https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

From text to geoinformation – A modular approach for extraction of disaster information from web text data

Vanessa Rittlinger [1], Johannes Mast [1], Stefan Voigt [1], Christian Geiß [1,2], Hannes Taubenböck [1,3]

[1]German Remote Sensing Data Center, German Aerospace Center (DLR), Oberpfaffenhofen, 82234, Germany
[2]Department of Geography, University of Bonn, Bonn, 53115, Germany
[3]Department of Remote Sensing, University of Würzburg, Würzburg, 97074, Germany

Correspondence to: Vanessa Rittlinger (vanessa.rittlinger@dlr.de)

Abstract. The implementation of effective disaster management measures requires comprehensive information about a given flooding situation. Text data from web news offer potentially large volumes of information for this purpose. However, the extraction and spatiotemporal analysis of flood event-related information is inherently demanding due to the immense volume of unstructured text. Addressing this challenge, we present a modular and scalable method that allows the extraction of disaster-relevant information from a large text corpus. This is accomplished by combining domain specific entity extraction with dictionaries, a machine learning model for toponym identification, and hand-crafted rules for entity linking in a modular workflow. The extracted information is augmented with geolocations in order to support spatial analysis. Using the West Germany flooding event 2021 as a case study, we evaluate the capacity of our approach to extract relevant geospatial information at a variety of spatial granularity levels and in the form of various thematic descriptors. By doing so, we outline the capabilities and limitations of this approach for text extraction and analysis. Furthermore, we demonstrate the potential for systematic utilization of text data for improved situational awareness and for disaster management support.

1 Introduction

Understanding the spatial and temporal dynamics of flooding events and their consequences through comprehensive analyses of past events offers valuable insights for enhancing risk management and mitigation strategies. Diverse, heterogeneous data sources like remote sensing-based information products, cadastral records, *in-situ* measurements, and census data are routinely integrated to assess hazards with catastrophic potential (De Albuquerque et al., 2015). Large textual corpora, including news articles, web pages, and social media posts, have emerged as a particularly rich resource for extracting disaster-related information, supporting rapid mapping and informed decision-making and future disaster mitigation and management. Textual data can supply complementary, and at times superior, insights as compared to traditional sources such as satellite-derived flood extent maps, and it is not subject to temporary data-delivery restrictions (Wieland et al., 2025). Notable types of text data that are used in disaster information extraction procedures are news data or social media (De Bruijn et al., 2020; Eyre et al., 2020; Klonner et al., 2016; Senaratne et al., 2023a). Social media remains a widely used textual source for disaster information extraction (Hanny et al., 2025; Suwaileh et al., 2022). However, platform-specific constraints, exemplified by Twitter's

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

changing data-usage policies over time, can limit the accessibility and completeness of these data streams (Navalkar et al., 2025). Beyond this, news websites and their text content are used as sources for disaster-relevant information extraction (Owuor et al., 2020). Frequently, the spatial and temporal dynamics of a given disaster event are particularly relevant to identify spatial patterns and event hotspots (De Albuquerque et al., 2015; Wieland et al., 2025). Especially the combination of localized text-based information with other geospatial information sources, such as satellite data, can enhance situational understanding and awareness for given disasters events (Zhu et al., 2022).

To extract and analyze information from text data, methods such as topic modelling, topic classification and Named Entity Recognition (NER) are frequently used in natural language processing (NLP) and information retrieval (De Bruijn et al., 2020; Fu et al., 2022; Kahle et al., 2022).

Currently applied NER methods range from rule-based methods to unsupervised and supervised machine learning (ML) models. Rule-based methods require predefined rules designed by an expert without the need of explicit integration of empirical prior knowledge. However, they are limited to rather static predefined settings. ML approaches, on the other hand, can model complex relationships in the data and generalize well to new and unexpected text content, however, they require large amounts of properly encoded prior knowledge (Lai et al., 2022).

A variety of these NLP approaches have successfully been applied in the natural hazard domain. A widely used application is document-level topic classification (Abraham et al., 2025). However, there is often much more relevant information than people, places, and organizations given in the text, and often also more than a singular topic per document. More fine-grained approaches have been used to identify entities within documents. Domain specific entity identification, such as for biological, legal, and geological domains, is already state-of-the-art (Fu et al., 2022; Leitner et al., 2019; Sänger et al., 2024). In the disaster domain, specific ontologies have been created to provide support for disaster management purposes (Shukla et al., 2023). Different natural hazard types can be attributed to texts via ML NER models (Sun et al., 2022): news coverage of a flooding event has already been researched in a descriptive way (Kahle et al., 2022). A newspaper article text analysis was also carried out to assess the spatial distribution of the economic impact of a flooding event (Madruga de Brito et al., 2025). These approaches generally overlook the nuanced content that resides in individual documents and words, and they have not yet focused on extracting fine-grained disaster-relevant information (Kahle et al., 2022; Madruga de Brito et al., 2025). Consequently, even basic keyword identification remains a useful strategy for pulling out pertinent details from disaster-related texts (Senaratne et al., 2023a). Within text-based disaster analysis and risk assessment, the location of a specific event is a critical source of information, as it allows linking the information about the event with its geospatial context (Lai et al., 2022). By linking the locations mentioned in a text to the corresponding event characteristics such as the type of disaster, its impact, or mitigation measures researchers can significantly enhance information relevance (Schiersch et al., 2020). Nevertheless, systematic identification of disaster-relevant information from text across diverse textual sources and their geospatial localization remains largely unexplored.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Considering the challenges mentioned above, we seek to address the following research questions:

- What added value can web text data from media and public websites offer for the extraction of disaster-relevant information and subsequent potential information for mitigation measures?
- At which spatial and temporal level of detail can unstructured web text data be used to extract structured, geospatially explicit disaster-relevant information?

Despite the general capabilities of applying text analysis methods in the disaster domain, it is still a challenge to extract temporally, spatially, and thematically fine-grained, geolocated disaster information directly from the related web texts. To address these challenges in the context of flooding events described above, we present a pragmatic approach for integrated extraction of georeferenced information of a flooding situation from web text data. Aiming to go beyond basic identification of the natural hazard type, we extract individual words that provide information possibly relevant in the context of a disaster, in the following described as 'disaster-relevant information' (DRI). To identify DRI in a scalable way, we combine natural language processing (NLP) methods to generate relevant and geolocated information directly from unstructured text. We implement these in a comprehensive pipeline that covers all necessary processing from the extraction of semantically relevant text to the spatial representation of the information. We exemplify the approach in a re-analysis of a flooding event in Germany as well as its spatial and temporal development through its web text-based footprint. Furthermore, we provide a synergistic comparison of our results with other geospatial data sets for the given flood situation. Ultimately, we aim to contribute to a better understanding of disaster and crisis situations in future situations, where, e.g., satellite-based mapping and web-text based situational awareness can be combined.

The remainder of the paper is organized as follows. Section 2 presents the integrated data, the study area, the approach is tested on, as well as the methodological approach and experimental setup. In section 3 experimental results are presented and discussed. Section 4 elaborates the potential and limitations of the approach and concludes with an outlook on possible future research work. Section 5 provides concluding remarks.

2 Study Area & Data

2.1 Study area

As a case study, we choose the West Germany 2021 flooding event. The large impact of the event led to high media coverage and the generation of several geospatial flood mapping products, making it a suitable event for our proof-of-concept. The event occurred between the 14th and 15th of July due to a large summer storm system called "Bernd" (Fekete and Sandholz, 2021). The heavy rainfall triggered flooding that resulted in more casualties and injuries than any other event in Germany over the last 50 years, while inflicting major damage to the economy, infrastructure, and housing. Geospatially, we focus on the counties that announced the state of emergency in West Germany during the event (Figure 1) (Schäfer et al., 2021). One particularly affected region in this area was the Ahr valley, which was extensively covered by media (Kahle et al., 2022). Due to the area's morphology, the heavy rain and the resulting flood caused soil erosion, material transport, and accumulation which led to the

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

capacity limits of retention systems and dams being exceeded, e.g., the Steinbachtalsperre (Schäfer et al., 2021). The strong impact of the event can be at least partly linked to higher flow velocities, the ad-hoc blockage of runoff by organic debris carried by floodwaters and man-made debris, the narrow riverbed, the natural course of the river narrowed by obstacles close to the riverbed by housing or other buildings and infrastructure such as low bridges and surface sealing in settlements that reduce the absorption capacity of the soil (Schäfer et al., 2021). Post-event analyses revealed a large disparity between the actual flood extent and the existing flood hazard map, underscoring a substantial underestimation of the event's magnitude (Schäfer et al., 2021). The flood and its extensive consequences triggered a nationwide debate about the likelihood of such a large-scale event, the unfolding of the disaster, and the associated warnings and evacuation protocols. In the aftermath, new prevention measures were introduced to mitigate the impact of future events (Fekete and Sandholz, 2021). The resulting long-term information stream, encompassing damage-limitation actions, policy responses, and public discourse, provides an ideal dataset for assessing the capability of our method to extract and georeference disaster-relevant content from text.

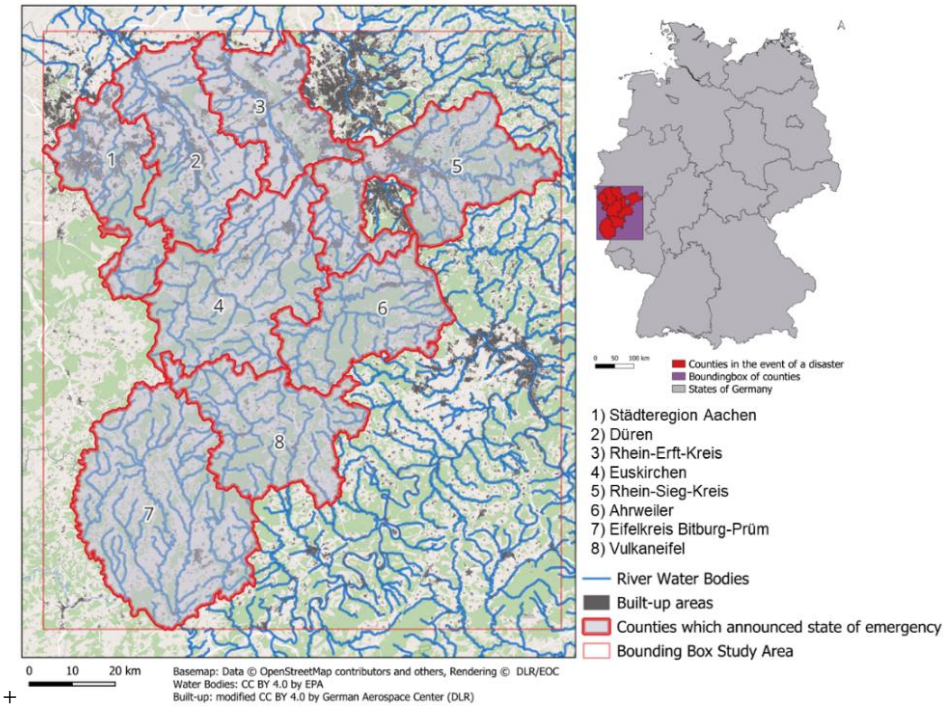


Figure 1: Map of Germany (top right) and in red the states which announced the state of emergency and their bounding box which is used as the focus area in this study

2.2 Web text dataset

Web text data, in the form of news, social media, and official public websites and reports, can provide information about a given disaster event. In this study, we use data from the Global Knowledge Graph (GKG) database of the GDELT project (GDELT Project, 2015; Leetaru and Schrodt, 2013), which systematically collects news articles from the web. The collection is enriched with the URL of the website, dates and other information about the content of the website, such as the topic and

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

location of the website. GDELT as a data source is commonly used for text analysis (Saz-Carranza, et al., 2020; Senaratne et al., 2023b) and has been used, for example, for event tracking of natural and man-made events (Owuor et al., 2020).

For the purpose of this study, the GDELT GKG served as a means to preselect relevant websites with text related to the West Germany 2021 flooding event. The GDELT metadata was used to identify articles that matched flood-related themes, the region of the study area, and the time frame from 12th of July 2021 to 31st of August 2024, to cover a time range from shortly before the onset of the disaster event up to a long-term period for reconstruction. In order to acquire the original text of the websites written in German for the spatiotemporal and thematic analysis of major floods, the websites were downloaded using the URLs of the articles. This enriched dataset was then passed to the pipeline to extract geolocated DRI from text.

2.3 Geospatial flooding event information and mapping used for comparison and analysis

During the extreme flooding event in West Germany in June 2021, the Copernicus Emergency Management Service (CEMS) produced a high-resolution spatial inventory of the event (EMSR517 - Copernicus EMS Mapping | Copernicus EMS On Demand Mapping, 2026). The dataset contains airborne and Sentinel-1 satellite imagery from which flood extent masks and building-level damage grades were derived. In the 17 event areas where the service was activated, flood extent and building-damage products were delineated on multiple dates (15.07, 16.07, 19.07, 20.07, 22.07, 8.11.) (Figure 2) (EMSR517 - Copernicus EMS Mapping | Copernicus EMS On Demand Mapping, 2026). These layers serve as the reference against which the spatial and contextual outputs of our web-text mining approach are subsequently evaluated. The event-specific activation maps indicate the anticipated relevance and impact in each area, thereby providing an objective baseline for validating the text-based damage assessments.

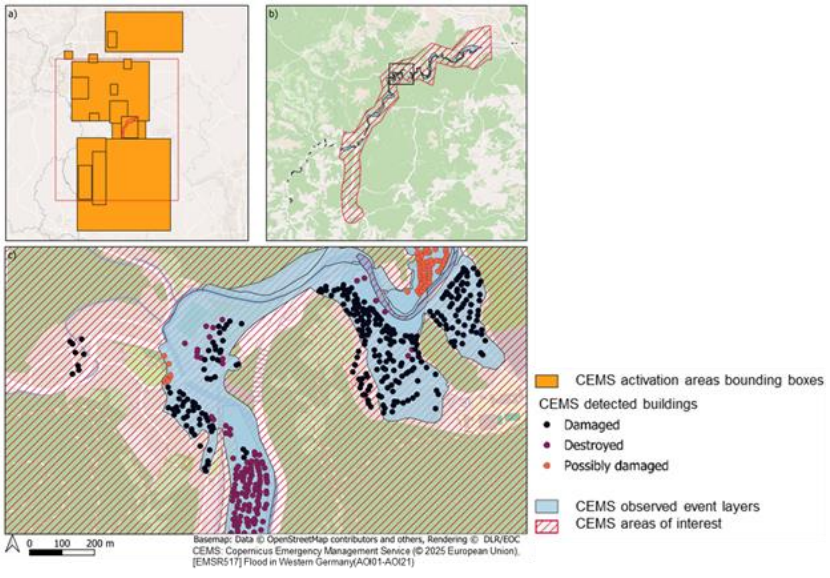


Figure 2: CEMS data products overview with a) bounding boxes of 17 activation areas, b) layer of observed event area of Bad Neuenahr-Ahrweiler and c) closer snapshot of CEMS classified building information, area of interest and observed event layer

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

3 Methodology

With the overarching objective of evaluating how text data can support disaster management during events and inform post-event mitigation, we design a modular workflow for extracting DRI from textual sources and geolocating it. The workflow proceeds in three stages:

- Firstly, the workflow identifies entities of DRI and toponyms within the text,
- geocodes the identified toponyms,
- and finally relates geolocation entities with DRI entities derived from the text.

The results comprise a set of point vector data containing the enriched DRI information at a specific location. Figure 3 provides an overview of the complete processing chain with the individual modular processing steps. We focus on automation of the pipeline to process large amounts of data. The individual modules draw on methodological approaches of various complexity, which we chose based on their empirical performance properties and the need for scalability, transferability, and transparency. For our special focus on extracting natural catastrophe information using NLP, we start with a rule-based approach that guarantees both validation and interpretability. Although recent studies report higher accuracies for ML models in related tasks, their transferability to disaster-specific settings is limited by domain-specific constraints and sparse training data (Leitner et al., 2020). Consequently, we adopt a dictionary-based method for extracting DRI, a strategy that has been shown to be efficient previously (Reveilhac and Morselli, 2022). For the processes of toponym detection and geocoding, we apply a proven state-of-the-art ML model and a geocoding service for our tasks (Akbik et al., 2019; Hu et al., 2024). Finally, we align toponyms to DRI using a rule-based linking scheme.

3.1 Identification of DRI within text data

News articles on a disaster event can contain valuable and diverse information of potential interest, such as people affected, information on damaged infrastructure, mitigation measures, and many more. Therefore, we apply a categorization system for DRI, reflecting the processes and sequence of the disaster cycle, starting from the initial phase, through the response phase, up to mitigation and adaptation measures (Bosher et al., 2021). We identify flood-related keywords and structure them into a dictionary comprising nine classes. These classes reflect the disaster cycle and the progression of a natural catastrophe as well as relevant information for each phase (Table 1). The identification of relevant keywords in the crisis management context is common practice (Reveilhac and Morselli, 2022; Senaratne et al., 2023a). Examples of our classification system are shown in Table 1. Our dictionary creation process identifies relevant keywords by applying a Term Frequency-Inverse Document Frequency (TF-IDF) approach (Senaratne et al., 2023a) to a curated corpus of flood-related reports and documents and complementing it by manually adding keywords based on our knowledge of the disaster domain. As our database consists of documents in German, there is a need to accommodate suffix variations of words based on the gender, case, and singular or plural (Kharis et al., 2021). Specifically, adding a certain word to the dictionary implies adding all of its possible variations to the dictionary. To keep the original text, we use an approach where we add variations of the keywords from their lemma rather

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

than lemmatizing all words and all text to compare the data with (Table 1). Our preset keyword list in a single language limits the number of potential DRI words that can be linked within the text, but this approach serves as a baseline in this proof-of-concept study.

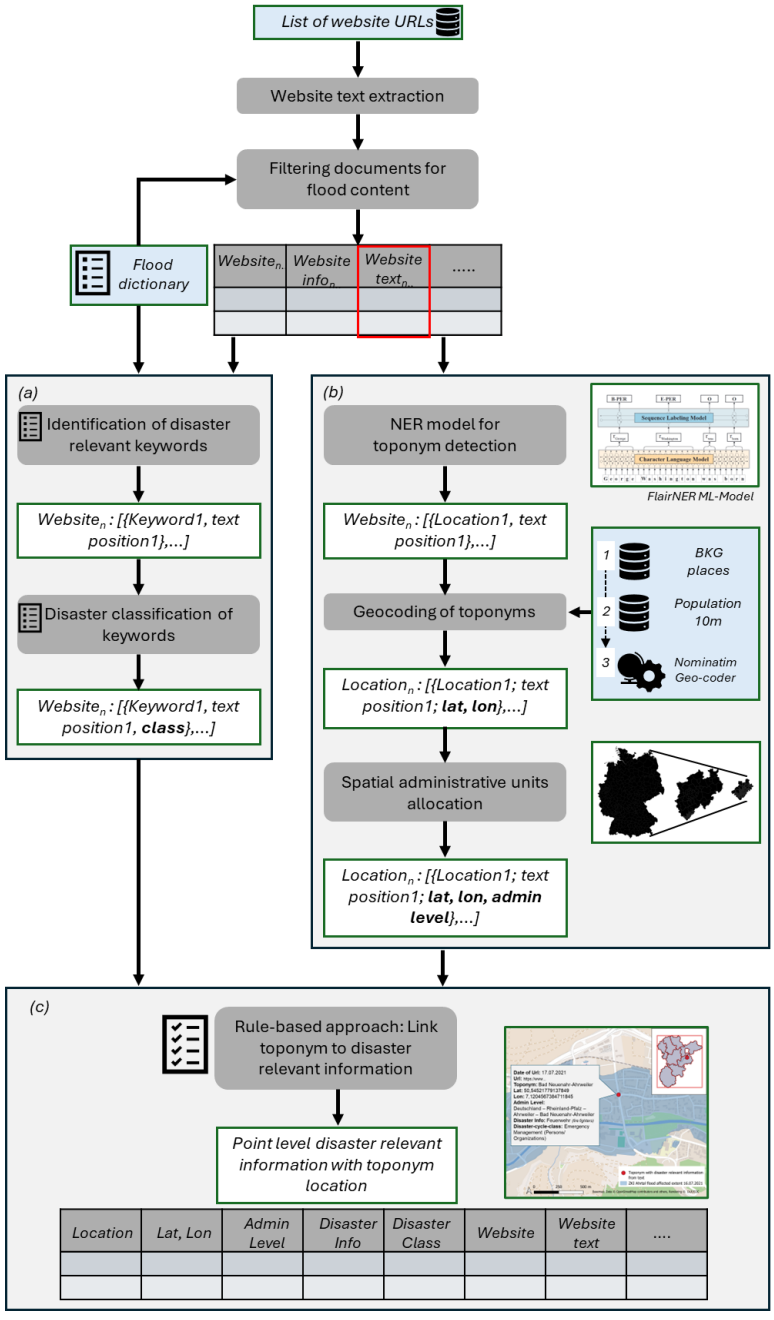


Figure 3: Schematic presentation of the workflow with (a) data preprocessing to derive the plain text from websites, disaster-relevant keyword extraction and classification, (b) toponym identification and geocoding and (c) linking of DRI to toponyms

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Disaster cycle phase (class)	Number of words per category	Keyword examples	Suffix variation examples
Flood-Synonyms	36	Hochwasser	Hochwassers
Disaster	75	Überflutung	Überflutungen
Prediction	80	Hochwassergefahr	Hochwassergefahren
Impact / Human loss	78	Hochwasseropfer	Hochwasseropfers, Hochwasseropfern
Impact Damage Infrastructure	251	Brückenschaden	Brückenschäden
Reconstruction	50	Aufbauhilfe	Aufbauhilfen
Emergency Management (Measurement)	151	Bergungsmaßnahme	Bergungsmaßnahmen
Emergency Management (Person/ Organization)	121	Polizei	Polizistin, Polizist, Polizisten.
Mitigation	61	Hochwasserschutzanpassung	Hochwasserschutzanpassungen

Table 1: Dictionary disaster classes (according to disaster cycle) and example keywords in the context of flood

In order to identify web news texts in which a flooding event plays a central role, we integrate a two-stage filtering process for the text content based on our dictionary. First, we filter for documents containing keywords from the category ‘disaster' and ‘flood-synonyms’. Second, we include keywords from the remaining phases of the disaster cycle. Only if both filtering steps are positive in the document, we consider this document as a relevant source for our analysis.

The objective of this approach is to extract fine-grained DRI, here seen as entities, from individual documents (see Figure 3(a)), rather than simply assigning a single topic at the document level. To this end, we employ a curated dictionary that enumerates key terms associated with the various phases of the disaster cycle. Each dictionary entry is matched against the full text retrieved from the target websites and the exact positions of each match within the text are recorded. This dictionary-based entity-recognition step thus provides the foundational, position-aware extraction upon which subsequent linkage and analysis can be built.

3.2 Identification and geocoding of toponyms

As stated previously, a large potential for enriching DRI lies in linking it with its geographic context via its geolocation. Therefore, we integrate a geoparsing procedure to identify and geocode toponyms (Hu et al., 2024) in the text around the DRI, as precisely as possible, including information on the hierarchical levels of administrative units (Figure 3(b)). We adopt the workflow introduced by Serere et al. (2023), where we apply a machine learning NER model for the toponym identification and a geocoding method to infer the location. We use the FlairNER model, which has already been shown to successfully integrate toponym identification in German texts (Akbik et al., 2019). In our use case, we apply it to identify the location, which is described as toponym entity.

The next important step is adding the latitude and longitude information to the identified places and thus the related DRI within the text. For the geocoding, the process of converting locations as text into latitude and longitude coordinates, we use a

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

combination of gazetteer matching and the use of geocoding models (Hu et al., 2024). A sequence of queries is made from different publicly available geospatial databases (here used as gazetteers) in order to maximize spatial coverage and obtain accurate location information. At first, the toponyms are being matched with the point database from local authorities GN250. Geographical names of municipalities, parts of municipalities, landscapes, mountains, hills, islands, rivers, canals, lakes and seas, etc. of Germany are stored in this layer (© GeoBasis-DE / BKG, 2024). Non-matched toponyms with the first database are passed to the second step, where two publicly available databases with global coverage of places in German are used for matching (GADM Project, 2025; naturalearthdam, 2025). The remaining non-matching toponyms are passed to a fine-grained geocoder, named Nominatim (OpenStreetMap contributors, 2025), to return coordinates with a parameter-limiting focus on our study area.

The geospatial precision of information within a text can vary. Some toponyms might refer to exact point locations while others refer to different levels of administrative units like counties, states, or countries. While all can be represented as point locations (usually via the centroid of the region or of its bounding box), it is difficult to compare different levels of administrative units as well to as represent them in the geocoding process at the same time. Therefore, we assign the respective administrative hierarchy information to each toponym in our focus study area using polygon vector files of the administrative boundaries of Germany. The hierarchy of German administrative levels thus represented ranges from national level to municipality level (Admin level 1-5: Country, state, districts, administrative associations and municipalities) (© GeoBasis-DE / BKG 2025).

3.3 Disaster-relevant information (DRI) extraction and entity toponym linking

Once the toponym entities and DRI entities are identified within a given text, corresponding toponyms and DRI are linked (Figure 3(c)). We apply a rule-based decision process to define a relational pair of toponym and DRI. If such a DRI occurs either within the very same sentence, in the sentence before or after a toponym, it is linked to this particular toponym (Figure 4(a)). This rule is based on the position distribution of toponym and DRI entities of the reference labels. Additionally, we implement two further rules to investigate post-processing steps to improve the precision of the entity linking approach (Figure 4(b)). Besides the initial entity linking rule (t_1), two further rules are applied:

- (t_2) The second, pruning rule removes relation pairs in the sentence positioned before and after, when there is already a descriptor linked with a toponym.
- (t_3) According to the third rule, only those descriptor-toponym links are accepted that occur within the same sentence.

It has to be mentioned that the process is computed for each toponym independently. Therefore, it is possible that more than one toponym can occur in a sentence. In this case, our approach accepts inter-sentence linking of all potential combinations of toponyms and DRI.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

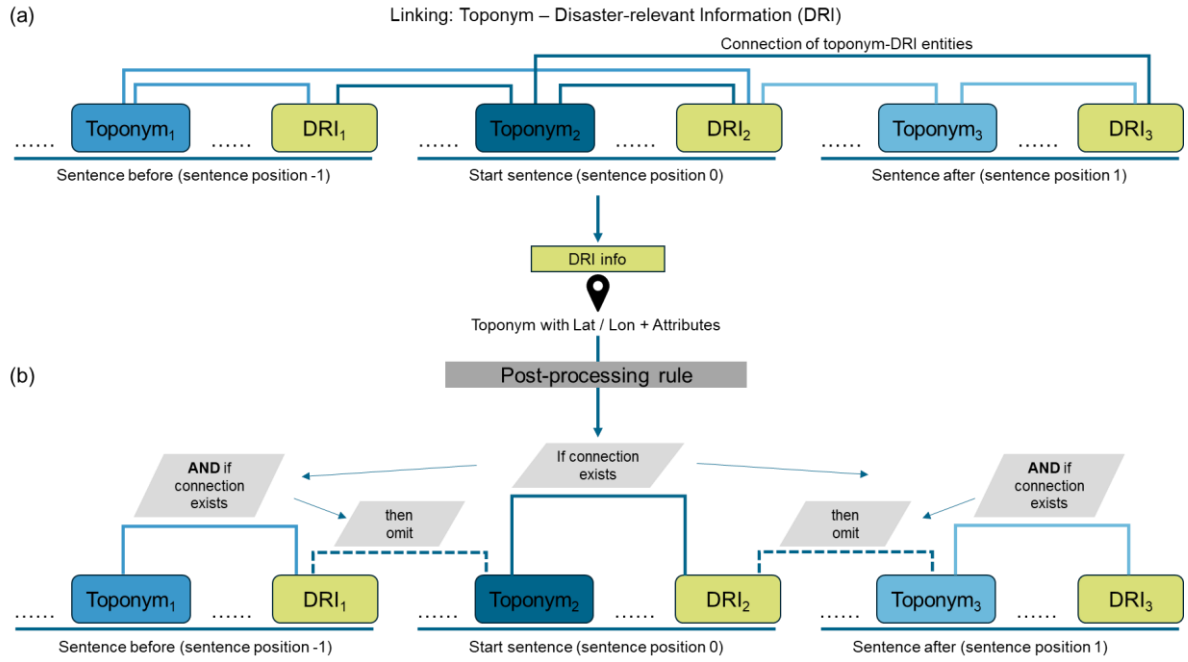


Figure 4: (a) Presentation of toponym and DRI linking process in the workflow and (b) integration of post processing rule of the output

To evaluate the validity of our workflow and its outputs, we constructed a manually annotated reference dataset by randomly sampling documents from the corpus. In order to ensure the coverage of all classes in the reference corpus, we used a stratified random sampling method (Thompson, 2012). Altogether, 122 documents were thus randomly selected and then manually labelled, which resulted in 5,087 toponym-DRI pairs. The resulting reference corpus includes relation pairs with a correct identification of the toponym, DRI, and the related disaster-cycle class and correct connection between toponym and DRI. This evaluation, focussing on the entity pairs, is similar to the approach presented in Schiersch et al. (2020).

To validate the workflow output quantitatively, the statistical measures of precision, recall and F1 score are computed (Warto et al., 2024). For recall, false negatives are considered missing labelled toponyms and their missing links to DRI. The emphasis of the evaluation is put on the correct identification of toponyms and its related DRI, which is also categorized into the nine predefined disaster-cycle classes. Errors or biases from the early steps (Toponym identification, toponym geocoding and DRI identification) are in this case not individually evaluated.

4 Experimental results

In this chapter, we first present the statistical validation of the outputs from the three system components. The best-performing rule is used for the data generation. Subsequently, we evaluate the potential of the DRI-extraction process using the finalized workflow output for the 2021 Western Germany flooding event. Finally, we present the potential and granularity of the information that can be obtained through this process.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

4.1 Statistical validation of the process pipeline

To validate our approach, we used the labelled reference dataset to assess its plausibility with statistical measures. We analyze the output of the three post-processing conditions after testing the entity link rules described above (t_1, t_2, t_3). For each document, we examine whether the toponym-DRI relation labelled in the reference document is also present in the results from our workflow. To do this, we calculate precision, recall and the F1 score of all three output variations (Figure 5).

With respect to absolute numbers of relation pairs the two postprocessing steps lead to a notable reduction in the absolute numbers of relation pairs. The total number of relation pairs for the original workflow is 12,699, with the integrated rule on adjacent sentences there are 9,418, and within the same sentence there are only 4,949 relations. In comparison, the reference corpus contains 5,087 relations distributed across the 122 documents. Precision, recall and F1 score are calculated per document for the number of matching relations compared to the reference corpus, in order to provide statistical information on the performance of the approaches. On a document level for all of the 122 documents, the precision is with 24.4% lower in the basic output that only uses one linking rule (t_1) compared to the other two outputs. This is attributable to the high number of relations. The highest precision values of 43% are found in the output with relations only in the same sentence (t_3). This elevated precision can be partly attributed to the distribution of relation distances in the reference corpus, as the vast majority of toponym–DRI associations are restricted in the same sentence. The recall value is higher in the original (t_1) and adjacency rule (t_2) output. Consequently, fewer relevant DRI items are omitted when the adjacency rule is applied, compared with the stricter intra-sentence filter. In terms of F1 score, the output with the same sentence rule(t_3) shows an average F1 score of 42.6%, followed by the adjacency rule output (t_2) at 38.7%, and lastly the original output (t_1) at 33.4%. We decided to use the adjacency rule applied output (t_2) for subsequent geospatial analysis, since it shows robust and balanced performance in all evaluated metrics.

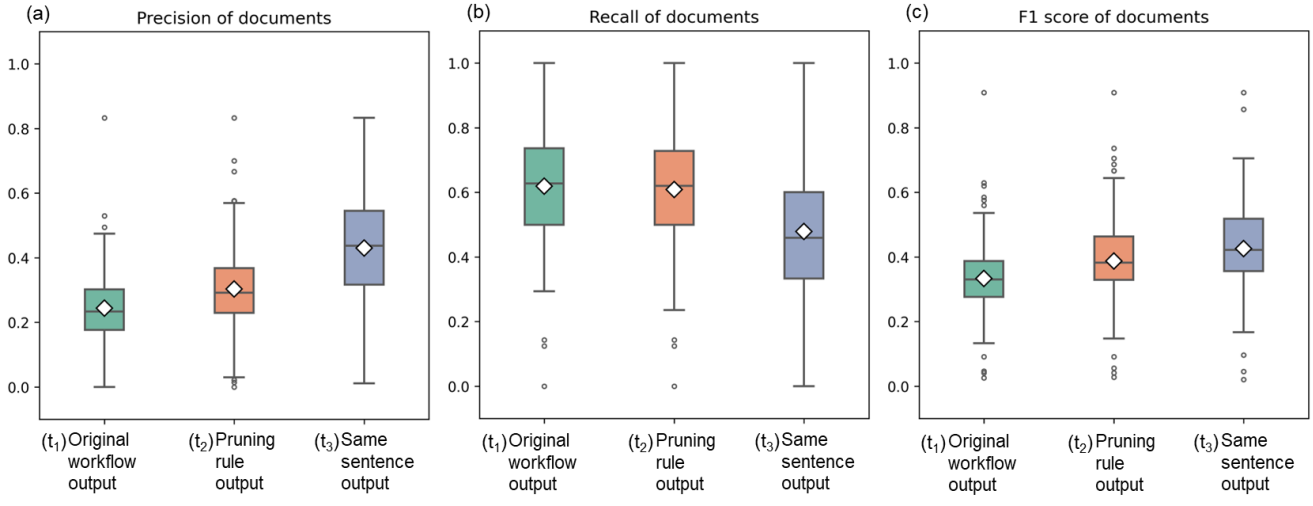


Figure 5: Boxplots for the three different workflow outputs in order to evaluate the entity linking process. The tested metrics are (a) precision, (b), recall and (c) F1 score of the results for each document tested against the reference corpus.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

4.2 Information content of the geolocated DRI

After validating the processing chain and the decision criteria for the final output dataset, we present the results that demonstrate the workflow's potential. In particular, we show that vector points enriched with DRI attributes can be reliably extracted from web text within a specific disaster context.

The initial starting point is 3,890 website links gathered from the GDELT database and after the preprocessing 3,874 website texts are suitable for the analysis. Web text data derived from news webpages can contain various text structures, as the text length varies from a minimum of 6 words to a maximum of 13,826 words. Depending on the text length the websites contain one to several toponyms and related DRIs. We extract a total number of 379,599 such relations. The extracted information is differentiated in its thematic content and specific DRI, the website publication date and the geolocation and its related administrative unit. Therefore, the overall DRI derived from the descriptors is analyzed in the dimensions of time, space and theme. Each point information itself contains a specific disaster-relevant keyword, its thematic classification as well as the toponym information and document origin (Figure 6).

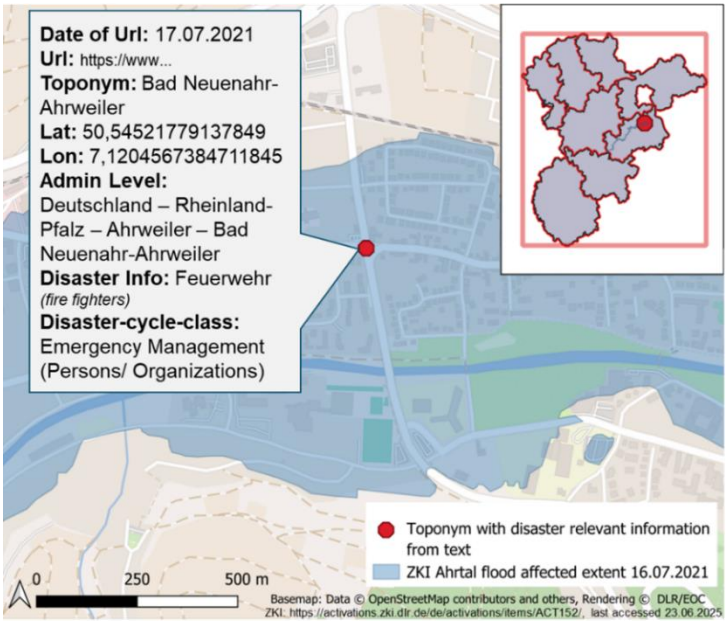


Figure 6: Example representation of the workflow output at the single point level

Beyond analyzing individual points, we summarize the spatial, temporal, and thematic distribution of all extracted information. Figure 7 shows the temporal evolution of all thematic classes in absolute numbers over the course of the disaster phase (dark grey), here defined as the dates of the precipitation and flooding impact and the following after event phase including intensive emergency and mitigation measures (light grey). Our after-event phase is set to 1 ½ months after the event, up to the 1st of September 2021, due to the involvement of major emergency measurements at that time (Gesamtverband der Deutschen Versicherungswirtschaft e. V. (GDV), 2021). The disaster phase exhibits a sharp concentration of events, followed by a steep

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

decline. The flood-related class, which aggregates flood synonyms, disaster, human loss, and infrastructure damage, is most frequently identified within the dataset, while the response class, which comprises emergency-management and people-related terms, accounts for the second-largest share. To examine inter- and intra-class dynamics, the daily proportional distribution of the classes over time can be useful (Appendix A1). These patterns illustrate the flow of information and provide a baseline for future, more detailed comparative analyses, e.g., of the periods of use of emergency management measures, which are beyond the scope of this paper.

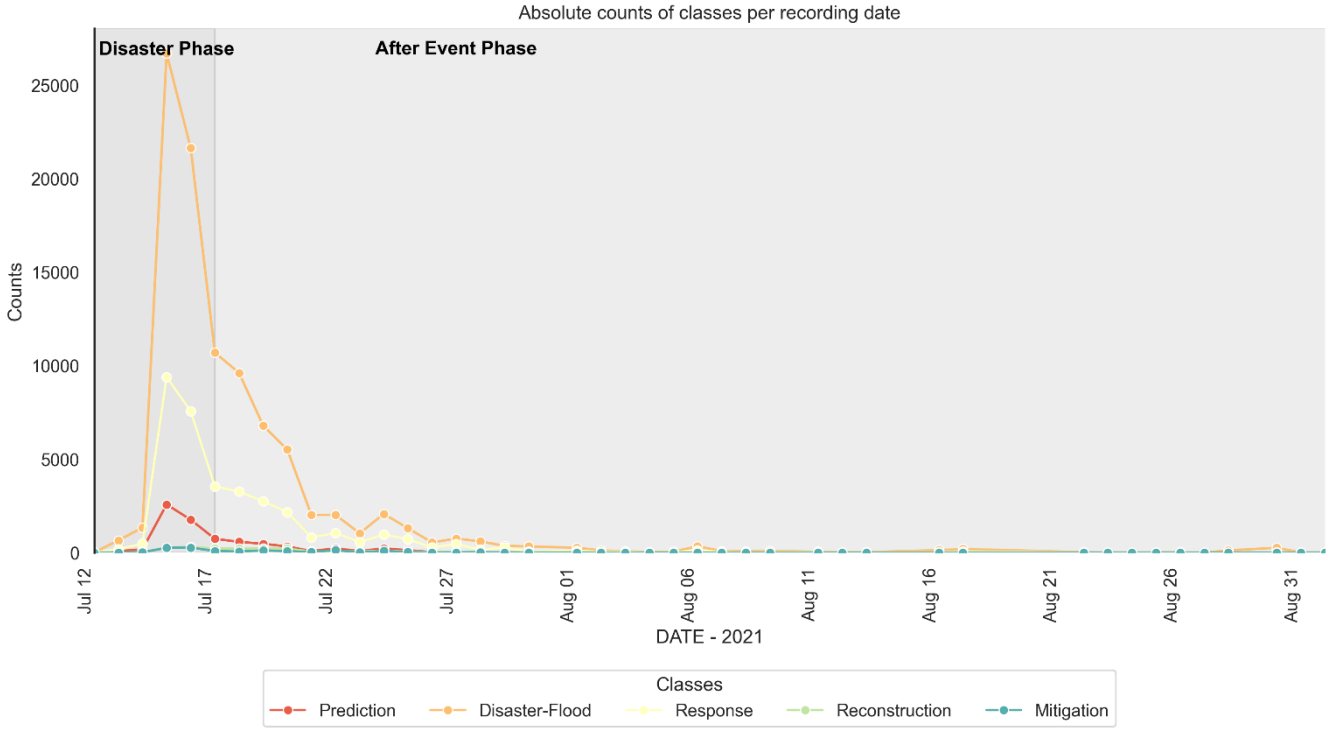


Figure 7: Distribution of DRI information, with the 9 disaster-cycle classes aggregated to five classes over time

4.3 Spatial dissemination of information

By attaching geographic coordinates to DRI extracted from text, we can explore both spatial and thematic patterns of disaster events. Table 2 shows that, within our analyzed documents, our workflow identifies toponym-DRI entities across several countries. This shows, that DRI of a specific and local event also includes DRI connections to toponyms with a global distribution and relevance. Figure 8 highlights that toponyms in texts connected to a disaster with DRI are returned at a variety of administrative levels, which should be taken into consideration when analyzing the disaster, as all levels provide DRI information. Looking at DRI information for the Waldporzheim district, we find direct mentions of this district. In the DRI results, we also have information on the higher administrative levels in which Waldprozheim is located. This includes the

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

municipality of Bad Neuenahr-Ahrweiler, which in turn is located in the district of Ahrweiler in the federal state of Rhineland-Palatinate in Germany. Within our study area, we defined a hierarchical set of administrative units (country, state, county, and municipality) and also incorporated locations that could not be assigned to any of the administrative category. For the latter we can see examples of the tunnel named "Engelsley-Tunnel" and the river Ahr (Figure 8c). Due to the spatial allocation of the toponyms, we have the possibility to aggregate the information spatially. A thematic comparison across these levels reveals that, at the state and county levels, human-loss and injury information is most prominent, whereas at the municipal level infrastructure damage dominates (See Appendix A2). Altogether, these results underscore the value of geocoded DRI for multi-scale disaster analysis and highlight the importance of incorporating all administrative tiers when aggregating information.

Country of occurrence	Absolute number of mentions
Germany	270,068
Belgium	4,218
Austria	3,019
Netherlands	1,928
Switzerland	1,070
Others (99 countries)	6,142 (all together)

Table 2: Global coverage of toponym identification

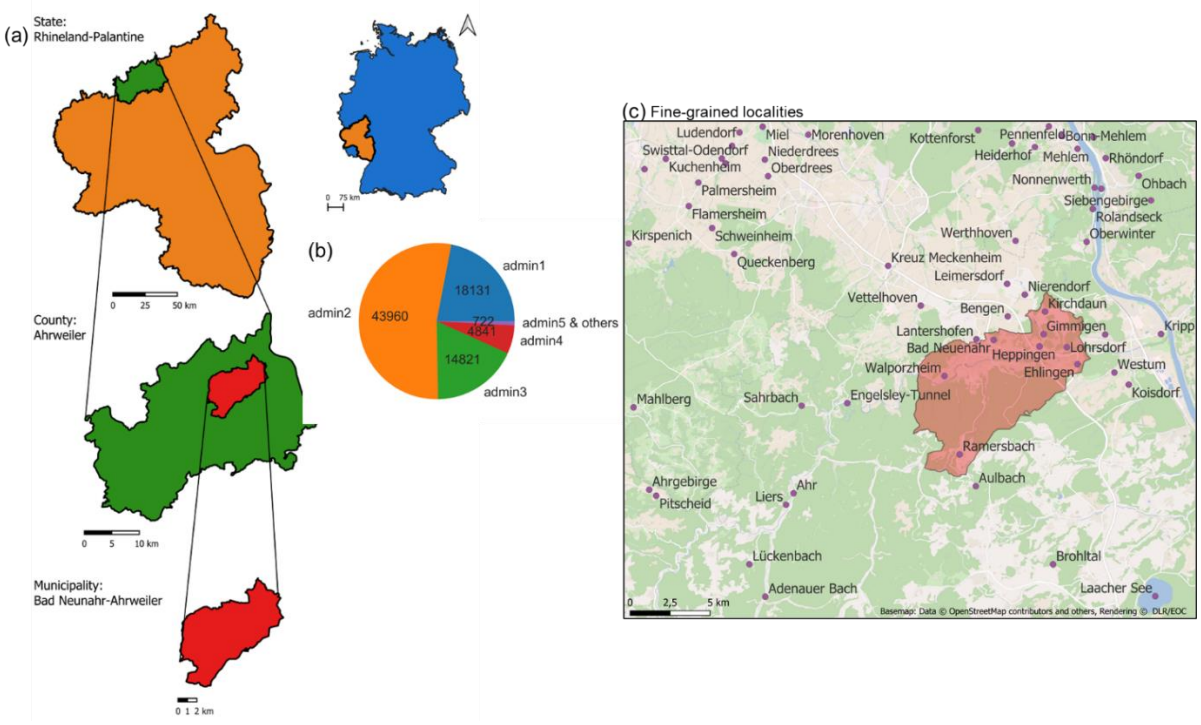


Figure 8: DRI given at different administrative levels. (a) Shows representative administrative units for the municipality Bad Neuenahr-Ahrweiler to the county Ahrweiler, state of Rhineland-Palatinate and the country of Germany. Total number of DRI mentions for the administrative unit are shown in b). (c) Shows examples of the fine-grained toponyms.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Beyond the analysis at individual administrative levels, we examined the spatial distribution of data points across disaster-cycle classes and individual DRI at the inter-unit level. We illustrate the spatial, thematic, and temporal distribution of two of our disaster-cycle classes. These can also be summarised or aggregated. In this way, we aim to demonstrate initial possibilities for spatial analysis and further utilisation of this data. Figure 9 shows the thematic patterns for the disaster phase and the subsequent after event phase at the municipal level, revealing that hotspots differ by disaster-cycle class and mention frequencies. These spatial patterns highlight regions of heightened relevance during the disaster phase and in the days that followed. These temporal shifts are highlighted by the change in the geographic centre of gravity of the respective DRI (black cross in Figure 9) which moves southeast over time. This shift of the geographic centre point between the two temporal snapshots underscores a progressive focus toward the Ahr Valley across all displayed thematic classes. In absolute terms, the disaster phase hotspots are centred on Erftstadt and Bad Neuenahr-Ahrweiler, a finding that holds consistently across all of our defined thematic classes.

In addition to the thematic classes relating to the disaster cycle, the geospatial metadata produced by our approach also allows to analyze the distribution of individual DRI words from the dictionary. As an example, we show in figure 10 the spatial distribution of the connected DRI-toponym pairs of volunteers, which we extracted from the text. We extracted a subset of our entity pairs with all German variations of the word volunteer in our dictionary. Mentions of volunteers can be found in texts about many different administrative units. Specific subsections of towns like “Waldporzheim” in “Bad Neuenahr-Ahrweiler” show an information connection to volunteers. Additionally, we can allocate locations like “Nürburgring” to the term volunteer, outside of areas of interest from the CEMS layer are displayed also in a higher frequency.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

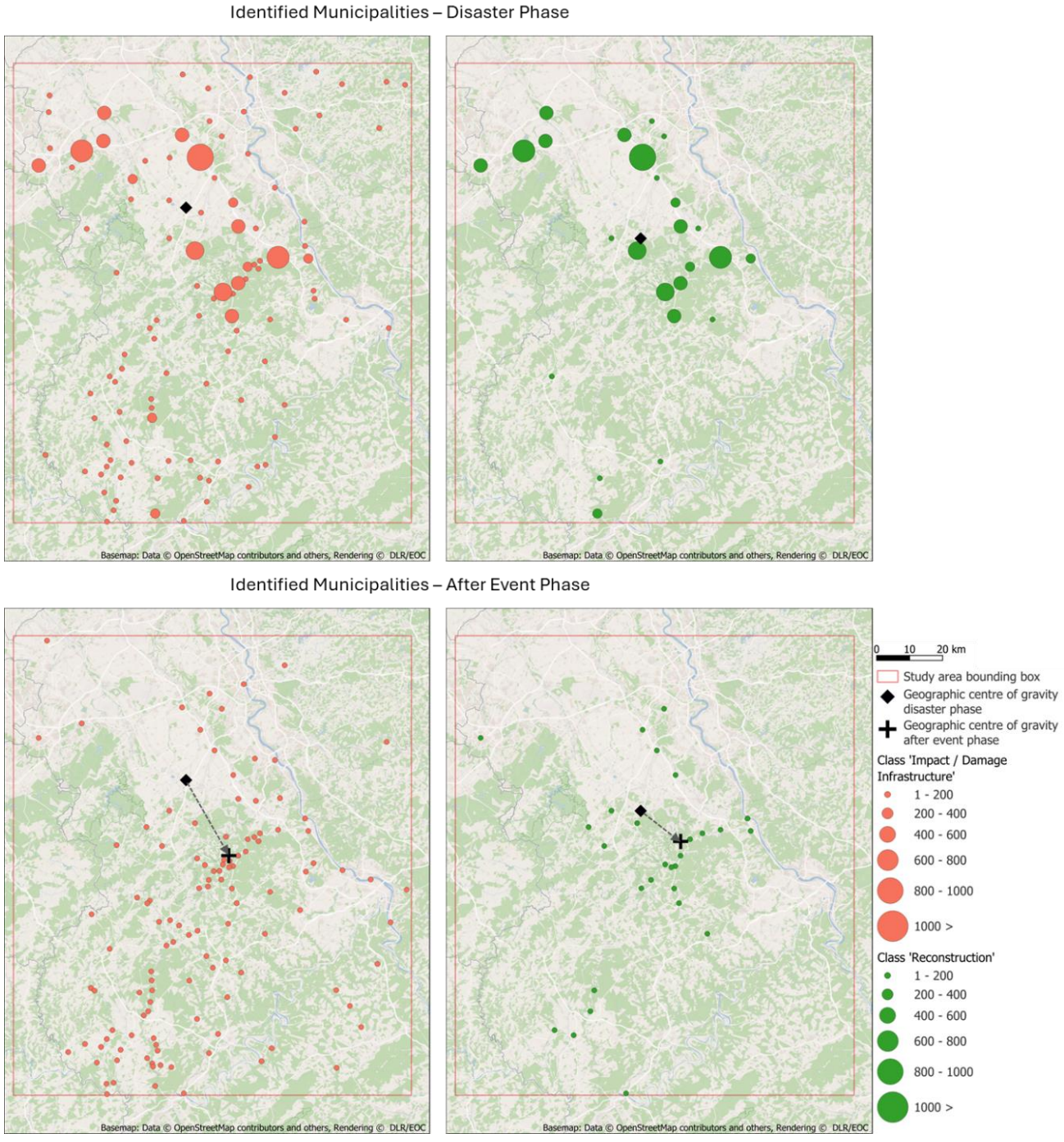


Figure 9: Thematic, spatial, and temporal distribution of two example disaster-cycle classes for municipalities (left: Impact, right: Reconstruction) aggregated for the disaster event period (top) and after event period phase (bottom). The bigger the point size, the more absolute DRI are linked to a toponym. To analyze the spatial distribution and change of focus area over time, the geographic centre of gravity (black symbols) and its change over two periods (dashed line) are shown.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

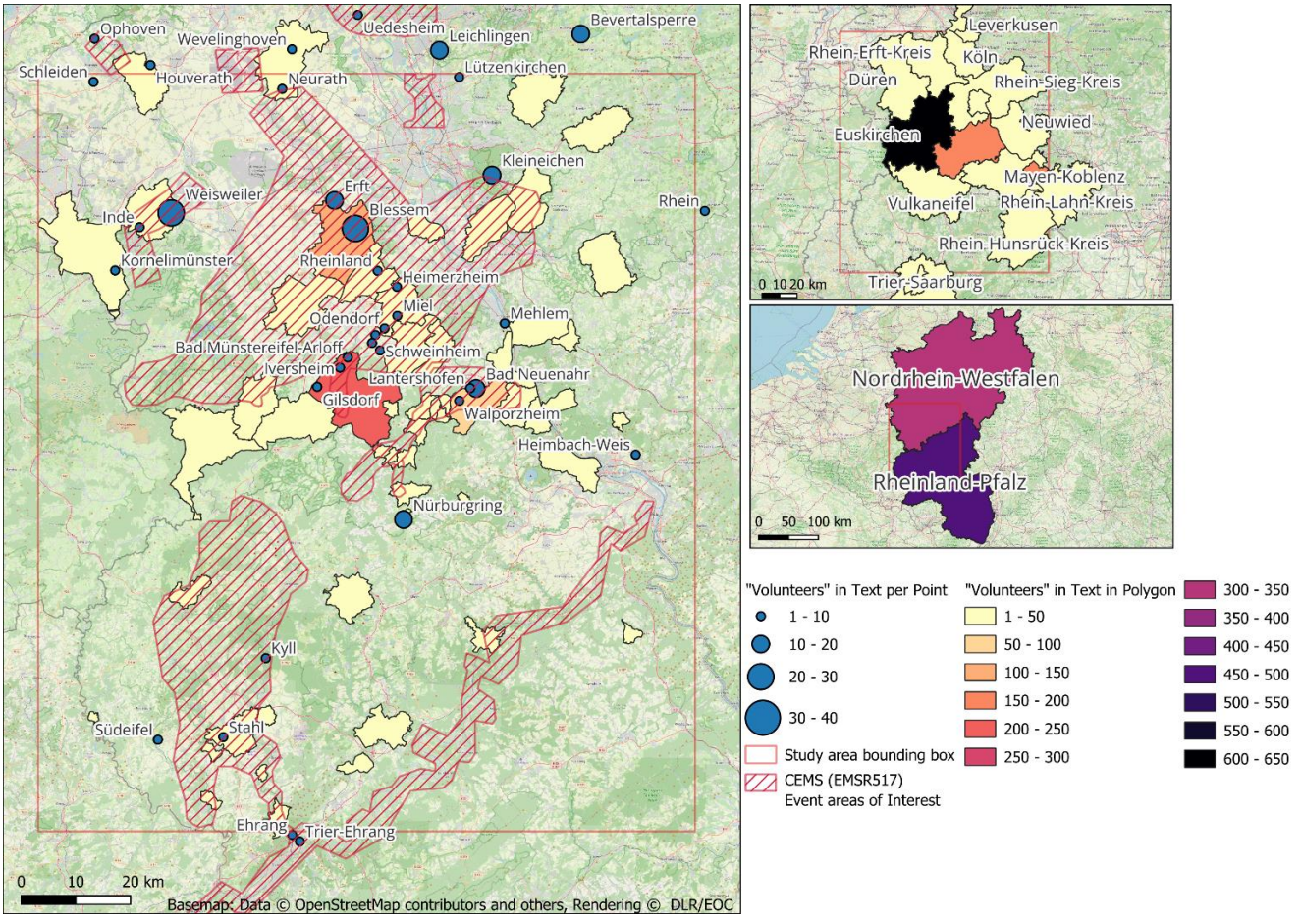


Figure 10: DRI information of volunteer aggregated over time and mapped at the different administrative levels of the linked toponyms

4.4 Linking web text data with other geospatial datasets of the event

The extracted DRI show a different spatial distribution at different administrative levels, as well as an accumulation of information about a specific event. This spatial distribution of the text data allows for joint analysis with additional spatial information about the event. To demonstrate this potential for our case study, we compare the findings from our text data analysis with other available geospatial datasets on building damage and flooding extent.

As an example, figure 11 relates the distributions of toponyms associated with the thematic class "damaged infrastructure" aggregated at the municipality level with building damage reported in the CEMS activation areas on a fine-grained building level. Both datasets show regions of relevance due to the frequency of mentions within an area. A spatial comparison at the municipality level reveals an added information layer to the datasets. Some municipalities like Erftstadt (1) and Bad Neuenahr-Ahrweiler (2) show a high relevance in both datasets. On the other hand, there are municipalities which show a higher relevance in either of the datasets (Schleiden (3) and Kall (4) for the CEMS layer and Cologne (5) and Blankenheim (6) for the text

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

output). Overall, the matching municipalities from both data sources are within the counties that announced the state of emergency. In the CEMS layer dark red spots relate to a high concentration of damaged buildings, whereas blue areas mark identified damaged buildings but of low intensity. This layer also shows the spatial limitation through the data acquisition measures. On the other hand, the results of our text analysis workflow provide us with more comprehensive spatial information over a longer period of time. The comprehensive mentions from the text data reveals a far greater extent of the event. The CEMS layer shows us the location of a damaged and destroyed buildings. On the other hand, the output of the text data includes damage information of various building descriptions in our classification but furthermore also includes bridges, and other damaged infrastructure components. For a deeper analysis, a bivariate comparison at the municipality level, normalized per 1,000 inhabitants, is provided in Appendix A3. This map highlights the spatial similarities in the relevance of the above-mentioned municipalities.

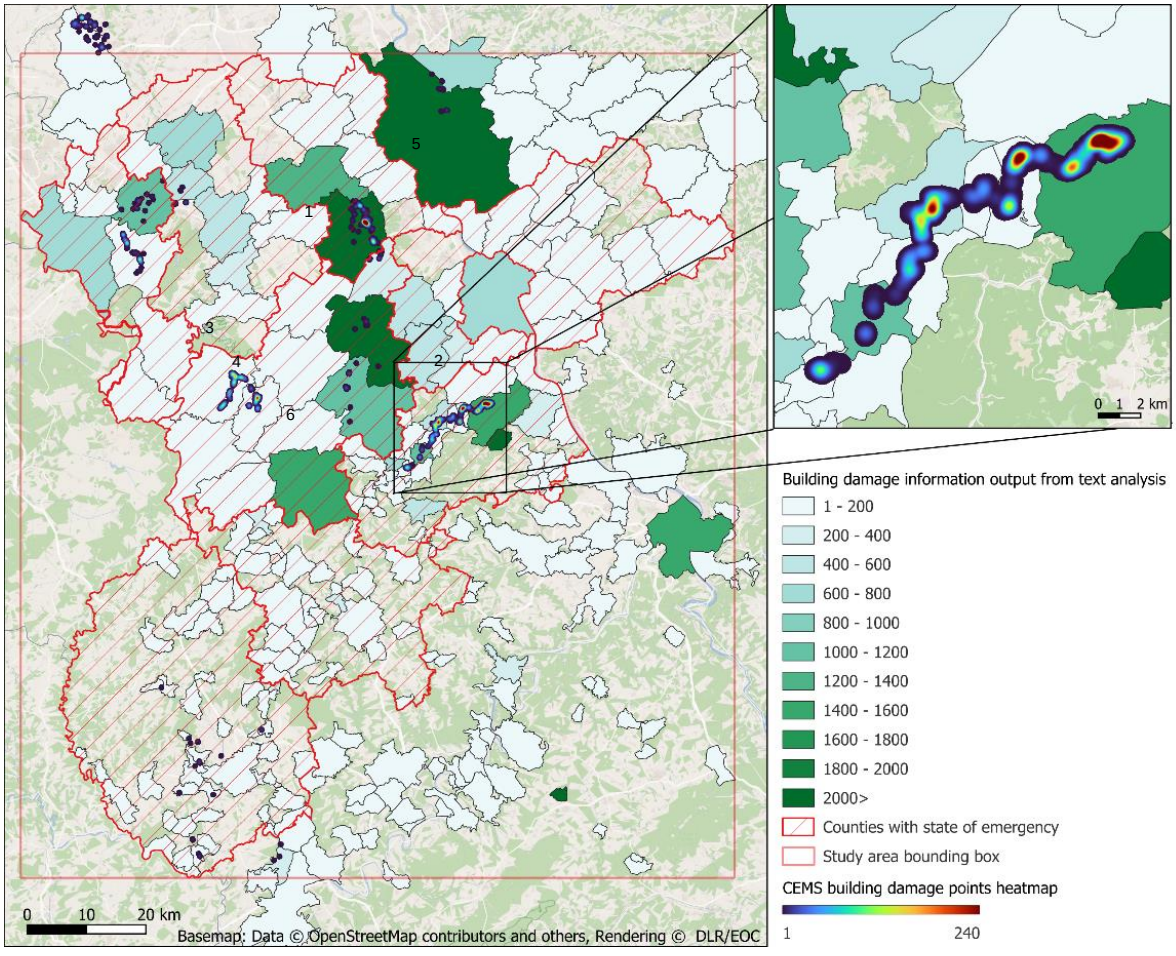


Figure 11: Comparison of disaster damage information from text data (text output in green) and from satellite imagery derived building damage information (CEMS heatmap layer). The numbers 1-6 refer to counties in the text above.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

We use various examples to show how we can obtain DRI from text data at different levels of administrative detail, ranging from nation to state level to counties, cities, and even more specific locations. Based on our disaster cycle -based DRI classification, spatial and temporal distributions and differences of DRI are extracted and highlighted. A comparison of the results from the text data with other spatial damage information shows similarities in their spatial distribution. Furthermore, our pipeline can obtain additional information, including its spatial distribution.

5 Discussion and Outlook

5.1 Methodological approach

A key goal of this study is to identify how disaster management can benefit from georeferenced DRI extraction from text. Our findings show that, through the development of a workflow to connect toponyms and DRI within the texts, we generate results including geospatial, temporal, and thematic information. This information is essential for the analysis of disaster incidences. The information extracted from the text shows spatial and thematic patterns, which can be determined using both quantitative analysis and visual interpretation. This enables a wide range of formats that are suitable for different types of analysis. Once geolocated, information can be aggregated at different spatial units, such as administrative units at different levels, making it possible to compare with other geospatial datasets. This flexibility, provided by spatial joining, is a key benefit enabled by geocoding the text information. With our approach, we can demonstrate how to extract disaster relevant information efficient and scalable.

To evaluate the accuracy, we applied a quantitative evaluation process in which we compared precision, recall and F1 score against manually labelled data. The best overall performance was demonstrated by the workflow output of toponym-DRI relationship pairs that included an integrated pruning post-processing rule. Along the modular workflow, several processing steps contribute individually to uncertainties, and their errors are propagated to subsequent steps. Naturally, our dictionary-based method is dependent on how comprehensive the dictionary is with respect to the disaster context. If a word variation is not in the list, it is not identified in the text. Similar words, synonyms, or other relevant information, that are not included in the dictionary are also missed (Reveilhac and Morselli, 2022), and cannot subsequently be linked to locations by the rule-based approach. Additionally, it is to mention that negations in the sentence structure are currently not identified. Large Language Model (LLM)-based approaches might be more robust when retrieving information from unstructured data and with less expert involvement. For this case study using news data, which is comparatively systematic and formal, our approach of using dictionaries and rule sets is suitable and excels at transparency and speed. However, the evaluation of the dictionary is limited because, to our knowledge, no complete dictionary of disaster-related terms exists.

Concerning the geoparsing, it is clear that toponyms can be missed and words can be incorrectly recognized as toponyms (Gritta et al., 2018b; Hu et al., 2024). 1356 unique toponyms could not be identified in the geoparsing process we used and without further post-processing steps and are therefore not included in our output dataset and results. These include, toponyms identified by the NER model for which there is a description of the surrounding area "Großraum Ahrweiler", a suffix

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

"Eifeldorfs" or a fine-grained geolocalization that is not located in our study area, such as "Felberntauerstraße". Words incorrectly identified as toponyms can therefore be passed to the geocoder, resulting in false or none coordinate allocations. Geocoders contain errors and biases in geographic coverage, and a single toponym can refer to many locations, which may lead to incorrect ambiguity resolution (Gritta et al., 2018a). Our approach first matches toponyms with an official geographic database and then passes unmatched toponyms to the Nominatim geocoding service, using high-quality local public reference information for places that can be matched to this database (Acheson et al., 2017). Out of the final output total of 2,303 unique DRI linked toponyms, 329 could not be geocoded by our first database match. Out those, 114 remained unmatched after the second international database. A short overview of the geocoding results of the databases independently, shows with an exemplary random sample of 230 toponyms (10% of the unique toponyms) different numbers per database are matched. Within the national local database alone, 35 places are not found, with the global open-source datasets "populated places" and GADM, 208 and 88 respectively are not found. Using Nominatim, with a parameter set to a study area focus, 8 toponyms cannot be found. However, it commits some notably erroneous allocations, for example, the toponym Qatar is located at the site of the Qatar embassy in Bonn. Our combination of all the databases and their successive matching allows us to match all of the 230 sample toponyms. Based on the known inaccuracies in the geocoding we furthermore acknowledge the subsequent bias in the allocation of administrative units to the toponyms. Incorrectly assigned coordinates are transferred there in our spatial matching of administrative levels.

In our last methodological step, the linking of the toponyms and DRI, the rule-based entity linking process shows limitations based on our evaluation metric values. We therefore use the statistical distribution of sentence distances between toponym and DRI, derived from our reference data, as a proxy to define our matching. We acknowledge a high false positive rate and low precision values in this method. Furthermore, the relationships among multiple toponyms and multiple DRI within a single sentence are currently not considered, which could lead to false allocation of DRI to toponyms. However, we achieve a straightforward, simple, and fast way to contextualise information that requires specific requirements from linking toponyms and DRI.

5.2 Data source quality and fake information

The substantial data volume generated by aggregating news content from the web presents significant challenges for systematic organization and quality assurance. The GDELT Project serves as a major aggregator of global news data, however, its data processing methods are not fully disclosed, putting constraints on the interpretability of derived analyses in research contexts (Hoffmann et al., 2022). In our study, we applied thematic pre-filtering during data collection and then extracted the text content of the websites. Therefore, we use the GDELT GKG database as a starting point for creating a list of websites, to which we then apply a processing pipeline that was fine-tuned for the context of the case study. To accomplish this, we also download the text content of the web pages independently. Another discussion point concerns the date linked to the output. We currently focus only on the date that is taken from the GDELT metadata for each of the websites. Specific date and time information in the text that relates directly to the toponym and DRI is currently not integrated in our approach. Accordingly,

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

we can use the time information in this analysis to show when the website with the text content was registered in the database, at the document level. However, this does not currently allow us to link the individual extracted DRI to an exact point in time but provides potential for further integration of NLP date extraction methods.

Given the documented risk of misinformation in online content (Abraham et al., 2025; Lemoine-Rodríguez et al., 2024), we recognize potential challenges in data reliability. This analysis focuses on analysing the full range of available text data to find out where, when and for which disaster cycle class information is available.

Currently, our data analysis is validated based on the labelling of toponym-DRI relations by experts and a comparison of the output to published geospatial datasets. In the future, more sophisticated data analysis of the web texts could be performed by comparing the data with various other event-related datasets, e.g. in-situ data with physio-geographical information (Madruga de Brito et al., 2025).

5.3 Results of workflow output and DRI-toponym information

While the analysis is constrained in its potential depth by the limitations of the data source and the discussed above dictionary-based DRI information extraction and classification, the case study reveals notable similarities between the geographic relevance inferred from the textual data, the officially declared states of emergency, and the spatial extents captured by ancillary geographic data products. This suggests that the approach, despite the methodological constraints, captures meaningful spatial patterns that are useful for disaster management applications.

Frequent mentions of emergency organizations, which are also found in the text data, especially during the disaster phase, align with findings from checking specific event reports. For example, the German Armed Forces, police, THW, BBK, and police officers are mentioned (BMI, 2021), which can be extracted as emergency organizations from our workflow. Moreover, a pronounced cluster of volunteers near the Nürburgring aligns with the known assembly point for helpers and emergency services (Fekete et al., 2022). These two examples illustrate the possibilities and variety of specific details that we can extract from a disaster and enable us to localize them spatially.

The results indicate a promising avenue for future work: automated detection of assembly points and emergency-service locations, which can then be leveraged in subsequent accessibility or network analyses to support operational decision-making in disaster scenarios. In addition, the observation of further events like the monitoring of slowly building events using this method could be considered. Beyond the spatial extraction performed here, the underlying text contains rich semantic structures that could be mined using advanced NLP techniques, such as sentiment analysis (Hanny et al., 2025). However, investigating these deeper semantic relationships lies outside the scope of the present study and is reserved for future research.

5.4 Outlook and future scope of study

For our analysis and its objectives, we methodically relied on rule-based and machine learning models that do not require a great deal of training to perform a proof-of-concept analysis. In this study, we outline how we can obtain further disaster-relevant information from texts in addition to the existing data products for an event and place this information in a spatial

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

context. Sophisticated NLP methods should be further developed for flood-related information extraction (Navalkar et al., 2025). Our results suggest substantial potential for text extraction with a focus on DRI and its georeferencing. Based on successful NER models in other domains (Fu et al., 2022; Leitner et al., 2019) and the demonstration of the benefits of text data in disaster analyses (Madruga de Brito et al., 2025), pursuing this approach with more automated methods could be beneficial. The labelled reference corpus of toponym-DRI pairs for the validation could therefore serve as the basis for training such models or could be integrated into an active learning process (Agarwal et al., 2024).

In the context of disasters, the automatic spatial localization of relevant information can be highly beneficial for understanding the complexity of the event. For this task, LLMs are already used to localize disaster-related social media posts (Yin et al., 2025). The extraction of relationships using LLM prompting and without fine-tuning for fine-grained toponym recognition has already been tested by Chen et al., (2022). As mentioned earlier, however, these models require a good base of training and validation data.

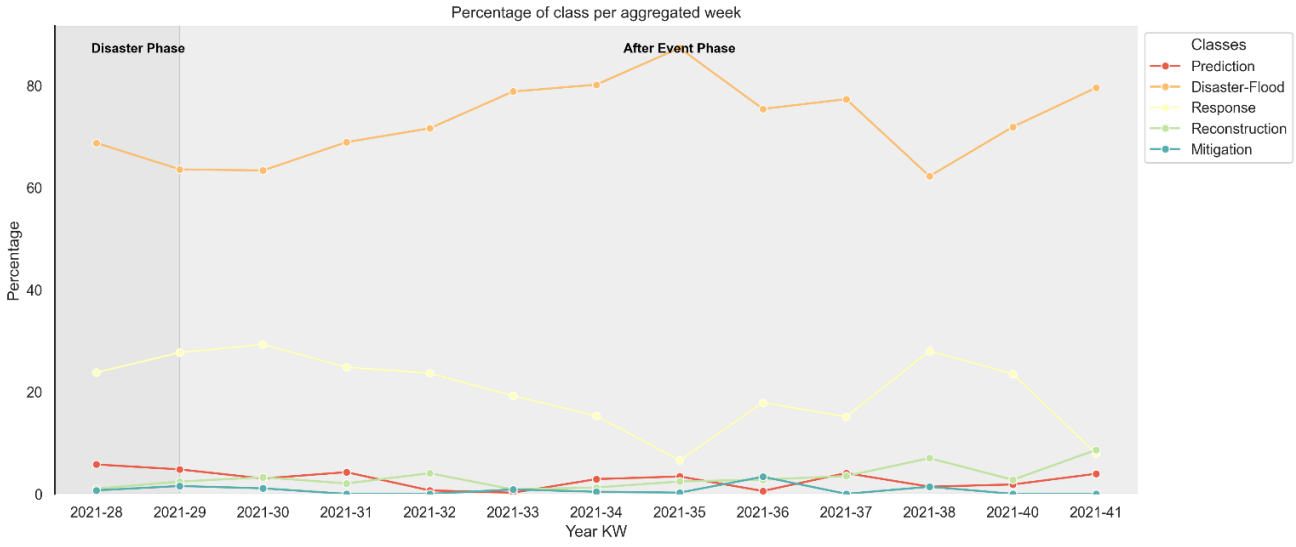
This case study focused on text data in German. This language restriction leads to the exclusion of data that could potentially be useful for the analysis, as flooding events occur worldwide and information is also published on websites in different languages. Our dictionary, and therefore our analysis, are however, only designed for German text. Future research could explore the use of multilingual LLM models (Yin et al., 2025) in approaches such as the one demonstrated in this study. This could enable the identification and linking of DRI and toponyms across a much wider range of languages and, therefore, enlarge and diversify the data base for disaster analyses.

6 Conclusion

This research demonstrates a systematic approach for extracting and geolocating relevant information about a flooding event from web text data. It identifies and structures this information in a detailed flood disaster-cycle-phase classification and geolocates it with fine-grained geolocations. Our methodological combination of rules, a dictionary, and a ML-model shows great potential for extracting specific disaster indicators from text and the potential of linking it to toponyms across a variety of administrative units. With improvements in the entity extraction and linking models, limitations can be mitigated in the future and new applications can build upon this study. In our case study on the 2021 floods in West Germany, it became evident that text data can complement existing geospatial data sources with additional detail, such as information on involved emergency organizations, emergency response measures, and mitigation measures. The gathered data can be used as an additional data layer for disaster management and response. This supports the collection and improvement of knowledge and preparedness for future events and related disaster situations.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Appendix



505 **Figure A1: Proportional distribution of the disaster cycle classes for each day**

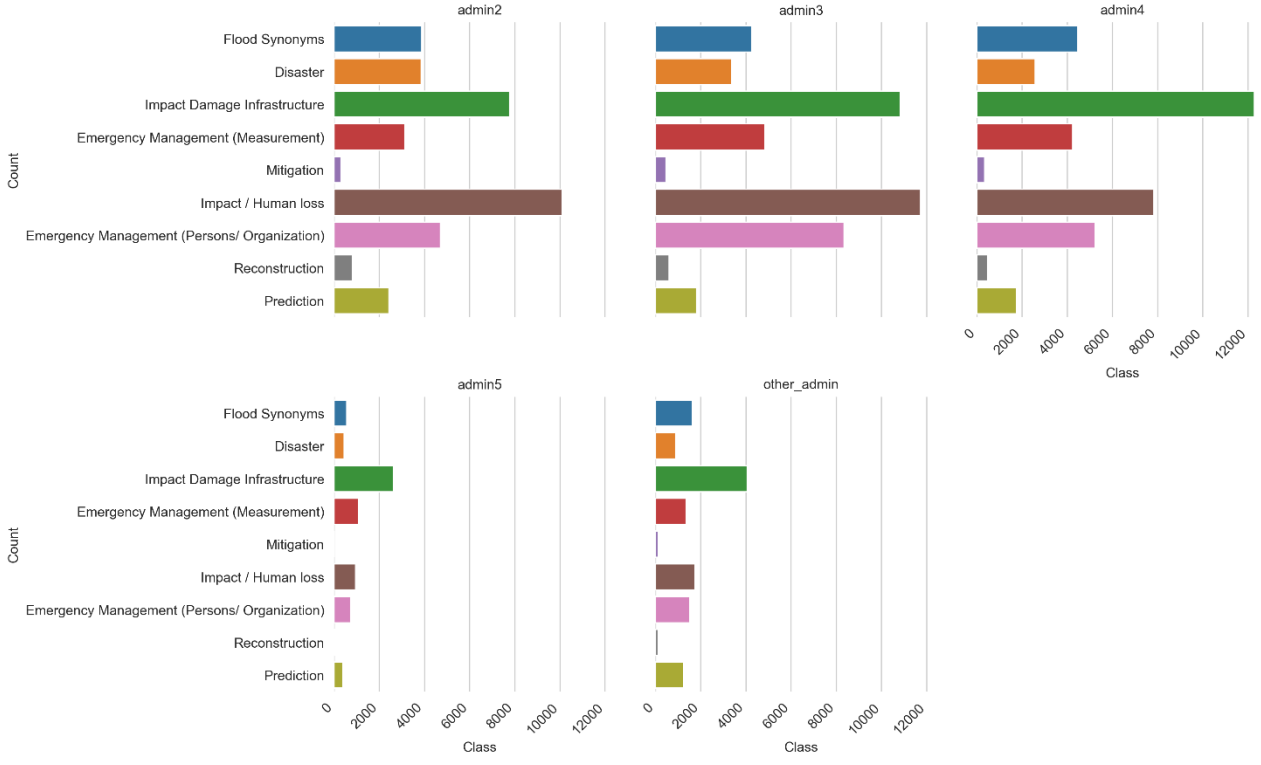


Figure A2: Thematic classes distribution across the different admin levels in the study area and aggregated of disaster- and after event phase

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

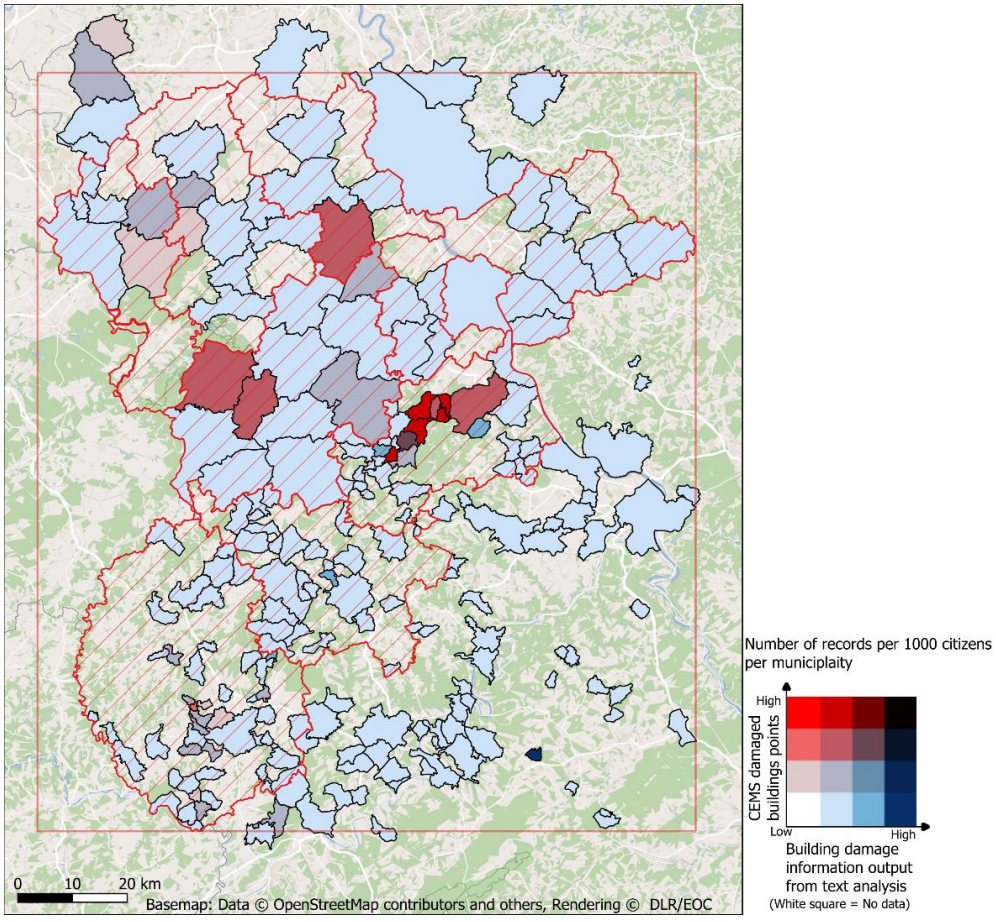


Figure A3: Bivariate comparison of CEMS building damage information and the building damage information output from the text analysis. The data is normalized per 1000 citizens per municipality.

Code and data availability

The code is available on request from the corresponding author. All the data are freely available and the sources are cited in the text.

Author contribution

VR: Conceptualization: VR, JM, SV, CG, HT; Methodology: VR, JM, SV; Software, validation, visualization, formal analysis: VR; Supervision: SV, CG, HT; Funding acquisition: SV, CG, HT; Writing-original: VR; Writing-reviewing & editing: VR, JM, SV, CG, HT. All authors have read and agreed to the published version of the manuscript.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

The authors would like to thank Patryk Pawel Gadziomski for his support with data processing and analysis.

Financial support

This study was conducted within the framework of the OpenSearch@DLR phase II project (internal DLR project).

References

Abraham, T. M., Wen, T., Wu, T., and Chen, Y.: Leveraging data analytics for detection and impact evaluation of fake news and deepfakes in social networks, Humanit. Soc. Sci. Commun., 12, 1040, https://doi.org/10.1057/s41599-025-05389-4, 2025.

Acheson, E., De Sabbata, S., and Purves, R. S.: A quantitative analysis of global gazetteers: Patterns of coverage for common feature types, Comput. Environ. Urban Syst., 64, 309–320, https://doi.org/10.1016/j.compenvurbsys.2017.03.007, 2017.

Agarwal, A., Kapuriya, J., Agrawal, S., Konam, A. V., Goel, M., Gupta, R., Rastogi, S., Niharika, N., and Bagler, G.: Deep Learning Based Named Entity Recognition Models for Recipes, 2024.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R.: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations), Minneapolis, Minnesota, 54–59, https://doi.org/10.18653/v1/N19-4010, 2019.

EMSR517 - Copernicus EMS Mapping | Copernicus EMS On Demand Mapping: https://mapping.emergency.copernicus.eu/activations/EMSR517/, last access: 8 January 2026.

BMI: Zwischenbericht zur Flutkatastrophe 2021, 2021.

Bosher, L., Chmutina, K., and Niekerk, D.: Stop going around in circles: towards a reconceptualisation of disaster risk management phases, Disaster Prev. Manag. Int. J., ahead-of-print, https://doi.org/10.1108/DPM-03-2021-0071, 2021.

Chen, Y., Harbecke, D., and Hennig, L.: Multilingual Relation Classification via Efficient and Effective Prompting, https://doi.org/10.48550/arXiv.2210.13838, 26 October 2022.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

De Albuquerque, J. P., Herfort, B., Brenning, A., and Zipf, A.: A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management, Int. J. Geogr. Inf. Sci., 29, 667–689, https://doi.org/10.1080/13658816.2014.996567, 2015.

De Bruijn, J. A., De Moel, H., Weerts, A. H., De Ruiter, M. C., Basar, E., Eilander, D., and Aerts, J. C. J. H.: Improving the classification of flood tweets with contextual hydrological information in a multimodal neural network, Comput. Geosci., 140, 104485, https://doi.org/10.1016/j.cageo.2020.104485, 2020.

Eyre, R., De Luca, F., and Simini, F.: Social media usage reveals recovery of small businesses after natural hazard events, Nat. Commun., 11, 1629, https://doi.org/10.1038/s41467-020-15405-7, 2020.

Fekete, A. and Sandholz, S.: Here Comes the Flood, but Not Failure? Lessons to Learn after the Heavy Rain and Pluvial Floods in Germany 2021, Water, 13, 3016, https://doi.org/10.3390/w13213016, 2021.

Fekete, A., Beckers, Daniel, and Hetkämper, Chris: Die Flut im Juli 2021 Erfahrungen und Perspektiven aus dem Rettungsingenieurwesen und Katastrophenrisikomanagement, TH Köln - University of Applied Sciences, 2022.

Fu, S., Lyu, H., Wang, Z., Hao, X., and Zhang, C.: Extracting historical flood locations from news media data by the named entity recognition (NER) model to assess urban flood susceptibility, J. Hydrol., 612, 128312, https://doi.org/10.1016/j.jhydrol.2022.128312, 2022.

GADM Project: Global Administrative Areas (GADM) Database, Version 3.6, 2025.

GDELT Project: GDELT-Global Knowledge Graph Codebook V2.1, 2015.

Gesamtverband der Deutschen Versicherungswirtschaft e. V. (GDV): Naturgefahrenreport 2021 - Die Schaden-Chronik der deutschen Versicheret, 2021.

Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N.: What's missing in geographical parsing?, Lang. Resour. Eval., 52, 603–623, https://doi.org/10.1007/s10579-017-9385-8, 2018a.

Gritta, M., Pilehvar, M. T., and Collier, N.: Which Melbourne? Augmenting Geocoding with Maps, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 1285–1296, https://doi.org/10.18653/v1/P18-1119, 2018b.

Hanny, D., Schmidt, S., and Resch, B.: Bluesky as a social media data source for disaster management: investigating spatio-temporal, semantic and emotional patterns for floods and wildfires, J. Comput. Soc. Sci., 9, https://doi.org/10.1007/s42001-025-00448-x, 2025.

Hoffmann, M., Santos, F. G., Neumayer, C., and Mercea, D.: Lifting the Veil on the Use of Big Data News Repositories: A Documentation and Critical Discussion of A Protest Event Analysis, Commun. Methods Meas., 16, 283–302, https://doi.org/10.1080/19312458.2022.2128099, 2022.

Hu, X., Zhou, Z., Li, H., Hu, Y., Gu, F., Kersten, J., Fan, H., and Klan, F.: Location Reference Recognition from Texts: A Survey and Comparison, ACM Comput. Surv., 56, 1–37, https://doi.org/10.1145/3625819, 2024.

Kahle, M., Kempf, M., Martin, B., and Glaser, R.: Classifying the 2021 'Ahrtal' flood event using hermeneutic interpretation, natural language processing, and instrumental data analyses, Environ. Res. Commun., 4, 051002, https://doi.org/10.1088/2515-7620/ac6657, 2022.

Kharis, M., Kisyani, Suhartono, Pairin, U., and Darni: How to Lemmatize German Words with NLP-Spacy Lemmatizer?:, International Seminar on Language, Education, and Culture (ISoLEC 2021), Malang, Indonesia, https://doi.org/10.2991/assehr.k.211212.036, 2021.

Klonner, C., Marx, S., Usón, T., Porto De Albuquerque, J., and Höfle, B.: Volunteered Geographic Information in Natural Hazard Analysis: A Systematic Literature Review of Current Approaches with a Focus on Preparedness and Mitigation, ISPRS Int. J. Geo-Inf., 5, 103, https://doi.org/10.3390/ijgi5070103, 2016.

Lai, K., Porter, J. R., Amodeo, M., Miller, D., Marston, M., and Armal, S.: A Natural Language Processing Approach to Understanding Context in the Extraction and GeoCoding of Historical Floods, Storms, and Adaptation Measures, Inf. Process. Manag., 59, 102735, https://doi.org/10.1016/j.ipm.2021.102735, 2022.

Leetaru, K. and Schrodt, P. A.: GDELT: Global Data on Events, Location and Tone, 2013.

Leitner, E., Rehm, G., and Moreno-Schneider, J.: Fine-Grained Named Entity Recognition in Legal Documents, in: Semantic Systems. The Power of AI and Knowledge Graphs, vol. 11702, edited by: Acosta, M., Cudré-Mauroux, P., Maleshkova, M., Pellegrini, T., Sack, H., and Sure-Vetter, Y., Springer International Publishing, Cham, 272–287, https://doi.org/10.1007/978-3-030-33220-4_20, 2019.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Leitner, E., Rehm, G., and Moreno-Schneider, J.: A Dataset of German Legal Documents for Named Entity Recognition, http://arxiv.org/abs/2003.13016, 29 March 2020.
Lemoine-Rodríguez, R., Mast, J., Mühlbauer, M., Mandery, N., Biewer, C., and Taubenböck, H.: The voices of the displaced: Mobility and Twitter conversations of migrants of Ukraine in 2022, Inf. Process. Manag., 61, 103670, https://doi.org/10.1016/j.ipm.2024.103670, 2024.
Madruga de Brito, M., Sodoge, J., Kreibich, H., and Kuhlicke, C.: Comprehensive Assessment of Flood Socioeconomic Impacts Through Text-Mining, Water Resour. Res., 61, e2024WR037813, https://doi.org/10.1029/2024WR037813, 2025.
naturalearthdata.com: Natural Earth » 1:10m Cultural Vectors - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales, 2025.
Navalkar, A., Tripathy, S. S., Gupta, M., Basu, S., Tripathy, P., Banerjee, A., Sekharan, S., Murtugudde, R., and Ghosh, S.: NLP-driven crowdsourcing for urban flood monitoring: insights from mumbai, Sustain. Cities Soc., 132, 106795, https://doi.org/10.1016/j.scs.2025.106795, 2025.
OpenStreetMap contributors: Nominatim (geocoding service), 2025.
Owuor, I., Hochmair, H. H., and Cvetojevic, S.: Tracking Hurricane Dorian in GDELT and Twitter, AGILE GIScience Ser., 1, 1–18, https://doi.org/10.5194/agile-giss-1-19-2020, 2020.
Reveilhac, M. and Morselli, D.: Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data, Polit. Res. Exch., 4, 2029217, https://doi.org/10.1080/2474736X.2022.2029217, 2022.
Sänger, M., Garda, S., Wang, X. D., Weber-Genzel, L., Droop, P., Fuchs, B., Akbik, A., and Leser, U.: HunFlair2 in a cross-corpus evaluation of biomedical named entity recognition and normalization tools, https://doi.org/10.48550/arXiv.2402.12372, 20 February 2024.
Saz-Carranza, A., Quer, X., and Maturana, P.: The Empirical Use of GDELT Big Data in Academic Research, 2020.
Schäfer, A., Mühr, B., Daniell, J., Ehret, U., Ehmele, F., Küpfer, K., Brand, J., Wisotzky, C., Skapski, J., Rentz, L., Mohr, S., and Kunz, M.: Hochwasser Mitteleuropa, Juli 2021 (Deutschland) : 21. Juli 2021 – Bericht Nr. 1 „Nordrhein-Westfalen & Rheinland-Pfalz", Karlsruher Institut für Technologie (KIT), https://doi.org/10.5445/IR/1000135730, 2021.
Schiersch, M., Mironova, V., Schmitt, M., Thomas, P., Gabryszak, A., and Hennig, L.: A German Corpus for Fine-Grained Named Entity Recognition and Relation Extraction of Traffic and Industry Events, https://doi.org/10.48550/arXiv.2004.03283, 7 April 2020.
Senaratne, H., Mühlbauer, M., Götzer, S., Riedlinger, T., and Taubenböck, H.: Detecting crisis events from unstructured text data using signal words as crisis determinants, Int. J. Digit. Earth, 16, 4601–4620, https://doi.org/10.1080/17538947.2023.2278714, 2023a.
Senaratne, H., Mühlbauer, M., Kiefl, R., Cárdenas, A., Prathapan, L., Riedlinger, T., Biewer, C., and Taubenböck, H.: The Unseen—An Investigative Analysis of Thematic and Spatial Coverage of News on the Ongoing Refugee Crisis in West Africa, ISPRS Int. J. Geo-Inf., 12, 175, https://doi.org/10.3390/ijgi12040175, 2023b.
Serere, H. N., Resch, B., and Havas, C. R.: Enhanced geocoding precision for location inference of tweet text using spaCy, Nominatim and Google Maps. A comparative analysis of the influence of data selection, PLOS ONE, 18, e0282942, https://doi.org/10.1371/journal.pone.0282942, 2023.
Shukla, D., Azad, H. K., Abhishek, K., and Shitharth, S.: Disaster management ontology- an ontological approach to disaster management automation, Sci. Rep., 13, 8091, https://doi.org/10.1038/s41598-023-34874-6, 2023.
Sun, J., Liu, Y., Cui, J., and He, H.: Deep learning-based methods for natural hazard named entity recognition, Sci. Rep., 12, https://doi.org/10.1038/s41598-022-08667-2, 2022.
Suwaileh, R., Elsayed, T., Imran, M., and Sajjad, H.: When a disaster happens, we are ready: Location mention recognition from crisis tweets, Int. J. Disaster Risk Reduct., 78, 103107, https://doi.org/10.1016/j.ijdrr.2022.103107, 2022.
Thompson, S. K.: Sampling, John Wiley & Sons, Incorporated, Newark, UNITED STATES, 2012.
Warto, Rustad, S., Shidik, G. F., Noersasongko, E., Purwanto, Muljono, and Setiadi, D. R. I. M.: Systematic Literature Review on Named Entity Recognition: Approach, Method, and Application, Stat. Optim. Inf. Comput., 12, 907–942, https://doi.org/10.19139/soic-2310-5070-1631, 2024.
Wieland, M., Schmidt, S., Resch, B., Abecker, A., and Martinis, S.: Fusion of geospatial information from remote sensing and social media to prioritise rapid response actions in case of floods, Nat. Hazards, https://doi.org/10.1007/s11069-025-07120-7, 2025.

https://doi.org/10.5194/egusphere-2026-2361
Preprint. Discussion started: 5 June 2026

Yin, W., Xue, Y., Liu, Z., Li, H., and Werner, M.: LLM-enhanced disaster geolocalization using implicit geoinformation from multimodal data: A case study of Hurricane Harvey, Int. J. Appl. Earth Obs. Geoinformation, 137, 104423, https://doi.org/10.1016/j.jag.2025.104423, 2025.

Zhu, X. X., Wang, Y., Kochupillai, M., Werner, M., Häberle, M., Hoffmann, E. J., Taubenböck, H., Tuia, D., Levering, A., Jacobs, N., Kruspe, A., and Abdulahhad, K.: Geoinformation Harvesting From Social Media Data: A community remote sensing approach, IEEE Geosci. Remote Sens. Mag., 10, 150–180, https://doi.org/10.1109/MGRS.2022.3219584, 2022.