparable to language model filtering. Those scores are comparable to a sophisticated language model (de Bruijn et al. 2019) and were acceptable for us.

3.2 Transformation of news text to database entry

3.2.1 Large Language Model makes news a database entry

The core of the text processing is the extraction of information nuggets from the texts and creation of the NHDB entries. Several types of text-processing language models exist. Named Entity Recognition (NER) searches and extract the target elements of the pre-defined classes. Natural Language Processing (NLP) extract some content according to the training examples without understanding of broad context. In turn, Large Language Model (LLM) analyzes a text along with its sense context and generates a responding text, e.g. summary about specific part of information. As compiling an NHDB by experts is a cognitively intensive task, we focused on a LLM, despite the higher resource cost and the previous attempt with NER and NLP.

A ready-to-use solution was privileged instead of an open-source LLM that had to be tuned and run on on-board computation resources. We adopted the commercial ChatGPT-4o-mini (OpenAI, USA) model because it demonstrated the best price-quality ratio among several models tested (see Supplementary Section S2). Average processing price for our data was about 1\$ per 1000 news. The quality estimates is given in Section 4.2. Beside simple and cheap usage, this choice allow the easy pipeline adaptation to other languages with no need of language model change, training data creation or model tuning. Along with that, the pipeline architecture permits for an easy replacement of the LLM if required.

Many texts about a single event can be published, requiring a cognitive effort to merge valuable information (Kron et al. 2012). In the same way, when multiple phenomena are described simultaneously, it becomes challenging to isolate their influences. To maintain reliability, source information is preserved as it is, instead of allowing a model to summarize or separate such multi-hazard and multi-source cases. Consequently, the resulting database entity is a news item rather than a hazard event. The identification of an individual hazard event within a collection of texts is done at the last step of the pipeline (Section 3.4).

The NHDB structure was meticulously crafted to balance user interests with the capabilities of the LLM. The structure consists of categories typical for a damage-oriented NH database (Kron et al. 2012; Froude and Petley 2018; Sudnitsyna and Shikhov 2024): the event's nature, timing, location, resulting

damage, and actions undertaken in response. Utilizing a testing dataset (Supplementary Section S4), an iterative prompt complication was realized: starting from 5 general questions, more detailed sub-rubrics were delineated if the recall and precision against the expert-defined responses did not drop. Ultimately, the NHDB structure incorporates 14 fields (Table 1). The overall quality and its variability among the fields at this design step were similar to those evaluated later with an extensive quality assessment (Section 4.2). An independent prompt includes news' source name, title, text body, and publication date.

Table 1 Structure of natural hazards database filled with using LLM

| Database field | Content | Comments and reasons of introduction |
|-------------------------|---|---|
| Text type | Either of the following: event that took place, a forecast, other | The processed sources publish also the weather forecasts. |
| Hazard type | One of 30 predefined types (see Supplementary Table S1) | List of 30 hazards is provided to LLM, and it is asked to select one of them or note "Not recognized". |
| Location: region | Name of the affected region. | We ask the LLM to align the toponyms hierarchy to manage better the coordinates search and mapping at the next step (see Section 3.3). |
| Location: district | Name of the affected district (subregional units). | |
| Location: other details | Any other geographical names, e.g. city/ village/ river/ area/ mountains /road. | |
| Time: year | Year of the hazard occurrence | Although the exact dates could be missing, the year is recognized anywhere with a high precision. Also useful for filtering for analysis. |
| Time: full start date | Date when phenomena started | Several types of hazards are long-lasting thus we are interesting to preserve its start and end. Note that many news text are publishes before an event ends so no end date can be derived. |
| Time: full end date | Date when phenomena ended | |
| Phenomena description | Details on natural phenomena manifestation, e.g. wind speed, precipitation type and amount, landslide size. | |
| Damage: description | Description of any kind of damage and losses to the economy, infrastructure, and personal properties. | The diversity of details is hardly embraceable for the prompt subsectioning, thus no commonly used fragmentation by recipient is applied. |

| | | |
|------------------|--|--|
| Damage: monetary | Reported losses in RUB (Russian currency). | |
| People: dead | Number of people dead. | |
| People: affected | Number of people affected in any way, e.g. injury, evacuation, blackout, transport cessation. | |
| Measures | Actions made preventively, consecutively or for the crisis-management by any actor (authorities, local communities, supplying companies, etc.) | Any measures can be recorded, from physical actions to legal acts. |

3.2.2 Non-generative verification of LLM outputs

Due to the fact that generative models can hallucinate, thus produce the content which did not exist in a source text, we integrated several independent features to verify LLM output. However note that no hallucinations were discovered at any stage of quality controls, as well as user experience (as for now, more than 3000 records read by the experts).

People tend for a limited vocabulary to describe specific types of NH. Thus, the RE approach, mentioned in Section 3.1.2, was adopted for identification of NH type. It achieves F1-scores of 0.85 for single-hazard and 0.77 for multi-hazard texts compared to expert labeling, what is compatible with the similar studies (Yzaguirre et al. 2015; Jongman et al. 2015; Taylor et al. 2015; van den Homberg et al. 2022). The results of RE and LLM recognition of NH type are compared and a flag is stores to speed-up the expert quality control.

The second feature, also driven by RE and resulting in flags, verifies textual descriptive fields of the database. With this approach we cannot control the non-distortion of the initial information nuggets. Instead, the nugget content is checked for the conformity to the expected roles. For instance, “evacuation” could be used to describe affected people or responding measures, but it should not appear in natural phenomena description.

We also tested the detection of geographic toponyms with the NER model Natasha, designed for Russian. Such a class of models was successfully used for the same task with less variable languages (de Bruijn et al. 2019; Yzaguirre et al. 2015; van den Homberg et al. 2022). Despite Natasha demonstrated a detection quality acceptable for the cross-checking goal, its output cannot be compared with LLM successfully: while the latter correctly normalize the names ("Moskvoy" in genitive form > "Moskva" in nominative form for Moscow city), the former extracts the toponyms as is or transforms according to the single masculine form ("Moskv"). Thereby, we do not integrate it into the processing pipeline.

3.3 Geocoding and mapping

To effectively meet user demands for mapping NHDB content, we geocoded the entries. Geocoding involves assigning geographic coordinates to geographic names. We opted for the open-source geocoding service Nominatim (<https://nominatim.org/>) that is driven by open-source geodata from OpenStreetMap crowdsourcing map. Among other tested commercial services (YandexMaps, Dadata, 2GIS), it demonstrated the similar quality and issues with the advantage of being free-of-charge.

Nominatim received a geographical name as text and returned a pair of X-Y coordinates of a point geometry on a map. Although this is a suitable logic for a settlement, it is a substantial misrepresentation for the large geographical units such as boroughs, regions or rivers, which in this case are represented by centroids instead of polygons and lines. In addition to coordinates, a Nominatim response contains extra data through descriptive tags linked to a source OpenStreetMap object. Given this well-structured indication about the type and nature of a geocoded location, we implemented the creation of polygon and polyline mapping geometries using the offline self-prepared geodata (see example on Fig.2b).

To prevent the mismatching of geographical names and their coordinates (de Bruijn et al. 2018), a textual request to Nominatim contains a region name in addition to the place name. Such a request design is possible thanks to the toponyms hierarchical structuring by LLM (see Table 1). When a text mentions several locations, each of them is geocoded independently.

3.4 Identification of events

The scale, damage, or societal impact of an event directly correlates with the volume of publications it generates. For instance, our NHDB contains a collection of over 300 news articles regarding the disaster flooding in the Southern Urals in spring 2024. To ensure accurate processing of the database records and to yield reliable statistics on event magnitude, texts related to the same incident must be linked.

To match the texts, we adopted the routine depicted at Fig.2. It is based on the idea of the spatial and temporal continuity of an event, reflected in the geographical and temporal proximity of publications. Specific criteria for defining “proximity” in space and time are based on the author's expertise (see Supplementary Section S3). For instance, the same day and same location is required for an avalanche, while the drought records may span a week and cover large distances. Note that the adopted spatio-temporal spans directly affect the “event” detection and are subject of discussion. Such a delineation of complex cases is not clear for experts as well (Kron et al. 2012).

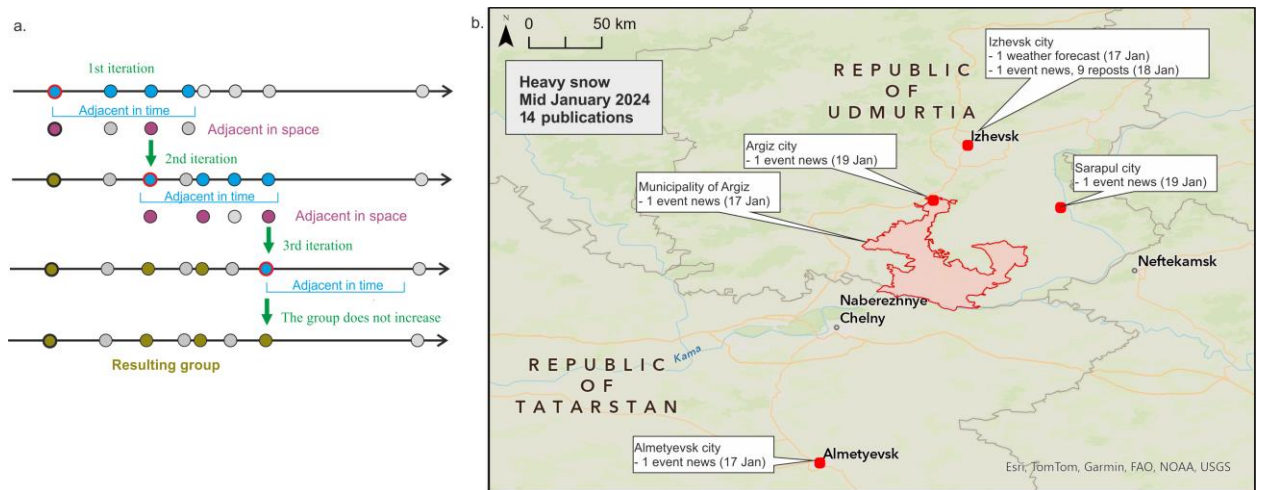


Fig.2 a. Grouping the news items about a same event. b. Map with an example of a news group.

4 Results and discussion

4.1 Selecting relevant and reliable sources of news

Russia regions vary of geography, natural phenomena, community traditions, socio-economic development, and internet penetration, i.e. the factors affecting exposure to hazards and also the online content production. To build a comprehensive NH database encompassing both local and national disasters, it is essential to compile a list of media that cover audiences of various scales and geographies (Taylor et al. 2015).

The NH information mining in newspapers has a long history (Guzzetti et al. 1994; Taylor et al. 2015; Froude & Petley 2018; CRED 2025). Initially, we relied on the federal and regional news agencies only. They were assumed to report socially impactful events, avoid fake news, and minimize other content types on NH topics such as interviews. With the modern shift towards digital, it was hypothesized that the variety and number of sources would adequately represent all regions. Thus, the strategy was to select several national-scale media and 2-3 local outlets per region, manually identifying the candidate sources that are popular and suitable for the scraping (see Section 3.1.1). Out of over a hundred popular resources examined, only 3 national and 7 regional media satisfied the scraping criteria. This revealed that use of mass media alone is inadequate. We kept further only 2 national media, which belong to different holdings and hence produce the original content.

Two other data sources with wide regional representation were revisited as additions: the social network VKontakte (also known as VK; <https://vk.com/>) and the websites of regional EMERCOM offices.

Microblogging platforms like Twitter are often used for ongoing crisis monitoring (de Bruijn et al. 2019; Jongman et al. 2015), but their short format limits usefulness for damage-focused databases (Sultanik & Fink 2012). VKontakte, popular among Russian-speaking users, is similar to Facebook and supports long-form content. While individual profiles and public pages and communities cannot be considered as the reliable sources (van den Homberg et al. 2022), we assume the official governmental pages to be sufficiently reliable. Using VKontakte as a governmental communication platform has been

encouraged during the last decades, and from December 2022, all local and regional administrations were obliged to have their own pages (Government of the Russian Federation 2022). In total, over 2700 official groups were listed for this project, targeting 2-3 sources per district.

EMERCOM is responsible for the management of emergency situations of any nature. Although it has uniformly collected a large amount of risk-management related information since the 1990s, those data are not publically accessible. However, the regional branches of EMERCOM publish daily and weekly forecasts of extreme events, event reports, and other risk-related information. This is a highly regulated, but consistent and approved content. All regional sites were included in our project.

4.2 Quality of the LLM text mining

In this section, we present a comprehensive assessment of the results obtained using the LLM ChatGPT-4o-mini on the 14-field database structure. For that, recall (R) and precision (P) were independently estimated for each testing entity per field, based on comparison with expert annotations, and then globally averaged across all testing entities. The test subset included up two thousand news items of the publication types of “occurred event” and “forecast”.

To create the expert annotations, the testing news were read by the annotators (one news by one person), who filled out the database fields according to predefined instructions. Another expert then compared the model and human answers (see details in Supplementary Section S5), computed R and P (Fig.3; Supplementary Table S4), and categorized the nature of the errors (Fig.4). For the fields evaluated via an “exact match” criterion (e.g., year), a semi-automated comparison approach was applied, enabling the processing of up to two thousand entries; for the multi-component or free-text fields, which require manual evaluation, the subsets of several hundred entries were analyzed (see Supplementary Table S4). If the annotator and expert disagreed, meaning the ambiguity of the guidelines or a text, either of the two answers was accepted from the model.

Figures 3 and 4 present the quality assessment results. The overall performance, error count, and error types are close to those observed during the manual data exploration, LLM selection (Supplementary Section S2), and database structure development (Section 3.2.1). Higher accuracy was observed for questions with concrete, explicitly stated answers, typically marked by specific verbal markers. For instance, the dead people are one of the easiest nuggets for the model thanks to the clear and scanty descriptive. Vague concepts and conclusions derived from the verbal constructions (e.g. date of the event end) more frequently lead to errors. The “Phenomena description” field, where the intensity or character of natural phenomena should be given, is the most ambiguous question for the model, even after the prompt refinements. The hierarchical sorting of the geographical locations between district and city levels was an expected trouble, because the actual Russian administrative division has a set of subregion districts with the literal naming “city’s district”. The multipurpose and multihazard news are the most difficult to process by the model, as well as by the humans. They can

provide at the same time a description of the current weather and a forecast, an announce cash aid to the people and a total monetary loss estimation, etc. A part of LLM errors is related to the clickbait links incorporated in the main text. Finally, one hazard type was consistently misinterpreted by the model, because the prompt contains the list of types' names without explanations: Russian naming of "aufeis" is close to the "ground frost" that non-experts usually mix them too. It should be highlighted that we found no instances of model hallucination, i.e., no generation of information absent from the source texts.

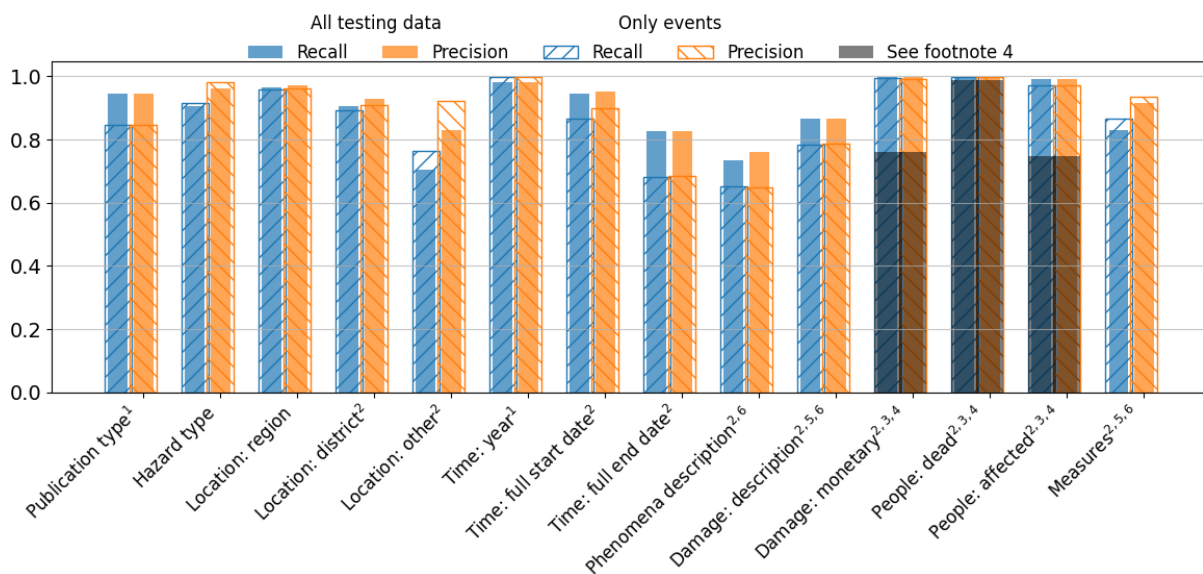


Fig.3 Recall and Precision of the LLM text mining compared to the expert labeling (see details in Supplementary Table S4). Fulfilled bars – the texts about occurred events and weather forecasts; hatched bars – the texts on occurred events alone.

Footnotes: (1) Answer with one element, R=P. (2) Relevant information can be not provided, "Not given" is thus a correct answer. (3) Can be given only in the news about occurred hazards, "Not given" is thus only correct option for the forecasts. (4) Because this information is very rare, R and P are additionally estimated on a special subset: only news containing the relevant information or the provided model answer are taken into account. (5) Taking into account the prompt tasks, the potential damage consequences and safety advices are accepted as correct answers for the forecast texts and evaluated here. (6) Because the prompt asks to keep details, very short and low-informative answers get a penalty of 0.5 scoring.

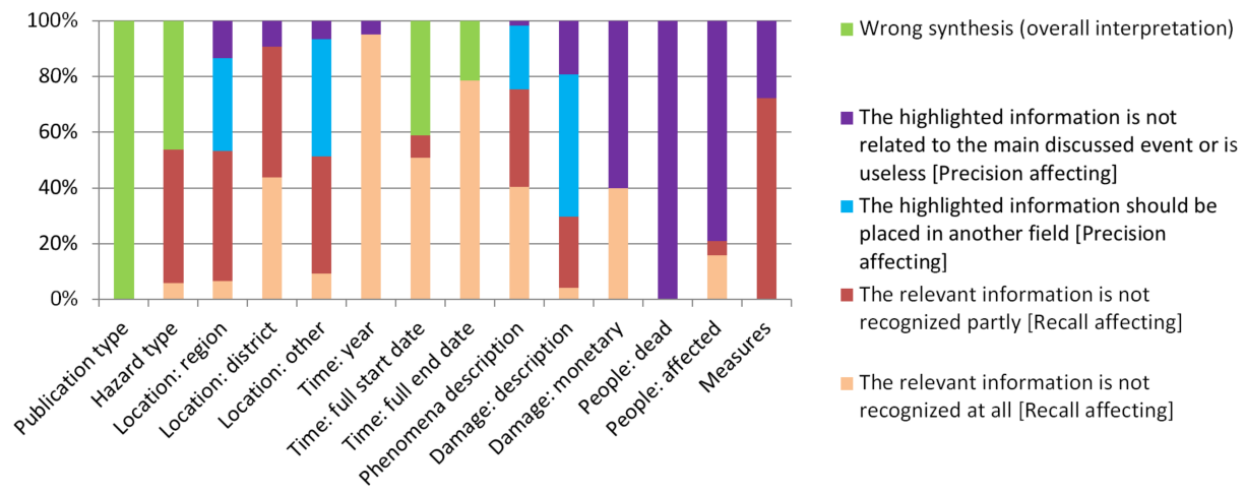


Fig.4 Nature of errors revealed during the LLM recall and precision assessment.

Note, that despite the widespread opinion that the manual text mining yields more accurate results (Kron et al. 2012; Froude & Petley 2018), this is not as much true. In practice, the quality of manual annotation declines as the number of processed texts increases. The readers mistake instructions, overlook required information, mistype, and misinterpret the text just like the model. For instance, the field “Location: other” was incorrect for 6% of the entities and disobeyed the instructions for 4% of the entities. The field “Time: full start date” contained typographical or semantic errors in 12% of cases (e.g. 2012 vs. 2021, June vs. July). Similarly, the “Hazard type” field included incomplete information in 13% of cases and entirely incorrect types in 2%.

Thereby, the trade-off between the quality and scalability make the LLM a highly competitive solution to a human in news mining, especially when a large amount of texts should be processes in limited time.

4.3 Geolocation and mapping

The geolocation of news, also called geoparcing or geotagging, is a task with long history (e.g. Woodruff and Plaunt 1994) and has received a lot of attention in NH-related text mining (Kordopatis-Zilos et al. 2015; van den Homberg et al. 2022; de Bruijn et al. 2018). It is usually done with two technically independent steps (de Bruijn et al. 2018; Yzaguirre et al. 2015): (i) find the toponyms in a text, and (ii) correctly match a location name with the map coordinates.

Recognition of the hazard location among all place names mentioned in a text and homonyms may become a challenging cognitive task (Woodruff & Plaunt 1994; Naaman 2011; Yzaguirre et al. 2015). Three main groups of methods are used for the toponyms search: the direct match of the words from the text with a predefined list of location names (e.g. Sultanik and Fink 2012), usage of NER language models that are trained for location name search (e.g. Imran et al. 2013), and usage of more complex language models which can guess toponyms without target training (e.g. Kordopatis-Zilos et al.

2015). Our pipeline, which uses ChatGPT for this task, is a third case. We found the result highly satisfactory (see Section 4.2 with quality metrics) and confidently helping to deal with the issues mentioned below. Without special training on Russian locations, ChatGPT recognizes toponyms, determines their relevance to a described event, transforms them to a normalized morphological form, and places them into one of three hierarchical database fields.

Once toponyms are extracted, the proper coordinates should be assigned. We selected the Nominatim service for this task (see Section 3.3). A typical list of issues during coordinate matching includes several places having the same name, typos, broad area names, and unofficial names (Sultanik & Fink 2012; Jongman et al. 2015; Battistini et al. 2013).

The three-field hierarchical structure of location storage in our database was designed specially to deal with the namesakes. In contrast with the short and personal Twitter messages (Jongman et al. 2015; de Bruijn et al. 2018), the media texts usually contains the regions and/or district name in the local media name, text title or text body. Although we did not overcome all such cases, the majority of ambiguous situations were managed correctly.

Considering the typos made or preserved by the model, Nominatim is very sensitive and could not resolve them itself. For the region names, which are a short known list of items, we tested the linguistic correction approach based on the similarity measure of a word and the elements of a predefined list. However, the approach did not success, in particular due to native similarity of several regional names, e.g. *Krasnoyarskiy Kray* and *Krasnodarskiy Kray*. Because this issue appears very rarely, we did not investigate its solution further.

Broad areas and semi-official names became the main unmanaged problems for our pipeline. Unofficial names rarely occurs in the news source used, and most cases are recognized and replaced by ChatGPT. However, for several areas the unofficial names are as widely used as official names, thus the model perceives them as acceptable but OpenStreetMaps and thus Nominatim does not contains. Geolocation of broad natural areas (e.g. the Caucasus) or historical areas (e.g. Povolgie for the territories along Volga River) are also troublesome for geocoding. Such areas are rarely mapped in OpenStreetMaps because they do not have clear borders, unlike administrative units. Thus, such toponyms are recognized by ChatGPT but not geocoded by Nominatim. Finally, there are news texts, mainly of weather forecast type, that indicate locations in a descriptive way, e.g. “the southern part of a region” or “the mountain districts of a region”. While an expert can more or less precisely delineate those areas on a map, the absence of toponyms fails the standard automated geocoding procedure. If one needs, a predefined list with the corresponding map geometries may be used to solve those issues. But there are relatively few for us thus we did not investigate it.

The coordinates-search step of our pipeline totally failed for 11% of the processed news items, with 89% being more or less successful. Among latter, some of many of the successfully recognized toponyms were not mapped in 8% of entities. In total, our pipeline achieves the results comparable with

the solutions of the same complexity (Battistini et al. 2013) and more sophisticated (de Bruijn et al. 2018). Note that similar to ChatGPT, Nominatim/OpenStreetMap covers many areas and local languages, thereby the described pipeline of toponyms recognition and geocoding is transferable for other countries. Depending on language specificities, the different issues may become more or less critical.

4.4 Overview of the created database

The created NHDB is based on 2860 individual Internet sources of news: 2 national-wide online newspapers, 85 sites of regional EMERCOM offices, 2768 pages of VKontakte social network. At this testing phase, we focused on 2018-2024 time frame. In total, over 8.3 million news items were collected. The RE-based filtering resulted in 52.4 thousand preserved texts from 1956 individual sources, i.e. 0.6%. A number of items were filtered out after LLM step as not relevant to NH topic, resulting in 48.9 thousand final database entries. The statistics provided below refers to this corps of news items.

4.4.1 News sources and spatio-temporal coverage

More than three quarter of the database entries originate from VKontakte (Fig.5a). They are mainly weather forecasts and thus do not contain data about the really occurred damage or response measures, which are the most valuable content for a damage-oriented NHDB. Mass-media and EMERCOM are more desirable sources in this point of view. Same time, EMERCOM publish a lot formal, low-content updates (grey color at Fig.5c). Many VKontakte announcements replicate EMERCOM weather alerts, thereby usage of those two sources do not increase commensurately the volume of original content. Summing up, mass media are the most informative source in terms of the goal relevance (Taylor et al. 2015) and information diversity (see below). Overall, 27% of the processed texts contained data on factually occurred events.

The temporal distribution of news shows a clear increase (Fig.5b). Nevertheless, it is not a shift in extreme events frequency (IPCC 2021) or rise of infrastructure affection (UNDRR 2015). It is an ongoing government digitalization and the increase of public communications via the social media (Sudnitsyna & Shikhov 2024; Chernokulsky et al. 2020; Government of the Russian Federation 2022).

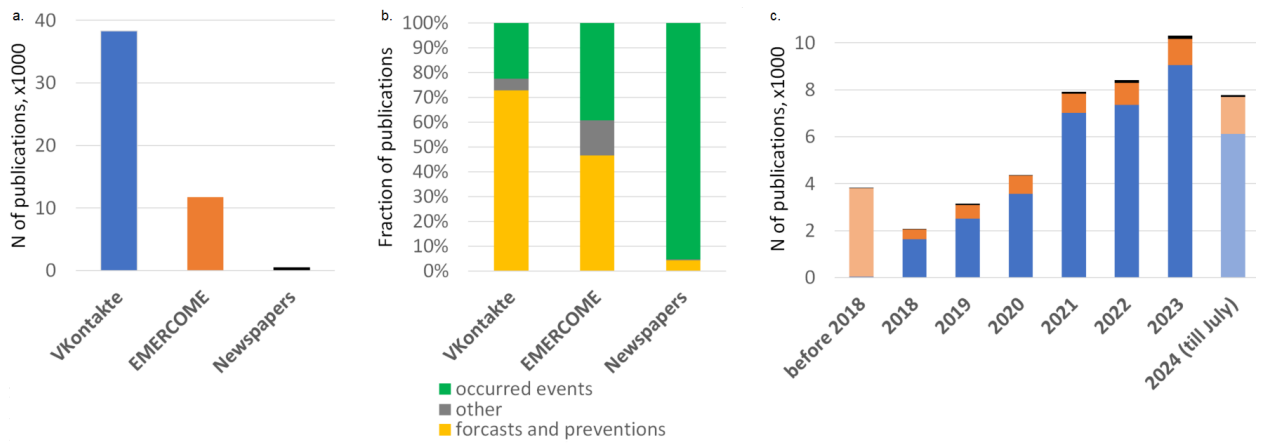


Fig.5. Statistics on the processed news per source, publication type and year (colors at the subfigure A and C are the same).

Geographically, the database covers the regions unevenly. For instance, Caucasus mountains are one of the most replete area of Russia in sense of frequency and diversity of natural hazards (Fuchs et al. 2017) but those regions provide very few content. The uneven Internet penetration, focus on more populated places and damage bearing events, as well as the novelty effect are the most frequently proposed explanations for this spatial bias (van den Homberg et al. 2022; Froude & Petley 2018; Taylor et al. 2015; Kron et al. 2012; Battistini et al. 2013; Jongman et al. 2015; de Bruijn et al. 2019). However, for the Caucasus case, we also suggest the cultural influence: the general reluctance to discuss problems of the state-controlled public discourse (Gavra & Glazkova 2015; Starodubrovskaya et al. 2011).

4.4.2 Frequency of information nuggets

The NHDB entries vary in data richness (Fig.6). The *hazard type* is almost always named and successfully detected by the model. The affected *region* could usually be inferred from the text or source name, though more *detailed location data* were often missing — especially in forecasts of large-covering severe weather. The *year* and *start date* are commonly provided by the texts, but the *end date* can be not evident or has not come yet at the moment of publication. For occurred events, around 67% of news describe the *damage* and 86% describe the *responding measures*. Note that according to a used prompt, detection of the *damage* and *measures* in the forecasting news is also allowed: EMERCOM forecasts may describe the possible hazard consequences and give some response advises. Data on *monetary loss estimate* are the rarest content being hard and long to provide. The *affected individuals* are rarely mentioned relieving that the majority of the local and middle scale weather events does not lead to victims or serious live interruption. Summing up, usually, the news relatively clear specify what, when and where happened or expected. Other information varies depending on incident nature and reporting intent. Despite the fact that the presented statistics is a mix of the actual text content and model mining quality, those patterns agree well with our preliminary manual investigation and similar researches (Imran et al. 2013; Kron et al. 2012; van den Homberg et al. 2022).

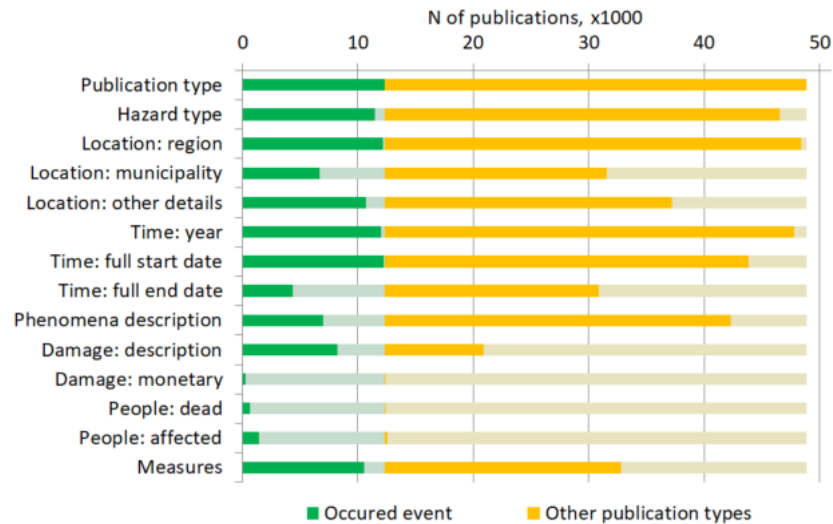


Fig.6 Frequency of information categories recognition by the LLM in the processed news. Bright colors refer the number of NHDB records with the fulfilled information; light colors refer the rest records of the corresponding type.

4.4.3 Types and frequency of natural hazards

The pipeline was designed to recognize 30 types of natural hazards, including cryosphere and geology domains (see Supplementary Table S1). However, only 18 types appeared in the processed texts (Fig.7). We assume the following explanations. First, while certain NH from our list cause damage that is widely perceivable and clearly attributed to a caused phenomenon by the non-specialists (e.g. floods, strong wind), the others do not (e.g. permafrost degradation, forest insects outbreak). Thereby, the latter processes had lower visibility in the news (Raška et al. 2014). This is partly responsible in a vicious circle: such NH also receives less attention in projects of database creation from the documentary (Raška et al. 2014). Another reason is the specificity of a phenomenon manifestation. For instance, a *costal abrasion* usually looks like a series of *landslides* and is described correspondingly by the eyewitnesses. Finally, the names of several hazards from our list are professional terms and unlikely used by a wide public as is.

Looking on the text content, 46% mention a single hazard, and the rest mentions two and more inter-related or cascade processes.

76% of the texts (36.9 thousand) met the technical requirements for matching a group of texts talking on same event (see Section 3.4). We identified 20.8 thousand of such unique events. The vast majority are weather forecasts. However, the texts on the occurred events show the important live impact of floods and wildfires: in contrast to the weather phenomena, they have higher fraction of factual texts against the forecasts. Among the formed text groups, 13.4 thousand events are described by a single message, 2.9 thousand events – by two messages, and the remaining 4.3 hazards – by three or more messages. The most publicized event was a disaster 2024 spring flood in South Ural (EM-DAT

database ID: 2024-0198-RUS): 165 texts for Kurgan region, 160 for Orenburg region, and 93 for Tyumen region, enlightened the water level forecasts, ongoing flooding extent, evacuation proposals, help measures, and so on during almost two months.

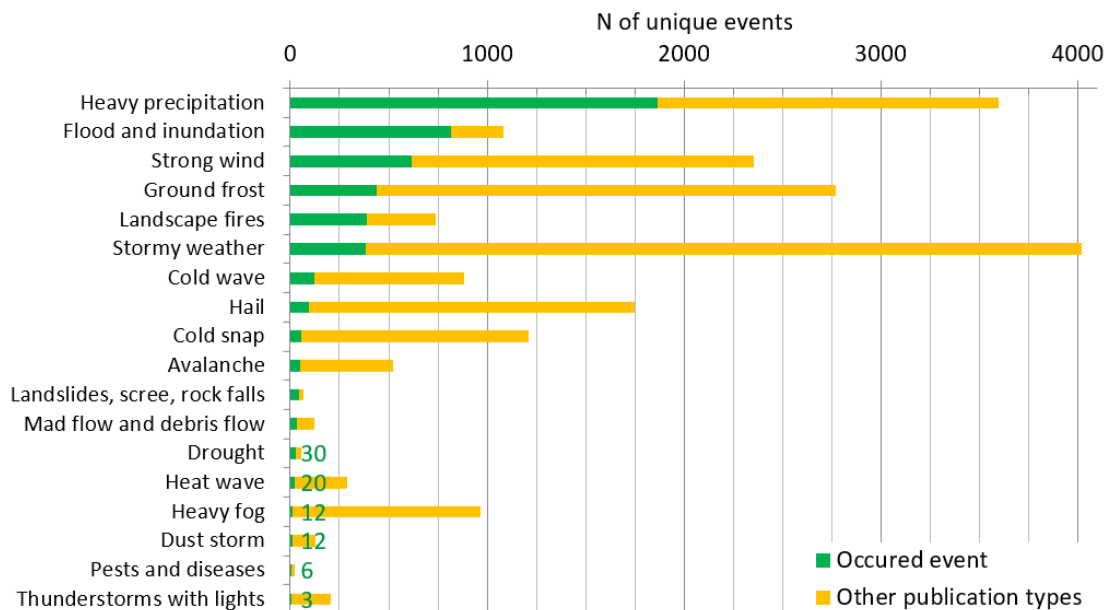


Fig.7 Number of unique events per hazard type. Green labels are the number of unique occurred events when the quantity is too low for the clear plot representation.

4.4.4 Comparison with the external databases

The news-based databases contain the biased sampling of events without clear reasons for inclusion. In fact, they highlight events that worry people and potentially disturb their lifestyle regardless the phenomena intensity or severity. This is in contrast with the traditional databases, which are collected according to some strict criteria. Nevertheless, we run the validation of the derived events against the traditional sources. Supplementary section S6 and Table S5 provide the sources used, methodology and detailed results.

First of all, the Russian government-maintained sources were examined. We found the correspondences for more than 1100 unique events (about 10%) of our database. The floods and the forest fires have the highest rate of the matched records, significantly ahead the weather phenomena.

The world-wide, internationally recognized EM-DAT database (<https://www.emdat.be/>) was also examined. We found the correspondences for all 17 events, traced by EM-DAT in 2018-2024. Less critical incidents were mentioned in 3-10 news, while large-scale weather events and spring riverine floods were noted in the dozens of forecasts, ongoing news and post-factum reports. EM-DAT contains much more frequently the monetary losses, but our sets of texts provide more descriptive insights and details. Another difference is about identification of the event duration: for the complex, large-scale hazards such as floods after stormy weather, we constantly trace the longer duration in the news than EM-DAT does.

5 Conclusion

For several decades, the international community of risk reduction takes efforts on establishing the national hazards databases everywhere. However, publicly accessible and reach-in-content data still stays a challenge for the majority of countries. One of the ways to fill this information gap is news texts collection and processing. Despite this is biased source and the build-on results should be used with the limitation understanding, such texts provide many insides on spatial affection, intensity and consequences of the occurred hazards.

A traditional method for news exploitation involves semi-automated searching and manual extraction of information. Recently, automated approaches utilizing language models have been developed. Here we present a fully automated pipeline for creating a news-based hazard database using a Large Language Model. The process encompasses text filtering, information extraction, geocoding, mapping, and quality checking, and can handle various text sources, including newspapers, social media, and official reports. The proprietary model ChatGPT-4o-mini from OpenAI LTD was chosen as the core processing engine due to its high performance, cheap usage price, and ease of use with many languages. We extract information on occurred phenomena, dates, locations, damage to the economy and individuals, and response measures. The accuracy of the resulting data ranges from 0.70 to 0.99 across different categories, which is on average comparable to the manual mining of large text amount. The geolocation process utilizes the open-source service Nominatim, driven by global OpenStreetMap geodata.

The produced trial database demonstrates a number of issues common for the similar news-based products. However, verification against the traditional databases demonstrates its value. This database more frequently provides the information about the local-scale incidents and describes richer the large-scale events. It also contains the responding measures that are rarely collected.

Combination of the multi-language ready-to-use model, the world-wide open-source geodata, and open-source programming language supports the easily transfer of this processing pipeline to other regions and languages. Thus, the presented solution can be widely used to enhance the factual data abundance required for the disaster management improvement.

Statements

Data and code availability

The resulting Hazard Database can be shared upon a request for the academic and education goals under a BY-CC-like license; a demo subset is freely available at <https://geography.hse.ru/georisks/results>. The list of the processes sources and the news items of EMERCOM can be shared for free for any goals upon a request; the news of VKontakte and newspapers

cannot be shared according to media holding policies. The codes of the processing pipeline, described in Methodology section, can be shared for usage and developments upon a request.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This research is part of Strategic Project “Human-Centered AI”, which is part of Higher School of Economics’ development program under the “Priority 2030” academic leadership initiative. The “Priority 2030” initiative is run by the Ministry of Science and Higher Education of the Russian Federation as part of National Project “Science and Universities”.

Authors’ contribution

A.D. contributed to conceptualization, funding acquisition, methodology elaboration, code writing, data processing, results analysis and verification, and writing and editing of the original paper draft.

M.S. contributed to data curation, methodology, software development, supervision, and validation.

G.K. contributed to conceptualization, formal analysis, funding acquisition, investigation, methodology, project administration, supervision, validation, visualization, and review and editing of the manuscript.

T.A. contributed to conceptualization, funding acquisition, project administration, methodology, resource provision, and supervision.

Acknowledgements

Authors would like to thank all collaborators who contributed to this study, including the team of Andrey Dorozhniy for the VKontakte scraping; a software engineer Maxim Ritikov for his help with Python code; the team of Petr Parshkov for the realization of LLM tuning experiment; Vsevolod Moreydo, Ekaterina Podolskaya and Tatiana Gorbacheva for expert verification with external databases; the students of the FGGT Faculty, HSE University, for execution of quality check routines. Ekaterina Sarapulova’s role for the continuous management support is greatly appreciated.

References

- Atefeh, F. & Khreich, W., 2015. A Survey of Techniques for Event Detection in Twitter. *Computational Intelligence*, 31(1), pp.132–164. Available at:
<https://onlinelibrary.wiley.com/doi/10.1111/coin.12017>.
- Battistini, A. et al., 2013. Web data mining for automatic inventory of geohazards at national scale. *Applied Geography*, 43, pp.147–158. Available at: <http://dx.doi.org/10.1016/j.apgeog.2013.06.012>.
- Brakenridge, G.R., Anderson, E.K. & Carlos, H., 2009. Dartmouth Flood Observatory: Global Active Archive of Large Flood Events. Available at:

- <http://floodobservatory.colorado.edu/Archives/index.html>.
- de Bruijn, J.A. et al., 2019. A global database of historic and real-time flood events based on social media. *Scientific Data*, 6(1), pp.1–12. Available at: <http://dx.doi.org/10.1038/s41597-019-0326-9>.
- de Bruijn, J.A. et al., 2018. TAGGS: Grouping Tweets to Improve Global Geoparsing for Disaster Response. *Journal of Geovisualization and Spatial Analysis*, 2(1), p.2. Available at: <http://link.springer.com/10.1007/s41651-017-0010-6>.
- Carrara, A., Crosta, G. & Frattini, P., 2003. Geomorphological and historical data in assessing landslide hazard. *Earth Surface Processes and Landforms*, 28(10), pp.1125–1142. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/esp.545>.
- Chernokulsky, A. et al., 2020. Tornadoes in Northern Eurasia: From the Middle Age to the Information Era. *Monthly Weather Review*, 148(8), pp.3081–3110. Available at: <https://journals.ametsoc.org/view/journals/mwre/148/8/mwrD190251.xml>.
- CRED, 2025. EM-DAT: The International Disaster Database. Available at: <https://www.emdat.be/>.
- Earle, P. et al., 2010. OMG earthquake! can twitter improve earthquake response? *Seismological Research Letters*, 81(2), pp.246–251.
- Enikeeva, K.R. et al., 2016. On the role of social media services to support decision-making in emergency situations [in Russian]. *Problems of risk analysis*, 13(1), pp.36–45. Available at: <https://cyberleninka.ru/article/n/o-rol-i-servisov-sotsialnyh-setey-dlya-podderzhki-prinyatiya-resheniy-v-chrezvychaynyh-situatsiyah>.
- Froude, M.J. & Petley, D.N., 2018. Global fatal landslide occurrence from 2004 to 2016. *Natural Hazards and Earth System Sciences*, 18(8), pp.2161–2181.
- Fuchs, S. et al., 2017. Editorial to the special issue on natural hazards and risk research in Russia. *Natural Hazards*, 88(s1), pp.1–16.
- Gavra, D. & Glazkova, S., 2015. Communicative strategies and misunderstandings. Discourse analysis of the Russian North Caucasus case. *Asian Social Science*, 11(19), pp.237–246.
- Government of the Russian Federation, 2022. On Designating VKontakte and Odnoklassniki as Information Systems Utilized by Government Bodies for Official Pages [in Russian].
- Guzzetti, F., Cardinali, M. & Reichenbach, P., 1994. The AVI project: A bibliographical and archive inventory of landslides and floods in Italy. *Environmental Management*, 18(4), pp.623–633.
- van den Homberg, M., Margutti, J. & Basar, E., 2022. Enriching impact data by mining digital media. *United Nations Office for Disaster Risk Reduction (UNDRR)*, (April).
- Imran, M. et al., 2013. Extracting information nuggets from disaster- Related messages in social media. *ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management*, (May), pp.791–801.
- Imran, M. et al., 2015. Processing Social Media Messages in Mass Emergency: Survey Summary. *ACM Comput. Surv.*, 47(4), p.38.

- IPCC, 2021. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. In V. Masson-Delmotte et al., eds. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.
- Jongman, B. et al., 2015. Early flood detection for rapid humanitarian response: Harnessing near real-time satellite and twitter signals. *ISPRS International Journal of Geo-Information*, 4(4), pp.2246–2266.
- Kordopatis-Zilos, G., Papadopoulos, S. & Kompatsiaris, Y., 2015. Geotagging social media content with a refined language modelling approach. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9074, pp.21–40.
- Kron, W. et al., 2012. How to deal properly with a natural catastrophe database - Analysis of flood losses. *Natural Hazards and Earth System Science*, 12(3), pp.535–550.
- Mazhin, S. et al., 2021. Worldwide disaster loss and damage databases: A systematic review. *Journal of Education and Health Promotion*, 10(1), p.329. Available at: https://journals.lww.com/10.4103/jehp.jehp_1525_20.
- Melnikova, A., 2019. *Specific features of the russian emergency journalism (1986 – 2018)*. Peoples' Friendship University of Russia.
- Ministry of Economic Development of the Russian Federation, 2021. On Approval of Methodological Recommendations and Indicators on Adaptation to Climate Change [in Russian].
- Ministry of the Russian Federation for Civil Defence, E. and E. of C. of N.D., 2021. On Establishing Criteria for Information on Natural and Technogenic Emergencies [in Russian].
- Naaman, M., 2011. Geographic information from georeferenced social media data. *SIGSPATIAL Special*, 3(2), pp.54–61. Available at: <https://doi.org/10.1145/2047296.2047308>.
- Petrova, E.G., 2009. Natural and Technogenic Emergencies in Russia: Experience in Compiling and Analyzing a Database. In A. L. Shnyarkov, ed. *Snow Avalanches, Mudflows, and Risk Assessment*. Moscow: Universitetskaya Kniga, pp. 152–162.
- Raška, P. et al., 2014. Documentary proxies and interdisciplinary research on historic geomorphologic hazards: A discussion of the current state from a central European perspective. *Natural Hazards*, 70(1), pp.705–732.
- Shamin, S.I., Bukhonova, L.K. & Sanina, A.T., 2019. Database of Hazardous and Adverse Meteorological Events.
- Shikhov, A. & Bykov, A., 2014. The database on hazardous and severe weather events in the Perm Region as a regional analogue ESWD [in Russian]. *Geographical bulletin*, 4(31), pp.102–109. Available at: http://gis.psu.ru/wp-content/uploads/2015/01/530_Shikhov_Bykov.pdf.
- Shnyarkov, A. & Gryaznova, V., Natural hazard database for Russia (1950x - 2017).
- Starodubrovskaya, I.V. et al., 2011. *The North Caucasus: a modernization challenge [in Russian]*,

- Moscow: "Delo" RANHiGS. Available at: <https://www.iep.ru/ru/publikacii/publication/4447.html>.
- Sudnitsyna, T. V. & Shikhov, A.N., 2024. Developing a gis database of hazardous hydrological events (with the Kama River basin as an example) [in Russian]. *Geographical bulletin*, 12(3), pp.178–189. Available at: <http://117.74.115.107/index.php/jemasi/article/view/537>.
- Sultanik, E.A. & Fink, C., 2012. Rapid geotagging and disambiguation of social media text via an indexed gazetteer. *ISCRAM 2012 Conference Proceedings - 9th International Conference on Information Systems for Crisis Response and Management*, (April), pp.1–10.
- Taylor, F.E. et al., 2015. Enriching Great Britain's National Landslide Database by searching newspaper archives. *Geomorphology*, 249, pp.52–68. Available at: <http://dx.doi.org/10.1016/j.geomorph.2015.05.019>.
- The World Bank, 2016. Solving the puzzle: Innovating to Reduce Risk. *Global Facility of Disaster Risk and Recovery*.
- U.S. Geological Survey, 2025. ANSS Comprehensive Earthquake Catalog (ComCat). Available at: <https://earthquake.usgs.gov/earthquakes/search/>.
- UNDRR, 2025. *Disaster Losses and Damages Data: A Review of Existing Applications and Use Cases*, Available at: <https://www.undrr.org/node/89954>.
- UNDRR, 2020. Hazard definition & classification review: Technical Report. *Hazard Definition and Classification Review. Technical report.*, p.88. Available at: <https://www.undrr.org/publication/hazard-definition-and-classification-review>.
- UNDRR, 2015. *Sendai Framework for Disaster Risk Reduction 2015 - 2030*,
- UNDRR, 2021. *Strategic Framework 2022-2025*, Available at: <https://www.undrr.org/media/49267/download?startDownload=true>.
- Wirtz, A. et al., 2014. The need for data: Natural disasters and the challenges of database management. *Natural Hazards*, 70(1), pp.135–157.
- Woodruff, A.G. & Plaunt, C., 1994. *GIPSY: Georeferenced Information, USA*: University of California at Berkeley.
- Yzaguirre, A., Warren, R. & Smit, M., 2015. Detecting environmental disasters in digital news archives. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 2027–2035. Available at: <http://ieeexplore.ieee.org/document/7363984/>.
- Zuzak, C. et al., 2022. The national risk index: establishing a nationwide baseline for natural hazard risk in the US. *Natural Hazards*, 114(2), pp.2331–2355. Available at: <https://doi.org/10.1007/s11069-022-05474-w>.

Supplementary materials

S1. The list of the target natural hazards

This is a list of natural hazards, targeted for the identification by the presented processing pipeline.

Table S1 List of natural hazards for the database creation.

| Hazard group | Hazard type | Comment |
|--|-------------------------------|--|
| Shallow geohazards (ground-related and slope-related) | Karst | |
| | Suffosion | |
| | Subsidence of loess ground | |
| | Landslides, scree, rock falls | Mass movement of relatively dry soil |
| | Mad flow and debris flow | Mass movement (flow) of liquefied soil or hyperconcentrated water flow |
| | Avalanche | |
| | Erosion | Planar, rill and gully erosion, thermoerosion in permafrost |
| Hydrology | Riverbed deformations | Vertical and horizontal channel deformations and coastal erosion in watercourses |
| | Inland coastal modification | Erosion and accretion of reservoir and lake shores |
| | Abrasion | Coastal erosion of seashores |
| | Flood and inundation | Riverine, coastal and pluvial floods |
| Cryosphere | Permafrost degradation | Including thermokarst |
| | Frost heaving | |
| | Solifluction | Mass movement of surface layer related to freeze-thaw cycle |
| | Aufeis | From river and group water |
| Weather and climate | Strong wind | Any kind – tornado, hurricane, cyclone, etc. |
| | Dust storm | |
| | Drought | |
| | Heat wave | Very hot weather (typically, above +35 °C) |
| | Cold snap | Short and sudden drop of temperature close to 0 °C, damageable mainly for agriculture |
| | Sudden temperature change | Inter-diurnal temperature change by 10 or more degrees |
| | Cold wave | Very cold weather (typically, below -35 °C) |
| | Hail | |
| | Heavy precipitation | Strong or long-lasting rain or snow |
| | Thunderstorms with lights | |
| | Heavy fog | |
| | Ground frost | Ice deposited directly on objects when temperature drops slightly below 0 °C |
| | Stormy weather | Set of two and more bad weather phenomena |
| Biosphere | Landscape fires | In any type of landscape (forest, steppe, tundra, etc.) independently of the cause of the fire |
| | Pests and diseases | Damaging forest and agricultural |

S2. LLM selection

To select the LLM, we tested following models from the Generative Pretrained Transformers (GPT) group:

two chatbots from Russia:

(1) YandexGPT (Yandex LLC, Russia; <https://ya.ru/ai/gpt>) and

(2) GigaChat (PJSC Sber, Russia; <https://giga.chat/>);

and two versions of ChatGPT (OpenAI, USA <https://chat.openai.com>):

(3) ChatGPT-4o and

(4) ChatGPT-4o- mini, which is a smaller and cheaper version of the former.

Gemini (Google, USA; <https://gemini.google.com/>) demonstrated low quality at the first test of the hazard type recognition, and was discarded from further model comparison. All tested models are free in manual mode, but paid when API automatization is used.

We used a simplified structure of a database for this experiment. It had five broad sections:

1. Hazard type, 2. Location, 3. Dates, 4. Damages and losses, 5. Measures. A prompt contained five corresponding questions, the news text with the metadata (source name and publication date), and a request to provide an answer in JSON format. Note that the prompt optimization was out of the experiment scope of LLM selection. The testing dataset contained 60 news items (see Supplementary Section S3), the same as were used to design the NHBD structure (Section 0). Similarly for the main database, a database entry was a news items. The recall (R, items detected by model among all correct items) and precision (P, correct items among all detected items) were evaluated per news item against the expert answers, and then averaged by the global mean.

Figure S8 shows the recall for two ChatGPT and the better of the two Russian models. Because P was consistently higher than R for all models, we do not show it here. ChatGPT demonstrated the best quality of text mining in all five questions, completed all the tasks, and did not fail to respond according to the prescribed JSON structure. GigaChat had on average 20% lower recall, and it refused to answer in 10% of requests (e.g. "I am tired and don't want to do this"). YandexGPT failed to respond to 80% of requests due to the unstable API connection and refusals, thereby we disqualified it from the quality comparison.

Despite the fact that ChatGPT-4o has slightly better results than ChatGPT-4o-mini, we choose the latter for our pipeline due to the 20 times lower cost. About 60\$ was paid for all tests and the processing of more than 52'000 filtered news items.

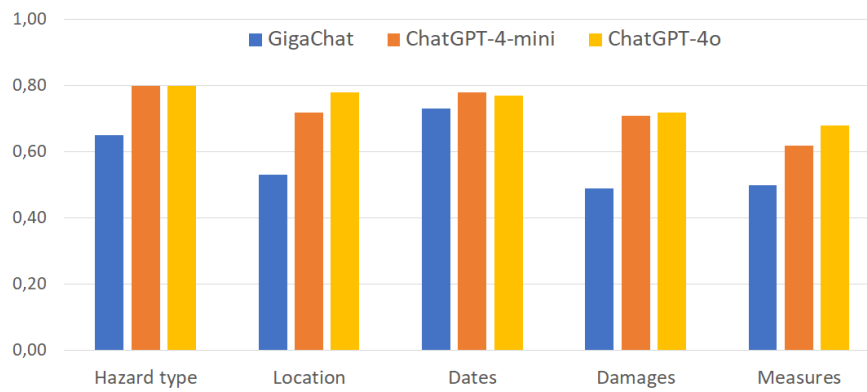


Figure S8 Recall of three candidate LLM in tests of mining hazard type, location, date of event, caused damage, and undertaken actions from the news items.

Several authors use the self-trained NER and NLP models for an analogous task with an appropriate satisfaction. After the integration of a ready-to-use LLM, we also explored the capability of LLM tuning. We expected that it could improve the quality of response for the complicated content of location, damage, and measures. This approach was tested with YandexGPT, the only tested LLM which can be tuned by end-users. The experiment ran after we had compiled NHDB with 14 fields. Two thousand training records (see Supplementary Section S5) from the NHDB created with ChatGPT-4o-mini were semi-randomly selected to proportionally represent the diversity of NH and source types, and had been verified by experts. Several experiments with different designs ran, including the best performing training of separate models for each thematic group of the NHDB fields. However, ChatGPT remained the better solution. Summing up our experience, we found the ready-to-use LLM to be more accurate, quicker, and monetarily and computationally cheaper.

S3. Spatio-temporal spans for the news grouping

The Table S2 provides the spatial and temporal spans used to find the news items about a same event as described in Section 3.4 of the main manuscript. All hazards are categorized with the attribution of the corresponding time and distance spans based on personal expertise. The time spans are based on a typical, expected duration of a natural phenomena. The following spans are used for the “forward” search (see Figure 2 of the main manuscript): immediate process – 0 day span (same day), short duration – 3 day span, long duration – 15 days span. The spatial spans depict the typical, expected spatial coverage of the damaging influence. The operation intersection between the buffered map geometries is used to match the news. Because a news can be geocoded on a scales from a settlement to a region (see Section #), two different buffer sizes are adopted for the points and polygons, correspondingly: local influence – 7 and 1 km, extended influence – 50 and 10 km, overregional influence – 70 and 20 km.

Table S2 Descriptive categories of the temporal and spatial spans for the news grouping.

| Hazard type | Temporal duration | Spatial influence |
|-------------------------------|--------------------------|--------------------------|
| Karst | Immediate | Local |
| Suffosion | Immediate | Local |
| Subsidence of loess ground | Immediate | Local |
| Landslides, scree, rock falls | Immediate | Local |
| Mad flow and debris flow | Immediate | Extended |
| Avalanche | Immediate | Local |
| Erosion | Immediate | Local |
| Riverbed deformations | Immediate | Extended |
| Inland coastal modification | Immediate | Extended |
| Abrasion | Immediate | Extended |
| Flood and inundation | Long | Overregional |
| Permafrost degradation | Immediate | Extended |
| Frost heaving | Immediate | Extended |
| Solifluction | Immediate | Extended |
| Aufeis | Immediate | Local |
| Strong wind | Short | Overregional |
| Dust storm | Short | Extended |
| Drought | Long | Overregional |
| Heat wave | Long | Overregional |
| Cold snap | Short | Overregional |
| Sudden temperature change | Short | Overregional |
| Cold wave | Long | Overregional |
| Hail | Short | Overregional |
| Heavy precipitation | Short | Overregional |
| Thunderstorms with lights | Short | Extended |
| Heavy fog | Short | Extended |
| Ground frost | Short | Extended |
| Stormy weather | Short | Overregional |
| Landscape fires | Long | Extended |
| Pests and diseases | Long | Overregional |

S4. Testing set of news for the processing pipeline elaboration

The experiment of LLM selection (Section 4.2 of the main manuscript) was realized on 60 news items. They were partly selected from the set of texts, used for the manual creation of a database prototype (Section 2 of the main manuscript) and enriched by the targetly searched news to balance the variability of several parameters. The final set covered 26 hazard types from 30 assumed (see Table S1). The characteristics of the set are provided in the Table S2. Note that at the moment of the models testing, EMERCOM was not yet envisaged as an information source, thus, no corresponding texts have been used.

For the elaboration of the database structure (Section 3.2.1 of the main manuscript), five news from EMERCOM have been added to the presented news set. According to the criteria from Table S1, the added texts were simple, short, with one hazard, and did not contained the toponyms irrelevant to a described event.

Table S3 Description of the testing news set, total size – 60 news items.

| Criteria | Options | Number of news |
|--|---|----------------|
| News Source | Newspapers | 30 |
| | Vkontakte | 30 |
| | EMERCOM | 0 |
| Complexity | Simple | 31 |
| | Complex (several hazards, historical notes, autor’s discussion, irrelevant information, etc.) | 29 |
| Size | Short | 52 |
| | Long | 8 |
| Number of hazards, mentioned in a text | One | 41 |
| | Many, all of them are relevant to the described event | 6 |
| | Many, some of them are not relevant to the described event | 13 |
| Several toponyms are not relevant to the described event | No | 41 |
| | Yes | 19 |

S5. Quality of automatic texts mining

This section provides the additional information for the Section 4.2 of the main manuscript.

To estimate the quality of the LLM text mining, the model answers was semi-manually compared with the human-provided annotations.

A total of 2065 news items, manually created during the LLM tuning experiment (see end of Section 4.2), were reused as the evaluation dataset. The news items were semi-randomly selected from the main corpus to ensure diversity in sources, hazard types, and text formats. The set included 21, 901, and 1173 items from Newspapers, VKontakte and EMERCOM sources respectively; this is 1%, 42%, and 57% of the set. The occurred events are noted in 687 items (33%), while the remainder consists of weather forecasts and other types of content. The set covers all 85 regions represented in the main database.

Several annotators received an instruction specifying the expected correct response for each database field. Each text was read once by one person, who then filled in the relevant database fields. After that, an independent evaluator compared the model-generated and human-provided answers. In cases of disagreement, the evaluator reviewed the original text to determine the correct outcome. Based on this verification, each instance was classified as either an LLM error, a human annotation error, or a case of ambiguity, where both answers were acceptable. The latter was introduced for two main

reasons. First, some texts are genuinely ambiguous. For instance, a phrase “...during the past day” in an afternoon article could plausibly refer to either “today” or “yesterday.” Another example is a text like “The ongoing storm is forecasted to continue for two more days”, which could be equally interpreted as either an “occurred event” or a “weather forecast” according to the model’s prompt. The second reason is that the manual text mining was initially conducted for a different purpose. Therefore, the annotation guidelines were slightly different from the LLM’s original prompt. For instance, the LLM’s “Text type” question included several options (e.g. “occurred event”, “weather forecast”, “other”), while the annotators have only two categories (“occurred event” and “no occurred event”). If the required information was not provided in the source news, a correct model answer should be “Not given”.

The table below provides the detailed values used for Figure 4 of the main manuscript including recall, precision, and the size of each test dataset. The largest evaluation set—where all items were read—corresponds to fields for which semi-automatic matching was possible, requiring minimal manual verification. Smaller evaluation subsets were used for fields involving free-text comparison or high mismatch rates, which required extensive manual review. The exceptions are fields on monetary damage assessment, died people and affected people, where the information is objectively rare.

Table S4 Recall, Precision, and the sizes of testing dataset per database field for the LLM text mining.

| Database field | Footnotes | Mean recall (R) | | | Mean Precision (P) | | | Number of news items in the dataset | | |
|-----------------------|-----------|----------------------|-------------|-----------------------|----------------------|-------------|-----------------------|-------------------------------------|-------------|-----------------------|
| | | Full testing dataset | Only events | Only special subset 4 | Full testing dataset | Only events | Only special subset 4 | Full testing dataset | Only events | Only special subset 4 |
| Publication type | 1 | 0,944 | 0,847 | | 0,944 | 0,847 | | 2065 | 668 | |
| Hazard type | | 0,904 | 0,914 | | 0,96 | 0,98 | | 1711 | 519 | |
| Location: region | | 0,966 | 0,959 | | 0,970 | 0,960 | | 2065 | 668 | |
| Location: district | 2 | 0,905 | 0,894 | | 0,929 | 0,908 | | 296 | 102 | |
| Location: other | 2 | 0,706 | 0,765 | | 0,830 | 0,923 | | 296 | 102 | |
| Time: year | 1 | 0,980 | 0,998 | | 0,980 | 0,998 | | 2065 | 668 | |
| Time: full start date | 2 | 0,945 | 0,865 | | 0,953 | 0,898 | | 2065 | 668 | |
| Time: full end date | 2 | 0,826 | 0,681 | | 0,828 | 0,685 | | 455 | 125 | |
| Phenomena description | 2, 6 | 0,734 | 0,652 | | 0,761 | 0,649 | | 408 | 223 | |
| Damage: description | 2, 5, 6 | 0,867 | 0,782 | | 0,867 | 0,786 | | 260 | 142 | |
| Damage: monetary | 2, 3, 4 | 0,998 | 0,994 | 0,760 | 0,997 | 0,991 | 0,760 | 2065 | 668 | 30 |
| People: dead | 2, 3, 4 | 0,999 | 0,999 | 0,988 | 0,999 | 0,999 | 0,988 | 2065 | 668 | 86 |
| People: affected | 2, 3, 4 | 0,99 | 0,971 | 0,747 | 0,99 | 0,971 | 0,747 | 2065 | 668 | 80 |
| Measures | 2, 5, 6 | 0,830 | 0,867 | | 0,914 | 0,934 | | 260 | 142 | |

Footnotes: (1) Answer with one element, R=P. (2) Relevant information can be not provided, “Not given” is thus a correct answer. (3) Can be given only in the news about occurred hazards, “Not given” is thus only correct option for the forecasts. (4) Because this information is very rare, R and P are additionally estimated on a special subset: only news containing the relevant information or the provided

model answer are taken into account. (5) Taking into account the prompt tasks, the potential damage consequences and safety advices are accepted as correct answers for the forecast texts and evaluated here. (6) Because the prompt asks to keep details, very short and low-informative answers get a penalty of 0.5 scoring.

S6. The existing databases used for the validation of the created database

This section provides detail for the validation of the created database against the existing relevant databases.

Except EM-DAT, the selected databases are maintained by the Russian governmental institutions according to the criteria of an event registration. They are fully or partly accessible for the wide public. The provided data contains only the “what-when-where” information, without damage estimates. Thereby, only a fact of a phenomenon occurrence is checked.

Validation was organised through the databases’ entities matching, using the same idea of spatio-temporal adjustment as for texts grouping, explained in Section 4.2 of the main manuscript. The databases have different aims and the related methodology, content, structure, and data formatting. Thereby, the matching was done automatically, semi-automatically or manually. Accordingly, we assume a different completeness of entity matching depending on the method: the unmatched entities are surely unique when compared manually, but with the automatic matching this may be due to both technical reasons and true uniqueness. In several cases only a subset of the data was validated due to the incompatibility of the databases structures. All the limitations and matching results are summarised in the Table S5. Note that because a news item is an entity of the created database, the many-to-one relationship is possible with the event-based validating databases.

Table S5. The technical limitations and the results of the created database validation against the existing similar databases.

| Maintaining authority; database name; reference | Comparison limitations and issues; matching approach | Compared hazards | Number of entities (if applicable): successfully matched in the created database / processed in the created database / processed in the external database |
|---|---|--|---|
| Roshydromet (hydrometeorological service); “Hazard weather conditions that caused economic losses”; http://meteo.ru/data/adverse-weather-conditions/ | 1. Only the phenomena exceeding a predefined intensity threshold. 2. Location usually is given at a region scale. 3. List of NH slightly differ from our list, the alignment was done. 4. Principle of the main NH | Strong wind Drought Heat wave Cold snap Cold wave Hail Heavy precipitation | 232 / 1162 / 543 4 / 58 / 33 5 / 22 / 19 2 / 72 / 40 8 / 169 / 26 13 / 122 / 111 229 / 4406 / 401 |

| | | | |
|--|---|--|--|
| | identification differ from the LMM-provided approach. Matching: automatic | Thunderstorms Heavy fog Ground frost Stormy weather | 0 / 3 / 1 0 / 13 / 5 18 / 582 / 5 7 / 599 / 458 |
| Roshydromet (hydrometeorological service); “Automated information system for monitoring of water bodies at hydrological station”; https://gmvo.skniivh.ru/ | 1. Fixes the everyday water level at a station point, contains a note if the level exceeding a predefined threshold; no traces about a flooding of the surrounding itself. 2. Semi-open access to the data till 2022. 3. Point-based information, the region-describing news cannot be matched. Matching: semi-automatic | Riverin flooding | 366 / 420 / - |
| RosNedra (geological service); “Yearly gazettes of exogenous geological processes” https://specgeo.ru/monitoring-sostoyaniya-nedr/gosudarstvennyy-monitoring-nedr-gmsn/production-gmsn/ | 1. Incidents of any nature (natural and antropogenic), with and without consequences. 2. At the moment of verification, data accrued for 2022-2024 3. List of NH slightly differ from our list, the alinement was done. Matching: automatic | Landslides, scree, and rock falls | 6 / 61 / 1921 |
| RosLeshos (forest management service), data processed by a side project; “Gazette of fires in forests”; https://tochno.st/datasets/fires | 1. Location as a point coordinate. 2. Start date – a date of the inventory registration Matching: manually | Landscape fire | 154 / 522 / - |
| EMERCOM (rescue ministry), data processed by a side project; “Gazette of landscape fires”; http://data-in.ru/data-catalog/datasets/202/ | 1. Location as a point coordinate. 2. Data till 2021. 3. Start date – a date of the inventory registration 4. Uneven in time quality and completeness of data. Matching: manually | Landscape fire | 135 / 522 / - |
| CRED; EM-DAT; https://www.emdat.be/ | 1. List of NH slightly differ from our list, the alinement was done. 2. Principle of the main NH identification differ from the LMM-provided approach. Matching: manually | Strong wind Stormy weather Riverin flooding Landscape fire Debris flow | 75 / - / 3 20 / - / 1 197 / - / 9 18 / - / 3 2 / - / 1 |