An automated approach for developing geohazard inventories using news: Integrating NLP, machine learning, and mapping.

Aydoğan Avcıoğlu[1], Ogün Demir[2], Tolga Görüm[3]

[1]BRGM, 3 avenue Claude Guillemin, Orléans, 45060, France
[2]Nezahat Gökyiğit Botanic Garden, Biodiversity Information Department, İstanbul, Türkiye
[3]Eurasia Institute of Earth Sciences, Istanbul Technical University, İstanbul, Türkiye

Correspondence to: Aydoğan Avcıoğlu (a.avcioglu@brgm.fr)

Abstract. Spatiotemporal inventories of ~~natural hazard~~geohazard~~s~~ are essential for comprehending the building of resilient societies; yet, restricted access to global inventories hinders the advancement of mitigation strategies. Consequently, we developed an approach that enhances the capability of online newspapers in the creation of ~~natural hazard~~geohazard inventory by utilizing web scraping, natural language processing (NLP), clustering, and geolocation of textual data. Here, we use the online newspapers from 1997 to 2023 in Türkiye to employ our approach. In the first stage, we retrieved 15~~,~~ 569 news by using our tr-news-scraper tool considering wildfire, flood, landslide, and sinkhole-related ~~natural hazard~~geohazard news. Further, we utilized NLP preprocessing approaches to refine the raw texts obtained from newspaper sources, which were subsequently clustered into 4 ~~natural hazard~~geohazard groups resulting in 3928 news. In the final stage of the approach, we developed a method, which geolocates the news using the Open Street Map (OSM) Nominatim tool, ending up with a total of 13 940 ~~natural hazard~~geohazard incidents derived from news comprising multiple incidents across various locations. As a result, we mapped 9609 floods, 1834 wildfires, 1843 landslides, and 654 sinkhole formation incidents from online newspaper sources, showing spatiotemporally consistent distribution with existing literature. Consequently, we illustrated the potential of online newspaper articles in the development of ~~natural hazard~~geohazard inventories with our approach from the web sources as text data to map by leveraging the capabilities of web scraping, NLP, and mapping techniques.

1 Introduction

Geohazards are direct threats to human life, ecosystems, and societies worldwide socio-economically, demanding ongoing innovation and development in the mapping, analysis, and monitoring of these events~~Natural disasters are vital and direct threats to human life, ecosystems, and societies worldwide socio-economically, demanding ongoing innovation and development in the mapping, analysis, and monitoring of these events~~. The average annual economic loss due to ~~natural hazard~~geohazards has been 34 billion US$ since 1900 according to The Emergency Management Database (CRED. 2023). According to the Sendai Framework, over 700 thousand people have lost their lives and >1.5 billion people have suffered from ~~natural hazard~~geohazards a period between 2005 – 2015 (UNISDR 2015, p. 12). However, the assessment of damage

and loss from different ~~natural disasters~~geohazards might be underestimated. Because massive events such as large earthquakes and extreme wildfire events tend to trigger and cause subsequent disasters, such as landslides, debris flows, flooding, soil erosion, etc. For instance, the Wenchuan earthquake, Mw 7.9 (USGS, 2008) triggered more than 60 ,000 landslides (Görüm et al., 2011) and its prolonged impact can influence even after years via landslide damming which is recognized hazard due to the ability to release outbursts floods (Delaney and Evans, 2015; Fan et al., 2012, 2019; Peng and Zhang, 2012). Therefore, the lack of thorough evaluations and databases customized for specific hazard types causes temporal delays that hinder the understanding of dynamics of temporal and spatial probability and overall socio-economic and environmental losses of ~~natural hazard~~geohazards. This obstacle results from the differences in data collection and monitoring practices among countries, each of which is subject to different legal frameworks.

Spatiotemporal archive inventorying is crucial for a better understanding of susceptibility, hazard, and risk assessment of ~~natural hazard~~geohazards (Tanyaş et al., 2017, 2022; Loche et al., 2022; Gómez et al., 2023; Stein et al., 2024; Bhuyan et al., 2024). Also, these inventories can provide an objective base for resilience and preparedness for disaster risk reduction strategies (Jones et al., 2022). For instance, a well-known database, the Emergency Events Database (EM-DAT) operated under the Centre for Research on the Epidemiology of Disasters (CRED) provides a wide range of ~~natural disaster~~geohazard inventory (subgroups: Geophysical, Hydrological, Meteorological, Climatological, Biological, and Extra-terrestrial) with their corresponding casualties and economic loss (Guha et al. 2015). However, some studies (Froude and Petley, 2018; Görüm and Fidan, 2021; Haque et al., 2016; Stein et al., 2024) highlighted that EM-DAT lacks a thorough assessment of the ~~natural disaster~~geohazards since it includes the events which resulted in the death of 10 or more people, 100 affected people, a declaration of a state of emergency, and a call for international assistance (Guha-Sapir et al. 2015). To overcome this constraint, some efforts to establish global or national spatiotemporal ~~natural hazard~~geohazard geo-databases are made, particularly for fatalities, utilizing systematic metadata search techniques that obtained from news articles, media sources, and national archives (Froude and Petley, 2018; Görüm and Fidan, 2021; Haque et al., 2016; Kirschbaum et al., 2015; Kirschbaum et al., 2010; Petley, 2012; Taylor et al., 2015). Beyond its value of comprehending ~~natural hazard~~geohazards within the spatiotemporal inventories, these geo-databases frequently necessitate and rely on a workforce for labor-intensive tasks, such as compiling, gathering, and analyzing the data in the creation of inventories.

Over the last two decades, there has been a noticeable advancement in the integration of artificial intelligence, namely machine learning (ML), deep learning (DL), and natural language processing (NLP), to create automated or semi-automated approaches for detection of ~~natural disaster~~geohazards (Meena et al., 2022), monitoring (Restrepo-Estrada et al., 2018), early warning systems (Kitazawa and Hale, 2021), prediction (Fang et al., 2023), as well as the compilation and database generation pertaining to ~~natural disaster~~geohazards. However, the spatiotemporal data gathered by government and private databases are usually restricted or owned by private enterprises for profit (Lai et al., 2022). The internet sources like online newspapers and social media have been widely used to overcome this limitation by applying ML and NLP tools. For instance, Sodoge et al (2023) proposed an approach for automatization of drought impacts and creation spatiotemporal database based on newspaper articles in Germany between 2001 - 2021 using lasso logistic regression for impact detection

and named entity recognition in location identification. The spatiotemporal distribution of historical floods and storms was extracted using online newspapers on the United States-country scale by employing a hybrid named entity recognition model (Lai et al., 2022). On the other hand, social media, particularly X (formerly known as Twitter), enabled the researcher to assess spatiotemporal patterns and to create a database of ~~natural disaster~~geohazards by using data crawling methods (Franceschini et al., 2024). For instance, Hickey et al. (2024) tracked variations in the social reaction of geo-tagged Twitter posts during the 2018 2018 eruption of 18 Kīlauea, Hawaii, and found the reflective patterns of volcanic activities in the posts using sentiment analysis and ML tools.

Despite the wide application spectrum of NLP, ML, and DL tools in the creation of inventories or databases and assessment of ~~natural disaster~~geohazard research, the studies mainly focus on single web sources and ~~natural disaster~~geohazards such as drought (Madruga De Brito et al., 2020; Sodoge, Kuhlicke, Mahecha, et al., 2023), landslides (Battistini et al., 2013), flood (Liu et al., 2020), typhoon (Kitazawa and Hale, 2021). Therefore, we developed an integrated method that retrieves, classifies, and geolocates multiple ~~natural disaster~~geohazards; landslide, flood, wildfire, and sinkhole formation using web gazette sources in Türkiye between 1997 and 2023. We chose Türkiye as our research focus due to its proneness to ~~natural disaster~~geohazards, leading to annual casualties (Görüm and Fidan, 2021) and socio-economic losses. Even though Türkiye highly suffers from earthquakes, we exclude the earthquake from our approach since geotagging problems due to the epicenter and news reporting distance (Battistini et al., 2013) ~~and also~~and the international and national services provide freely available near-real-time spatial data of earthquake distribution.

The key goal of this research, therefore, is to develop an ~~semi-~~automated approach for building spatiotemporal inventories~~y~~ and maps of ~~natural hazard~~geohazards in Türkiye, such as sinkhole formation, wildfires, floods, and landslides. Our research aims to develop a system capable of parsing newspaper articles regarding geohazards from online sources, automatically classifying the news by hazard type, extracting relevant spatial coordinates and dates of occurrence, and subsequently mapping and storing the gathered geohazard data.

~~Our research focuses on creating a system that can parse newspaper articles about natural hazards from internet sources, classify the news automatically according to the type of hazard, extract pertinent spatial coordinates and times of occurrence, and then map and store the collected natural disaster hazard data~~.

2 Methods and Data

To accomplish our targets, the general concept of the proposed approach includes five integrated main steps: (1) data collection from newspaper websites using a web scraper tool that we developed, (2) NLP preprocessing which cleans data and extracts locations with named entity recognition (NER), (3) modeling; non-negative metric factorization for topic modeling, (4) geolocator which we developed an algorithm using Nominatim, (5) final inventory mapping.

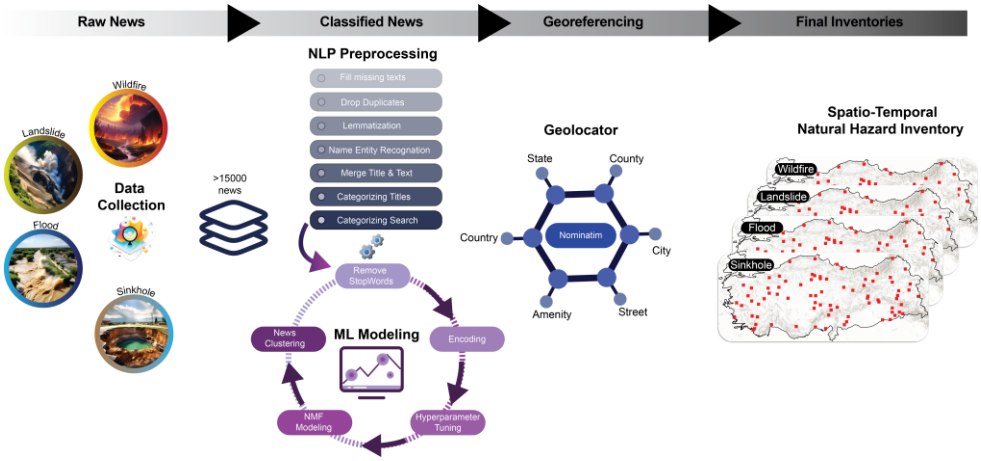

Figure 1: The systematic flowchart of the methodology summarizing the steps employed from the data gathering to the final inventory mapping stages. The illustration of ~~natural hazard~~geohazards by photos in the Raw News section of the figure ~~were~~ was created with the assistance of OpenAI's DALL·E model.

2.1. Data collection

We developed a Python tool "tr-news-scraper" (Demir and Avcioglu, 2024), to retrieve ~~natural disaster~~geohazard-related news from newspaper websites. However, it's a flexible library that enables parsing any news by feeding some keywords as input. This application fetches HTML content from predefined websites of news by using the requests library. We targeted the wide-established national newspaper agencies that have been publishing for at least 10 years; Sabah, Milliyet, Hürriyet, CNN Türk, and Posta.

A list of keywords associated with ~~natural hazard~~geohazards (Table 1) has been input into the tr-news-scraper. It fetches the URLs of newspaper articles which include each term by going through many pages of each news website. The

scraper adds a delay between requests to avoid flooding the servers and getting blocked. The tool records metadata for every news article URL it fetches, containing the article's keyword and publishing date of the news. It also utilizes caching techniques to prevent retrieving the same URL twice.

Table 1: The keywords associated with ~~natural hazard~~geohazards in Turkish are used to fetch newspaper articles in Türkiye.

Categories	Keywords in Turkish	Keywords in English
Wildfire	"orman yangını", "orman yandı", "çalılık yangını", "makilik yandı"	"the forest fire", "forest burned", " bushes burned", "maquis burned"
Flood	"sel", "sel meydana geldi", "taşkın oldu", "nehir taştı", "çamur aktı", "dere taştı"	"flood", "flood occurred", "there was a flood", "the river overflowed", "the mud flowed", "the creek overflowed"
Landslide	"heyelan", "toprak kayması", "kaya düştü", "kaya düşmesi", "toprak aktı"	"landslide", "earth slide", "the rock fell", "rockfall", "the earth flowed"
Sinkhole	"obruk meydana geldi", "obruk oluştu"	"a sinkhole occurred", "a sinkhole formed"

Subsequently, the content of each news article is retrieved by the scraper tool after gathering ~~all of~~all the URLs. It takes advantage of the newspaper library to make article extraction and processing easier. The scraper gathers relevant data for every URL, including the title of the article, the author(s), the date of publication, the primary text content, related keywords, and any images that are accessible. Following extraction, the data is structured into a Pandas DataFrame (Supplementary Table 1) to ensure that it might be examined further. The scraper eliminates duplicate entries based on both textual content and URL to ensure data integrity. In essence, the tr-news-scraper tool retrieves news articles about ~~natural hazard~~geohazards automatically from several websites, giving a large dataset for further analysis and classification.

2.2. NLP Preprocessing

The unrefined retrieved data from web sources, as stated above, typically lacks spatial coordinates and is unstructured. The context of news is transformed multiple times throughout the preprocessing stage before being analyzed. NLP preprocessing requires multiple steps to get text (i.e., news) input ready for additional modeling or analysis. In advance of analysis and modeling, Hickman et al., (2022) portrayed the following common steps for text preprocessing: text identification (i.e., tokenization), content removal (i.e., stop words and nonalphabetic characters removed), agglomeration of semantically related terms to reduce data sparsity and improve predictive power (i.e., lowercase conversion, misspelling correction, contraction/abbreviation expansion, and stem/lemmatized), and capturing more semantic information (i.e., handling negation).

We have followed the steps shown in Figure 1 to complete the cleaning and preprocessing of the unrefined data parsed from web sources to get ready for further analyses. Initially, we filled in the missing contents (the main body of the

news) with the meta description text which summarizes the parsed news related to our target keywords. This step was necessary owing to contents where news articles contain titles but lack corresponding content during parsing procedures. The duplicates that originate from the same media sources yet differ in keywords have been removed from the retrieved dataset. We created a new column by merging the titles and contents of the individual columns we parsed, which better represented each event (such as a flood, landslide, etc.) or unrelated news concerning our target. Because titles provide succinct, event-specific information that highlights the distinctive features such as type, location, and date of ~~natural hazard~~geohazards. Subsequently, we utilized the lowercase conversion which is a standard application in NLP (Hickman et al., 2022) to new merged content which includes titles and the main body of the news. Also, the removal of punctuation of content was a necessary step in achieving our objective of cleaning noisy text by using specific filters (noun, adjective, adverb, and verb), thus enabling the elimination of conjunctions, punctuation marks, articles, etc. After that, in order to improve the coherence and consistency of our text analysis, we lemmatized our content using the TrSpaCy pipeline (Altinok, 2023) to break words down to their most basic forms utilizing linguistic processing approaches. Lemmatization is a step that helps to standardize terminology and makes semantic analysis across the corpus more accurate.

When identifying the stopwords for the ~~natural disaster~~geohazard ~~hazard~~ news articles, it is necessary to compile a detailed list of commonly used terms that are unrelated to the incident. By removing noise and unimportant information from the original stopword compilation, content (i.e., type, date, location of event) directly related to the disaster of interest can be identified more accurately. We may fine-tune our search criteria and improve the accuracy of our data retrieval efforts by deliberately eliminating irrelevant terms from our analysis.

Vectorization is another essential step for text classification in NLP. We utilized the Term Frequency (TF) - Inverse Document Frequency (IDF) technique (TF-IDF) which is a widely used statistical method in NLP and feature extraction. We perform the TfidfVectorizer using the scikit-learn (Pedregosa et al., 2011) in scaling the words. TF is the number of times it appears in a document in relation to the total number of words in that document, Eq 1.:

$$TF = \frac{\text{number of times the term appears in the document}}{\text{total number of terms in the document}} \quad (1)$$

A term's IDF shows how many documents in the corpus contain that term. Words that are specific to a limited subset of papers (for example, technical jargon terms) are given a greater relevance value than words that are used in all publications (a, the, and), Eq 2.:

$$IDF = log\left(\frac{\text{number of documents in the corpus}}{\text{number of documents in the corpus contain the term}}\right) \quad (2)$$

It is important to note that the IDF equation continues to prioritize relevant terms such as “forest fired”, the lack of scaling would result in an excessive focus on infrequent words such as “and, or, the, etc.”. By employing a logarithmic transformation, we mitigate the influence of excessively common words that lack utility for classification, while preserving significant distinctions among pertinent terms. By making this change, the model is guaranteed to concentrate on informative terms instead of giving common ones an excessive amount of weight.

Then TF-IDF can be calculated by multiplying TF and IDF scores:

$$TF\text{-}IDF = TF * IDF \quad (3)$$

The n-grams, a hyperparameter in the TfidfVectorizer, were applied as “ngram_range = (1,2)” for unigrams and diagrams capturing both single words and pairs of words within the specified range to better represent nuanced text data. By doing so, we benefit from more contextual meaning, for example, by maintaining word sequences, enabling models to differentiate phrases such as "sinkhole occurred" from the individual terms "sinkhole" and "occurred," which may possess distinct meanings when analyzed in apart from one another.

Named Entity Recognition (NER) is the process of locating specific words or phrases, so-called “entities”, in a document and categorizing them into groups like people, places, or events. It helps to grasp the context and meaning of the text, which is important for a variety of natural language processing applications, such as sentiment analysis and information extraction. We used the NER component of the TrSpaCy which is the first spaCy model trained for the Turkish language produced by using diverse sources in the Turkish language: Wikipedia articles, crawling of e-commerce, and movie review websites for different genres (Altinok, 2023). This model essentially includes a tokenizer, trainable lemmatizer, POS tagger, dependency parser, morphologizer, and NER pipelines.

2.3 Modeling

Once the extracted unrefined data has been successfully preprocessed, the next step is to feed these refined contents into several models for additional processing and prediction. First, the most relevant keywords related to ~~different natural disasters~~geohazards are identified using the Nonnegative Matrix Factorization (NMF) technique, which was first pioneered by Paatero and Tapper (1994, 1997) as well as Lee and Seung (1999, 2001).

2.3.1 Nonnegative matrix factorization (NMF)

NMF refers to a set of linear algebra and multivariate analysis techniques where a matrix $\boldsymbol{X}$ is divided into two matrices, $\boldsymbol{W}$ and $\boldsymbol{H}$, each of which only has non-negative elements by minimizing the distance $\boldsymbol{d}$ between $\boldsymbol{X}$ and the product of WH. The most widely used distance metric is the squared Frobenius norm, which is a simple matrix application adaption of the Euclidean norm (Lee and Seung, 1999), Eq. 4:

$$d_{\mathrm{Fro}}(X,Y) = \frac{1}{2}\|X - Y\|_{Fro}^{2} = \frac{1}{2}\sum_{i,j}(X_{ij} - Y_{ij})^2 \quad (4)$$

This approach, which is an unsupervised learning technique, reduces the dimensionality of data into spaces of fewer dimensions. We used the NMF model as an additional step for data filtration which is initially categorized into different types of ~~natural disasters~~geohazard in the preprocessing section. This model identifies and finds efficiently the different clusters of ~~natural disasters~~geohazard. Because the original data set parsed through our tr-news-scraper method contains

Formatted: English (United States)
Formatted: English (United States)
Formatted: English (United States)
Field Code Changed
Field Code Changed

irrelevant and noisy information that originated from search engine optimization (SEO) practices of the news websites. This method incorporates news information that can increase news visibility to search engines and their user base.

NMF model suggests clusters that are either ~~natural hazard~~geohazards or not by using the parsed and preprocessed news content as input data. Further, we expand this division to encompass different categories of ~~natural hazard~~geohazards by increasing the number of natural clusters within the NMF model that we seek to autonomously detect. As mentioned above, the primary goal of the NMF model is to identify and differentiate the themes related to different ~~natural hazard~~geohazard~~s~~ and others (i.e., not related to ~~natural hazard~~geohazard~~ss~~; health, politics, sport, etc.) that also need to be evaluated by further validation procedures. Therefore, we utilized the coherence score and expert-based evaluation to validate the results of the NMF model. The coherence score measures the semantic similarity between high-scoring words within each topic, evaluating how interpretable and meaningful the topics are (Rehurek and Sojka, 2010; Röder et al., 2015). To assess the coherence of the news topics produced by the NMF model, we employed the Coherence Model from the Gensim library (Rehurek and Sojka, 2010). We utilized *Cv* as the coherence option in CoherenceModel. *Cv* uses a sliding window approach, grouping the top words into a single set and employing an indirect validation metric that combines normalized pointwise mutual information (NPMI) with cosine similarity (Rehurek and Sojka, 2010; Röder et al., 2015; Syed & Spruit, 2017). Furthermore, true (i.e., incident) and false (i.e., not incident) tests are determined by authors based on textual content that compares with those produced by the NMF model results to perform an expert-based evaluation. The classifications created by authors were regarded as actual data. The accuracy of the NMF-generated categories was then evaluated, and their alignment with the human-defined categories was used to determine whether they were correct, yielding the evaluation score. For this task, we have chosen random 2000 news over 10593 total news within all-~~natural hazard~~geohazard-news, and we have evaluated the final score by reading and determining their "incidents" criteria (i.e., true and false). It is essential to highlight that if the news meets our criteria—that is, provides a clear explanation of the occurrence of specific events with date and geographical attributes—we have classified it as an "incident" in our inventories (Supplementary Table 2). Therefore, we have eliminated the news such as regional or temporal reviews and repeated news.

2.4. Geolocator

In this study, we developed a geospatial data processing and localization method leveraging the locally hosted Nominatim geocoding service. The primary objective of this method is to determine geographic locations from textual locality descriptions accurately. Because the information structured in the RSS or Atom format does not have a native geographical location; the news itself is not associated with any structured geo-location (Battistini et al., 2013). The process is implemented in Python. Our method processes the textual locality descriptions and retrieves geolocation information. It constructs and sends queries to the Nominatim API for each locality entry, parsing the results to identify geographic components such as states/provinces, counties, cities, amenities, and streets. The method follows a systematic approach. It begins by searching for states/provinces within the locality descriptions. Upon finding a match, it then sequentially searches

for counties, cities, amenities, and streets within the remaining locality descriptions. The implemented method efficiently processes large datasets of locality descriptions, accurately identifying and organizing geographic components.

3 Results and Discussion

A total 15 569 number of articles from 1997 to 2023 have been fetched from newspaper websites through our web scraping tool "tr-news-scraper" by using the keywords listed in Table 1. The raw inventory includes 5510 floods, 4262 wildfires, 5255 landslides, and 542 sinkholes. Following the first filtering, which involves eliminating redundant, repeated, and unnecessary news, a total of 10 593 news remains for the subsequent stages of NMF modeling and geolocalization (Table 2). Geohazard news from NMF is grouped into 2236 floods, 655 wildfires, 766 landslides, and 271 sinkholes, resulting in 3928 news remaining which include multiple locations. Nevertheless, following a thorough semantic analysis using TrSpaCy (Altinok, 2023), we identified 13 940 distinct locations (i.e., cities, counties, villages, etc.). As a result, we have determined that these areas have at least 13 940 geohazard incidents.

Table 2: The analytical processes of different classes of news at various stages of the analysis from obtaining unrefined data to geolocalized news as an inventory.

Class	Flood	Wildfire	Landslide	Sinkhole	Total
Unrefined News	5510	4262	5255	542	15 569
Filtered News	4270	2123	3860	331	10 593
NMF Groups	2236	655	766	271	3928
Geolocalized Incidents	9609	1834	1843	654	13 940

Figure 2 shows the WordCloud of different ~~natural hazard~~geohazards and Supplementary Table S2 summarizes the most important 20 words with their frequencies related to the NMF model results. As expected, the total number of words is highest in the flood news and lowest in the sinkhole, it is correlated with the number of news parsed. "Yangın" (fire) and

"orman" (forest) are the two most commonly (3.28% and 2.59%, respectively) used terms about wildfires. Given how frequently "orman" appears in the news, it implies that we have compiled the news from forested areas which matches the wildfire criteria for this study. Because we haven't included news on fire occurrences involving buildings, homes, etc. Furthermore, bigram combinations of the most frequent words reveal that the term "orman yangını," which is translated as "wildfire," is the most often occurring noun group in NMF grouping analyses. This shows that urban fires—such as those that occur in buildings, homes, etc.—were eliminated from the inventories and analyses.

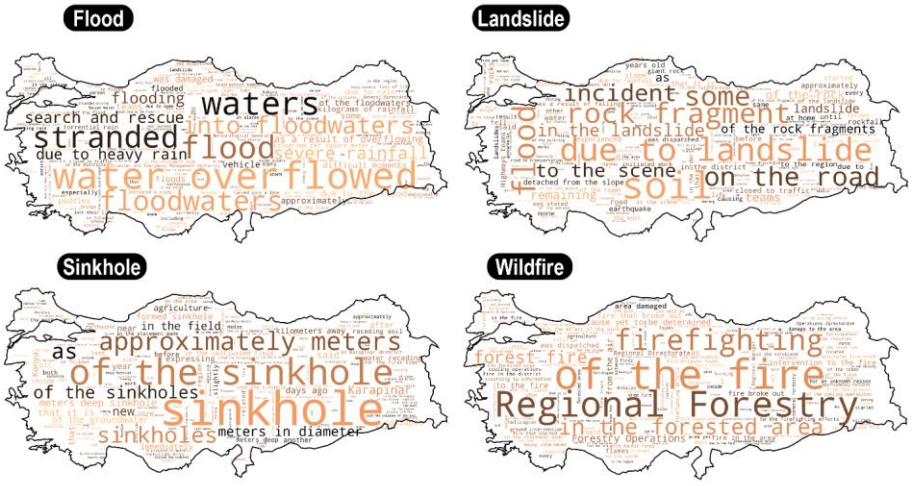

Figure 2: The world clouds illustrate the most frequently seen words in the filtered for different geohazard news. The sizes of each word denote its relative frequency or significance within the dataset; larger words, such as "waters, of the sinkhole, due to landslide, of the fire" signify principal themes, whereas smaller words offer supplementary context pertaining to details of geohazards (for example "meters depth sinkhole") within the news. The color variations serve solely for visual differentiation without indicating any categorical distinctions. Additionally, the spatial location of the words was arbitrarily positioned and does not indicate geographic relation with geohazards.~~The world clouds for different natural hazards emphasize the most frequently seen words in the filtered news~~.

The phrases "sel" (flood), "su" (water), "yağış" (precipitation), and "sağanak" (downpour) are associated with flooding, indicating a strong correlation between heavy rainfall (the term "sağanak") and subsequent flooding incidents. As indicated by the prominence of these phrases, which highlight the importance of water-related disasters and their direct relationship to precipitation patterns. Despite being documented as less frequent occurrences, sinkhole events can be identified by their characteristic phrases within the news. The term "metre" (meter) and the term "obruk" (sinkhole) are

closely associated, it is possible that understanding the significance of sinkholes depends critically on their size or depth. These terms are frequently used, which emphasizes how crucial it is to measure and monitor sinkholes because they can cause major disruption to infrastructure and public safety.

The terms connected to landslides are "kaya" (rock), "toprak" (soil), and "heyelan" (landslide). These keywords also emphasize the geological origin of the component of landslide events by indicating the role of rock and soil. It is clear to identify the type of landslide activity from the most frequent terms from the table that “düştü” (fell), “kopan” (broken off), and “parçala” (pieces of) point particularly rockfall activities. These terms may also point to the selective importance of landslides, which are especially newsworthy because they directly affect people's lives, urban areas, and vital infrastructure. It is generally acknowledged that journalists often ~~agreed~~ agree on exclusive matters of public relevance (Harcup and O’Neill, 2017; Pita Costa et al., 2024). Therefore, the number of landslides in our inventory may primarily reflect landslides that impact humans because most landslides occur in remote mountainous areas or away from human infrastructure, etc., and are less essential than others, causing some socio-economic losses. It is important to emphasize that certain phrases may give rise to confusion when it comes to grouping analysis. For instance, the term "yağış" (precipitation) is significant for both landslides and floods. Since most news reports highlight landslide incidents by mentioning predisposing elements like precipitation—which is also important for flood events, this could result in mis-clustering, primarily in landslide events.

Hu (2018) highlighted the significance of ambiguous connections between texts and locations, meaning that geo-text data can contain both from and about locations (MacEachren et al., 2011). This problem implies that many location names have turned prominence in the news, which can lead to georeferencing complications. With this problem in mind, after acquiring the 3928 total clustered news (i.e., NMF Groups in Table 2), we used TrSpaCy (Altinok, 2023)—a geoparsing technique that extracts explicit place names from implicit geo-text data (Gregory et al., 2015; Hu, 2018)—to acquire place names. To further address the point raised by Hu (2018), the spatial filters also removed the names of other cities from the news, leaving only one.

It should be taken into account that the number of geohazards represents the minimum number of incidents that occurred during the period in our analyses. The geohazards events we fetched from the online gazettes are newsworthy with several aspects that cause economic losses in many ways: damage to critical infrastructures, urban areas, and agricultural activities. Given that floods and wildfires have numerous effects on human life and are more frequently observed, there may be less variation between actual occurrences and the incidents we record compared to landslides and sinkholes even if they cause severe loss. The reason behind this is that the internet gazettes or online sources do not consider landslides that occur remotely from metropolitan areas or vital infrastructure to be noteworthy.

3.1 Uncertainty assessment and limitations

Formatted: Indent: First line: 0 cm

The Non-negative Matrix Factorization (NMF) model was applied to the news dataset using a range of topic numbers from 2 to 20 components. The coherence score was calculated for each model configuration to evaluate the coherence and interpretability of the generated topics. Supplementary Figure 1 shows that the coherence score generally

increased with the number of topics, reaching its plateau at 4 components with a coherence score of 0.80. This suggests that the 4-topic model provides the most meaningful and coherent topics for the given news dataset. Beyond the 4 components, the coherence scores vary or decrease slightly, indicating that additional topics do not contribute significantly to the model's overall interpretability. The selected 4-topic model thus strikes an optimal balance between topic coherence and topic interpretability, providing a robust representation of the underlying thematic structure of the news dataset. Also, our expert-based evaluation score, which we performed by 2000 randomly selected news over 10 593, showed overall good consistency with a coherence score of 0.81 evaluation score. On the other hand, when it comes to evaluating each type of ~~natural hazard~~geohazard, flood, wildfire, landslide, and sinkhole, they resulted in different scores, 0.84, 0.7, 0.85, and 0.74, respectively. This evaluation is mainly based on the incident identification criteria. For instance, we have determined four major categories, leading to misclassification in our incident identification. These are the news, that is categorized as "not incident" as shown by the 0 values in Supplementary Table S3. The first category is the "common words" with incidents like "a person lost her/his balance and fell while rock climbing" (in Turkish: "*kaya tırmanışı yaptığı sırada dengesini yitirerek düşen*"). The second category is the "review news" which essentially compiles multiple incidents over the course of time. The third category is "warning alerts" such as "AFAD issues forest fire warning for 6 provinces…" (in Turkish: "*AFAD'dan 6 il için orman yangını uyarısı…*"). The last category is "misinterpretation" as given in the example: "*… They encountered a sinkhole that was 7 meters wide and 10 meters deep due to a meteorite fell.*" (in Turkish: *"... Meteorite düşmesinden kaynaklı 7 metre genişliğinde ve 10 metre derinliğindeki obruk ile karşılaştılar...*").

On the one hand, to enhance the reliability of our study, we incorporated a ground truth evaluation step, a manual verification method utilized in related research (Madruga de Brito et al., 2025; Stein et al., 2024). We employed random sampling for this evaluation, selecting 500 geohazard incidents to assess mapping performance. This sampling yielded 284, 97, 76, and 43 incidents of flood, landslide, wildfire, and sinkhole, respectively. We manually verified these incidents by cross-checking the locations of mapped geohazards with contextual news information from which we extracted location data. The uncertainty assessment for mapping performance overall resulted in good performance which is 82.4 % of geohazards accurately were mapped. On the other hand, an important limitation of our approach is the enhancement of spatial accuracy to achieve a level pertinent to streets, roads, and topographical features. This challenge arises from the heterogeneous writing style prevalent in journalism, which hampers our access to consistently uniform "administrative level" information (city, county, and village) and the associated specifics, such as streets, neighborhoods, and roads. Consequently, our primary objective for this study was to map geohazards within these administrative levels by geolocating incidents to the centers of these locations, as represented in the Open Street Map. We adhere to this methodology since the subsequent refinement of these inventories necessitates precise delineation of the targeted geohazards through the integration of geomorphological interpretation, particularly for landslides, and inundation area extraction for flood inventories utilizing high-resolution satellite imagery.

Formatted: English (United States)

~~The temporal distribution of the hazards shows an increase after 2005 (Fig. 3). This result implies that internet sources became more widely available after 2005, which is in line with the increase reported by Gorum and Fidan (2021) for fatal landslide cases in Türkiye. Throughout the study period (1997–2023; Fig. 3), sinkholes and wildfires fluctuated, but the frequency of floods has shown an increase, particularly after 2016, when the annual number of events increased approximately from 400 to 1600 by 2023. Similarly, even though the landslide numbers showed an increase after 2016, the frequency of landslides (Fig. 3) has remarkably decreased (~60 % and ~36 % last two years, respectively). The European Forest Fire Information System (EFFIS) (San-Miguel-Ayanz et al., 2012) database, which records the largest burnt area in Europe at 27848.33 hectares (https://forest-fire.emergency.copernicus.eu/reports-and-publications/annual-fire-reports, last accessed August 2024), suggests that although wildfire occurrences exhibited sporadic pattern, their peak occurrence in 2008 was consistent with our database (Fig. 3).~~

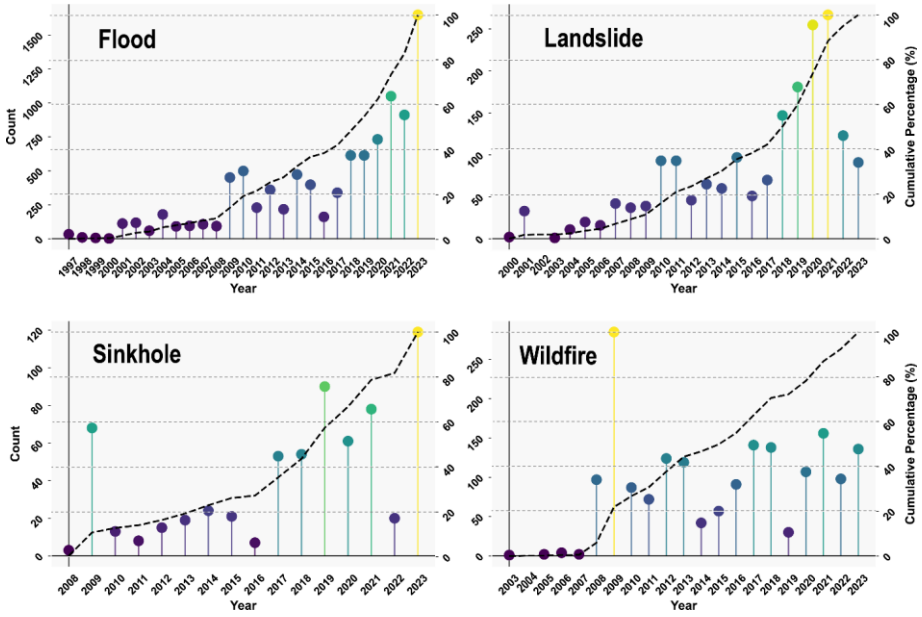

Figure 3: The temporal distribution of the flood, landslide, sinkhole, and wildfire incidents that we mapped. The beginning year varies for different ~~natural hazard~~geohazards and the dashed line indicates cumulative incidents over the study period as a secondary axis. *Count refers to the number of incidents of ~~natural hazard~~geohazards.

~~Hu (2018) highlighted the significance of ambiguous connections between texts and locations, meaning that geo-text data can contain both from and about locations (MacEachren et al., 2011). This problem implies that many location names have turned prominence in the news, which can lead to georeferencing complications. With this problem in mind, after acquiring the 3928 total clustered news (i.e., NMF Groups in Table 2), we used TrSpaCy (Altinok, 2023) a geoparsing technique that extracts explicit place names from implicit geo-text data (Gregory et al., 2015; Hu, 2018) to acquire place names. To further address the point raised by Hu (2018), the spatial filters also removed the names of other cities from the news, leaving only one.~~

~~It should be taken into account that the number of natural hazard~~geohazards ~~represents the minimum number of incidents that occurred during the period in our analyses. The geohazards events we fetched from the online gazettes are newsworthy with several aspects that cause economic losses in many ways; damage to critical infrastructures, urban areas, and agricultural activities. Given that floods and wildfires have numerous effects on human life and are more frequently observed, there may be less variation between actual occurrences and the incidents we record compared to landslides and sinkholes even if they cause severe loss. The reason behind this is that the internet gazettes or online sources do not consider landslides that occur remotely from metropolitan areas or vital infrastructure to be noteworthy.~~

3.~~1~~2 Spatiotemporal implications ~~variations on~~ on ~~natural hazard~~geohazard inventories

Here, we demonstrate the spatiotemporal distribution of ~~natural hazard~~geohazard inventories that have been ~~examined~~ gathered using ~~thorough semantic analysis~~web scraping ~~and,~~ Natural Language Processing (NLP), clustering, and mapping techniques at the national level in Türkiye and we compare spatiotemporal consistencies of our inventories with existing literature.

The temporal distribution of the hazards shows an increase after 2005 (Fig. 3). This result implies that internet sources became more widely available after 2005, which is in line with the increase reported by Gorum and Fidan (2021) for fatal landslide cases in Türkiye. Throughout the study period (1997–2023; Fig. 3), sinkholes and wildfires fluctuated, but the frequency of floods has shown an increase, particularly after 2016, when the annual number of events increased approximately from 400 to 1600 by 2023. Similarly, even though the landslide numbers showed an increase after 2016, the frequency of landslides (Fig. 3) has remarkably decreased (-60 % and -36 % last two years, respectively). The European Forest Fire Information System (EFFIS) (San-Miguel-Ayanz et al., 2012) database, which records the largest burnt area in Europe at 27848.33 hectares (https://forest-fire.emergency.copernicus.eu/reports-and-publications/annual-fire-reports, last accessed August 2024), suggests that although wildfire occurrences exhibited sporadic pattern, their peak occurrence in 2008 was consistent with our database (Fig. 3).

Formatted: Indent: First line: 1,27 cm

Formatted: Indent: First line: 1,27 cm

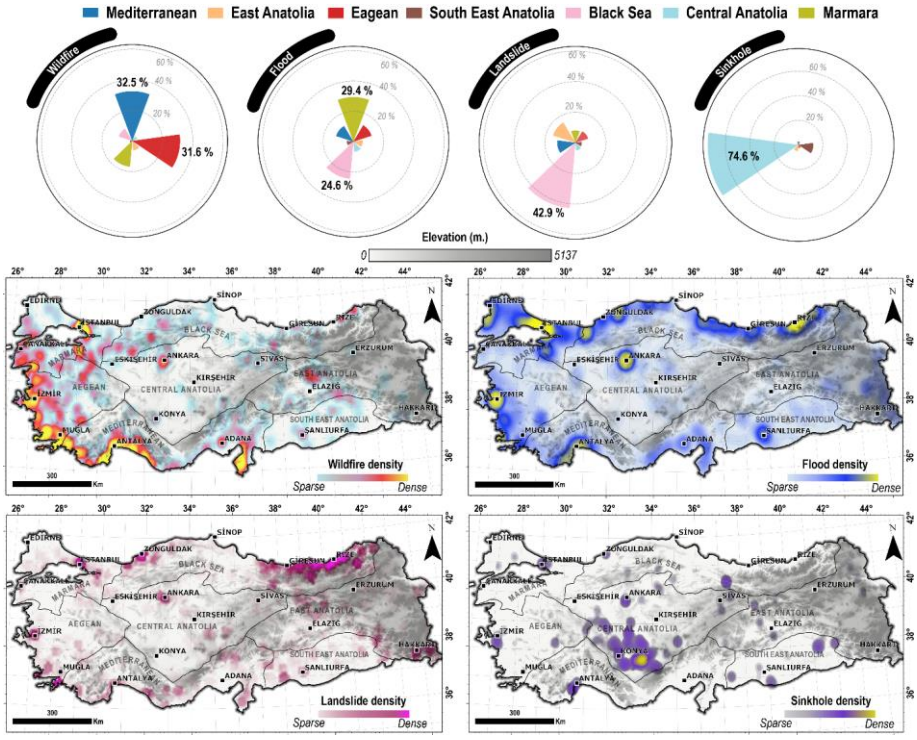

Figure 4: The spatial distribution of ~~natural hazard~~geohazards over different administrative regions of Türkiye. The circular bar plots depict the percentages of events across various regions, accompanied by density maps illustrating spatial hotspots of ~~natural hazard~~geohazards across different regions in Türkiye.

Given the primary concentration of ~~natural hazard~~geohazard occurrences (Fig. 4), Türkiye displays a particular spatiotemporal tendency to geohazards (Fig. 5). The spatial distribution of ~~natural hazard~~geohazards reveals that flood events are relatively well distributed over the country to others. But the majority (54 %) of the events happened in the Marmara (29.4 %) and Black Sea (24.6 %) regions, less common in the Central Anatolia Plateau (Fig. 4), the driest region in Türkiye, especially around the Konya province (average ~400 mm/y precipitation). Some hotspots for flood events appear near big cities like Istanbul, Ankara (the capital city), and İzmir (Fig. 4). This may suggest that large cities can readily get the attention of gazettes, even for minor incidents. Ankara, for instance, may appear as a hotspot, yet the Central Anatolia Region, which includes it, has had 6.7 % of all floods. On the other hand, Istanbul is not only the biggest city in the Türkiye which receives major attention from journalists, but also it is, particularly in north-faced basins, geographically part of the

Formatted: Indent: First line: 1,27 cm

Black Sea region resulting in higher rainfall which potentially impacts the number of floods. Therefore, attention should be given to these inventories for the real number and accuracy of the events. It is important to note that our inventory primarily captures the urban floods that mainly occur where there is a construction area in flood-prone areas (Brown et al. 2007; Mason et al. 2007) and poorly engineered flood control infrastructure (Gallegos et al. 2009; Ozdemir et al. 2013). Our findings on the spatial distribution of flood incidents align with the results of a recent study on flood inventories – also considered news as source data - (Akbaş et al., 2025, under review), which further supports the accuracy of our inventory.

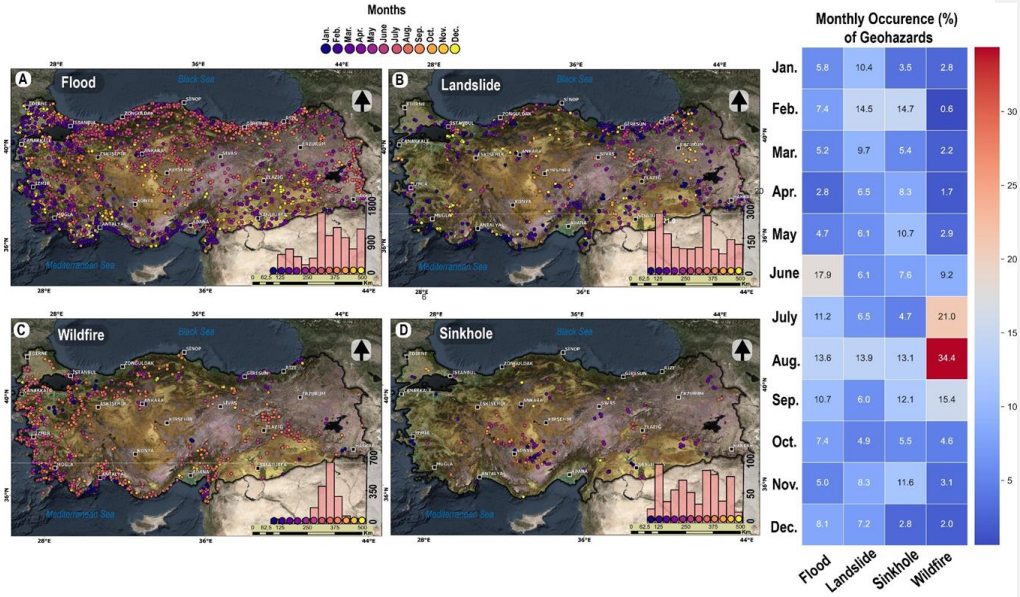

Figure 5: The spatiotemporal distribution (monthly) maps of ~~natural hazard~~geohazards, which are aligned with bar plots showing the total number of months over the study period from January to December in the color gradient from blue to yellow, respectively. The heat map chart demonstrates the monthly occurrence (in %) for each ~~natural hazard~~geohazard group.

The temporal distribution of the flood events also shows different patterns by region (Fig 5.). The summer season accounts for 42.7 % of all floods in Türkiye, and this temporal pattern shifts to the winter season in cities along the Mediterranean coastline (Fig. 5). Additionally, the spatial distribution of flood events is consistent with the findings of earlier research (Haltas et al., 2021). For example, Koç et al. (2020) showed that fatalities and economic losses due to flood events in the Marmara and Black Sea regions where we identified the majority of the events (54 %).

The spatial distribution of landslides in our inventory shows that the Black Sea Region is the most susceptible area to landslides by having the majority (42.9 %) of all events over Türkiye. The current literature indicates a comparable spatial distribution, with fatal landslides (Görüm and Fidan, 2021) and Turkish landslide inventory (Duman et al., 2011) predominantly concentrated in the Black Sea Region, especially in the eastern section. As indicated in the first section regarding the flood and landslide events being wrongly classified due to the common keywords (e.g, "yağış"), Figure 4 also portrays a supportive spatial relationship in the East Black Sea region where the higher slope and rainfall and the higher flood and landslide events occur. This misclassification mainly affects landslide incidents since most of the news highlights meteorological conditions in the news such as "*landslide occurred as a result of prolonged precipitation*." (in Turkish; "*uzun süren yağış sonucu heyelan meydana geldi*."). On the other hand, whilst this study does not concentrate on triggering or predisposing factors of the ~~natural hazard~~geohazards, it is important to note that the landslides where we obtained fewer occurrences; Marmara, Central Anatolia, and Southeast Anatolia regions are more likely to be associated with anthropogenic origin activities such as mining, road cutting, and other related activities (Fidan and Görüm, 2020; Görüm and Fidan, 2021).

There has been a spatial accumulation of wildfire incidents in our inventory mainly along the Mediterranean coast from Adana to Çanakkale, (Fig. 4 and 5). Conversely, the vicinity of Istanbul exhibits a concentrated region for wildfire occurrences (Fig. 4), indicating increased frequency; these wildfires are notably characterized by a smaller magnitude (i.e., areal coverage) compared to the Mediterranean coast. It is noteworthy to highlight that the Mediterranean and Aegean regions show a distinct tendency with their topography (Avcioglu et al., 2024), climatology (Tatli and Türkeş, 2014; Akbas, 2023), and predominant vegetation cover; *Pinus brutia* Ten. (kızılçam) (known as also Turkish pine) and *P. halepensis* Mill. (halep çamı) (Ekberzade et al., 2022), accounting for the majority of wildfire cases (~ 65%). Türkiye exhibits an evident seasonal pattern throughout the summer months, especially along the coastline that stretches from Marmara to the Mediterranean (Fig. 5). A recent study by Öztürk et al. (2024) identified significant wildfire areas attributed to lightning, demonstrating spatial consistency with our inventory mapping, namely hotspots of wildfires over the Mediterranean and Aegean regions of Türkiye. On the other hand, an interesting finding points out that wildfire incidents occur during the winter and spring in the Eastern Black Sea Region (one of the less frequent regions). This might suggest the influence of the natural phenomenon "foehn winds" a type of dry, relatively warm downslope wind that occurs in the lee (downwind side) of a mountain range which elevates wildfire risk as a potential driver and predisposing factor. This phenomenon also has been shown in the studies (Yetmen and Aytaç, 2017; Arslan et al., 2024) highlighting the importance of foehn winds in the wildfire case in the Eastern Black Sea Region.

The sinkhole formation is the least common ~~natural hazard~~geohazard in Türkiye compared to others in our inventory. In fact, literature has demonstrated that sinkholes are among the most significant ~~natural hazard~~geohazards (Waltham and Fookes, 2003; Parise et al., 2008) because of their rapid and unexpected occurrences, which restrict the certainty of their spatiotemporal forecast (Newton, 1987). Consistent with the literature (Doğan and Yılmaz, 2011; Gökkaya et al., 2021; Orhan et al., 2023)~~)~~, Figure 4 clearly shows that sinkholes predominantly occur in the Central Anatolia Region with no particular temporal tendency (Fig. 5), specifically, the Obruk Plateau (subregion) that surrounds the Konya province.

Formatted: Indent: First line: 1,27 cm

In addition to this region, sinkholes have also been reported in the news as incidents from other regions. Although geological settings have made it less likely for sinkholes to occur in other regions, we maintained all incidents because our approach primarily depends on context truth (i.e. article within news) rather than geologic or geomorphological accuracy. For example, the piping phenomena are predominantly characterized by journalists as sinkhole formation due to their insufficient scientific background in geoscience. This may imply that, despite the news's significant potential and benefits for comprehending ~~natural hazard~~geohazards, it is also necessary to carefully consider the news' scientific underpinnings.

Conclusion

It is essential to have a comprehensive and long-term understanding of how, when, and where ~~natural hazard~~geohazard~~s~~ have affected societies in recent years to inform policymakers about how to overcome and mitigate multiple hazards. Therefore, with this study, we developed an automated approach to build spatiotemporal inventories from online gazette news for multiple hazards; flood, landslide, wildfire, and sinkhole formation by combining the web-scraping, semantic analysis, clustering, and geolocating algorithms on a national scale of Türkiye. The news parsing tool "tr-news-scraper" has been developed and 15_569 news articles have been fetched with this tool from the selected online gazettes in Türkiye by employing the keywords associated with ~~natural hazard~~geohazards. After NLP processing; a total of 13_940 incidents of ~~natural hazard~~geohazard~~ss~~ have been recorded and geolocated. Consequently, we mapped 9609 floods, 1834 wildfires, 1843 landslides, and 654 sinkhole formation incidents, that occurred during the period between 1997 to 2023 in Türkiye. Our inventories show spatiotemporally distinct patterns in flood, landslide, wildfire, and sinkhole events, consistent with previous studies. The ground truth validation revealed 82.5 % ~~-~~incident mapping accuracy from the 500 randomly selected incidents. Although the clustering and incident identification findings show 0.80 confidence scores, the contextual similarities, the contextual similarities (e.g., "yağış" term for flood and landslide), review news, and misinterpretations of the news may give rise to confusion either in fetching or clustering appropriate categories of ~~natural hazard~~geohazards. Consequently, our study demonstrated the necessity for standardizing context writing related to geohazard news to more effectively capture information regarding location and incident specifics.

~~Overall, the~~The approach provided in this study expands to existing inventories by investigating the potential and limitations of using web scraping, NLP, and machine learning methods, as well as providing an open alternative to creating inventories where others are inaccessible owing to national restrictions. Furthermore, we can more accurately portray ~~natural hazard~~geohazard events with these inventories because local news is less prevalent than global news but covers more events. Hence, further research is required to expand the spatial scale of similar approaches to other regions and multiple languages using advanced large language models.

Formatted: Indent: First line: 1,27 cm

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Author contributions

AA and OD designed the study together with the contributions from TG. OD and AA performed data analyses, data visualization, and interpretations. AA prepared the manuscript with contributions from all co-authors.

Financial Support

This study is supported by the 2247-A National Fellowship for Outstanding Researchers Program of the Scientific and Technological Research Council of Turkey (TUBITAK) [grant number 1199B472343092].

Competing interests

The contact author has declared that none of the authors has any competing interests.

References

Akbas, A. (2023). Seasonality, persistency, regionalization, and control mechanism of extreme rainfall over complex -terrain. *Theoretical and Applied Climatology*, *152*(3), 981-997.

Akbas, A., Gorum, T., Ozdemir, H., (2025) . FlooDOT (FlooD inventory Of Türkiye): A comprehensive flood inventory and its spatio-temporal analyses, Journal of Flood Risk Management (Under Review)

Altinok, Duygu. 2023. "A Diverse Set of Freely Available Linguistic Resources for Turkish." Pp. 13739–50 in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1. Stroudsburg, PA, USA: Association for Computational Linguistics.

Arslan, H., Baltaci, H., Demir, G., & Ozcan, H. K. (2024). Spatiotemporal changes and background atmospheric factors associated with forest fires in Turkiye. Environmental Monitoring and Assessment, 196(10). https://doi.org/10.1007/s10661-024-13027-w

Avcıoğlu, A., Akbaş, A., Görüm, T., & Yetemen, Ö. (2024). The compound effect of topography, weather, and fuel type on the spread and severity of the largest wildfire in NW of Turkey. *Natural Hazards*, 1-19.

Battistini, Alessandro, Samuele Segoni, Goffredo Manzo, Filippo Catani, and Nicola Casagli. 2013. "Web Data Mining for Automatic Inventory of Geohazards at National Scale." *Applied Geography* 43:147–58. doi: 10.1016/j.apgeog.2013.06.012.

Bhuyan, Kushanav, Kamal Rana, Joaquin V. Ferrer, Fabrice Cotton, Ugur Ozturk, Filippo Catani, and Nishant Malik. 2024. "Landslide Topology Uncovers Failure Movements." *Nature Communications* 15(1):2633. doi: 10.1038/s41467-024-46741-7.

CRED. (2023). EM-DAT [Dataset]. CRED/UCLouvain, Brussels, Belgium. Retrieved form www.emdat.be

Delaney, Keith B., and Stephen G. Evans. 2015. "The 2000 Yigong Landslide (Tibetan Plateau), Rockslide-Dammed Lake and Outburst Flood: Review, Remote Sensing Analysis, and Process Modelling." *Geomorphology* 246:377–93. doi: 10.1016/j.geomorph.2015.06.020.

Demir, O., & Avcıoğlu, A. (2024). tr-news-scraper: Turkish news articles scraper based on specified keywords (Version 0.1.0) [Software]. https://github.com/demirogun/tr-news-scraper.

Doğan, Uğur, and Mutlu Yılmaz. 2011. "Natural and Induced Sinkholes of the Obruk Plateau and Karapınar-Hotamış Plain, Turkey." *Journal of Asian Earth Sciences* 40(2):496–508. doi: 10.1016/j.jseaes.2010.09.014.

D. Lee and H. Seung, "Learning the Parts of Objects by Non-and its variants are based on the SED objective Negative Matrix Factorization," Nature, vol. 401, no. 6755, pp. 788- function, which basically sets the keynote of the 791, 1999.

D. Lee and H. Seung, "Algorithms for Non-Negative Matrix whole NMF framework. It claims attention that the Factorization," Proc. Advances in Neural Information Processing notion of optimizing the objective functions is not Systems, pp. 556-562. 2001.

Ekberzade, Bikem, Omer Yetemen, Omer Lutfi Sen, and H. Nuzhet Dalfes. 2022. "Simulating the Potential Forest Ranges in an Old Land: The Case for Turkey's Forests." *Biodiversity and Conservation* 31(13–14):3217–36. doi: 10.1007/s10531-022-02485-8.

Fan, Xuanmei, Gianvito Scaringi, Oliver Korup, A. Joshua West, Cees J. van Westen, Hakan Tanyas, Niels Hovius, Tristram C. Hales, Randall W. Jibson, Kate E. Allstadt, Limin Zhang, Stephen G. Evans, Chong Xu, Gen Li, Xiangjun Pei, Qiang Xu, and Runqiu Huang. 2019. "Earthquake-Induced Chains of Geologic Hazards: Patterns, Mechanisms, and Impacts." *Reviews of Geophysics* 57(2):421–503. doi: 10.1029/2018RG000626.

Fan, Xuanmei, Cees J. van Westen, Qiang Xu, Tolga Gorum, and Fuchu Dai. 2012. "Analysis of Landslide Dams Induced by the 2008 Wenchuan Earthquake." *Journal of Asian Earth Sciences* 57:25–37. doi: 10.1016/j.jseaes.2012.06.002.

Fang, Zhice, Hakan Tanyas, Tolga Gorum, Ashok Dahal, Yi Wang, and Luigi Lombardo. 2023. "Speech-Recognition in Landslide Predictive Modelling: A Case for a next Generation Early Warning System." *Environmental Modelling & Software* 170(September):105833. doi: 10.1016/j.envsoft.2023.105833.

Fidan, Seçkin, and Tolga Görüm. 2020. "Türkiye'de Ölümcül Heyelanların Dağılım Karakteristikleri ve Ulusal Ölçekte Öncelikli Alanların Belirlenmesi." *Türk Coğrafya Dergisi* (December). doi: 10.17211/tcd.731596.

Franceschini, Rachele, Ascanio Rosi, Filippo Catani, and Nicola Casagli. 2024. "Detecting Information from Twitter on Landslide Hazards in Italy Using Deep Learning Models." *Geoenvironmental Disasters* 11(1):22. doi: 10.1186/s40677-024-00279-4.

Froude, Melanie J., and David N. Petley. 2018. "Global Fatal Landslide Occurrence from 2004 to 2016." *Natural Hazards and Earth System Sciences* 18(8):2161–81. doi: 10.5194/nhess-18-2161-2018.

Gökkaya, E., Gutiérrez, F., Ferk, M., & Görüm, T. (2021). Sinkhole development in the Sivas gypsum karst, Turkey. Geomorphology, 386, 107746. https://doi.org/10.1016/j.geomorph.2021.107746

Görüm, Tolga, Xuanmei Fan, Cees J. van Westen, Run Qiu Huang, Qiang Xu, Chuan Tang, and Gonghui Wang. 2011. "Distribution Pattern of Earthquake-Induced Landslides Triggered by the 12 May 2008 Wenchuan Earthquake." Geomorphology 133(3–4):152–67. doi: 10.1016/j.geomorph.2010.12.030.

Görüm, Tolga, and Seçkin Fidan. 2021. "Spatiotemporal Variations of Fatal Landslides in Turkey." *Landslides* 18(5):1691–1705. doi: 10.1007/s10346-020-01580-7.

Guha-Sapir D, Below R, Hoyois Ph (2015) EM-DAT: International Disaster Database. Université Catholique de Louvain,

Brussels, Belgium. www.emdat.be.

Gómez, Derly, Edwin F. García, and Edier Aristizábal. 2023. *Spatial and Temporal Landslide Distributions Using Global and Open Landslide Databases*. Vol. 117. Springer Netherlands.

Haltas, Ismail, Enes Yildirim, Fatih Oztas, and Ibrahim Demir. 2021. "A Comprehensive Flood Event Specification and Inventory: 1930–2020 Turkey Case Study." *International Journal of Disaster Risk Reduction* 56(February):102086. doi: 10.1016/j.ijdrr.2021.102086.

Harcup, Tony, and Deirdre O'Neill. 2017. "What Is News?" *Journalism Studies* 18(12):1470–88. doi: 10.1080/1461670X.2016.1150193.

Hickey, James, James Young, Michelle Spruce, Ravi Pandit, Hywel Williams, Rudy Arthur, Wendy Stovall, and Matthew Head. 2024. "Social Sensing a Volcanic Eruption: Application to Kīlauea 2018." *Natural Hazards and Earth System Sciences*.

Hickman, Louis, Stuti Thapa, Louis Tay, Mengyang Cao, and Padmini Srinivasan. 2022. "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations." *Organizational Research Methods* 25(1):114–46. doi: 10.1177/1094428120971683.

Hu, Yingjie. 2018. "Geo-text Data and Data-driven Geospatial Semantics." *Geography Compass* 12(11):1–19. doi: 10.1111/gec3.12404.

Jones, Rebecca Louise, Debarati Guha-Sapir, and Sandy Tubeuf. 2022. "Human and Economic Impacts of Natural Disasters: Can We Trust the Global Data?" *Scientific Data* 9(1):1–7. doi: 10.1038/s41597-022-01667-x.

Kirschbaum, Dalia Bach, Robert Adler, Yang Hong, Stephanie Hill, and Arthur Lerner-Lam. 2010. "A Global Landslide Catalog for Hazard Applications: Method, Results, and Limitations." *Natural Hazards* 52(3):561–75. doi: 10.1007/s11069-009-9401-4.

Kitazawa, Katsushige, and Scott A. Hale. 2021. "Social Media and Early Warning Systems for Natural Disasters: A Case Study of Typhoon Etau in Japan." *International Journal of Disaster Risk Reduction* 52(September 2019):101926. doi: 10.1016/j.ijdrr.2020.101926.

Koç, Gamze, Theresia Petrow, and Annegret Thieken. 2020. "Analysis of the Most Severe Flood Events in Turkey (1960–2014): Which Triggering Mechanisms and Aggravating Pathways Can Be Identified?" *Water* 12(6):1562. doi: 10.3390/w12061562.

Koç, Gamze, Theresia Petrow, and Annegret H. Thieken. 2020. "Analysis of the Most Severe Flood Events in Turkey (1960-2014): Which Triggering Mechanisms and Aggravating Pathways Can Be Identified?" *Water (Switzerland)* 12(6). doi: 10.3390/W12061562.

Lai, Kelvin, Jeremy R. Porter, Mike Amodeo, David Miller, Michael Marston, and Saman Armal. 2022. "A Natural Language Processing Approach to Understanding Context in the Extraction and GeoCoding of Historical Floods, Storms, and Adaptation Measures." *Information Processing & Management* 59(1):102735. doi: 10.1016/j.ipm.2021.102735.

Loche, Marco, Massimiliano Alvioli, Ivan Marchesini, Haakon Bakka, and Luigi Lombardo. 2022. "Landslide Susceptibility Maps of Italy: Lesson Learnt from Dealing with Multiple Landslide Types and the Uneven Spatial Distribution of the National Inventory." *Earth-Science Reviews* 232(June):104125. doi: 10.1016/j.earscirev.2022.104125.

MacEachren, A. M., Jaiswal, A., Robinson, A. C., Pezanowski, S., Savelyev, A., Mitra, P., ... & Blanford, J. (2011, October). Senseplace2: Geotwitter analytics support for situational awareness. In *2011 IEEE conference on visual analytics science and technology (VAST)* (pp. 181-190). IEEE.

Madruga De Brito, Mariana, Christian Kuhlicke, and Andreas Marx. 2020. "Near-Real-Time Drought Impact Assessment: A

Text Mining Approach on the 2018/19 Drought in Germany." *Environmental Research Letters* 15(10). doi: 10.1088/1748-9326/aba4ca.

Madruga de Brito, M., Sodoge, J., Kreibich, H., & Kuhlicke, C. (2025). Comprehensive Assessment of Flood Socioeconomic Impacts Through Text-Mining. Water Resources Research, 61(1). https://doi.org/10.1029/2024WR037813

Meena, Sansar Raj, Lucas Pedrosa Soares, Carlos H. Grohmann, Cees van Westen, Kushanav Bhuyan, Ramesh P. Singh, Mario Floris, and Filippo Catani. 2022. "Landslide Detection in the Himalayas Using Machine Learning Algorithms and U-Net." *Landslides* 19(5):1209–29. doi: 10.1007/s10346-022-01861-3.

Newton, J.G., 1987. Development of Sinkholes Resulting from Man's Activities in the Eastern United States. US Geological Survey Circular 968.

Orhan, O., Haghshenas Haghighi, M., Demir, V., Gökkaya, E., Gutiérrez, F., & Al-Halbouni, D. (2023). Spatial and Temporal Patterns of Land Subsidence and Sinkhole Occurrence in the Konya Endorheic Basin, Turkey. Geosciences, 14(1), 5. https://doi.org/10.3390/geosciences14010005.

OpenAI. (2025). DALL·E: AI Image Generation Model. Retrieved from https://openai.com/dall-e.

Ozdemir, H., C. C. Sampson, G. A. M. de Almeida, and P. D. Bates. 2013. "Evaluating Scale and Roughness Effects in Urban Flood Modelling Using Terrestrial LIDAR Data." *Hydrology and Earth System Sciences* 17(10):4015–30. doi: 10.5194/hess-17-4015-2013.

Öztürk, Mehmet Göktuğ, İsmail Bekar, and Çağatay Tavşanoğlu. 2024. "Rethinking Lightning-Induced Fires: Spatial Variability and Implications for Management Policies." *Forest Ecology and Management* 572(September):122262. doi: 10.1016/j.foreco.2024.122262.

P. Paatero and U. Tapper, "Positive Matrix Factorization: A nonnegative Factor Model with Optimal Utilization of Error Estimates of Data Values," Environmetrics, vol. 5, no. 2, pp. 111- combining the objective function

P. Paatero, "Least Squares Formulation of Robust non-Negative Factor Analysis," Chemometrics and Intelligent Laboratory Systems, vol. 37, no. 1, pp. 23-35, 1997.

Parise, Mario, Jo De Waele, and Francisco Gutierrez. 2008. "Engineering and Environmental Problems in Karst — An Introduction." *Engineering Geology* 99(3–4):91–94. doi: 10.1016/j.enggeo.2007.11.009.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Peng, M., and Limin Zhang. 2012. "Analysis of Human Risks Due to Dam Break Floods-Part 2: Application to Tangjiashan Landslide Dam Failure." *Natural Hazards* 64(2):1899–1923. doi: 10.1007/s11069-012-0336-9.

Petley, David. 2012. "Global Patterns of Loss of Life from Landslides." *Geology* 40(10):927–30. doi: 10.1130/G33217.1.

Pita Costa, Joao, Luis Rei, Nejc Bezak, Matjaž Mikoš, M. Besher Massri, Inna Novalija, and Gregor Leban. 2024. "Towards Improved Knowledge about Water-Related Extremes Based on News Media Information Captured Using Artificial Intelligence." *International Journal of Disaster Risk Reduction* 100(April 2023):104172. doi: 10.1016/j.ijdrr.2023.104172.

Poumay, J., & Ittoo, A. (2023, September). Evaluating Unsupervised Hierarchical Topic Models Using a Labeled Dataset. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 846-853).

Rana, Kamal, Ugur Ozturk, and Nishant Malik. 2021. "Landslide Geometry Reveals Its Trigger." *Geophysical Research Letters* 48(4):1–8. doi: 10.1029/2020GL090848.

Rehurek, R., & Sojka, P. Software Framework for Topic Modelling with Large Corpora.(2010). In *Proceedings of the LREC*

Formatted: English (United States)

2010 Workshop on New Challenges for NLP Frameworks. University of Malta.

Restrepo-Estrada, Camilo, Sidgley Camargo de Andrade, Narumi Abe, Maria Clara Fava, Eduardo Mario Mendiondo, and João Porto de Albuquerque. 2018. "Geo-Social Media as a Proxy for Hydrometeorological Data for Streamflow Estimation and to Improve Flood Monitoring." *Computers and Geosciences* 111(October 2017):148–58. doi: 10.1016/j.cageo.2017.10.010.

Röder, M., Both, A., & Hinneburg, A. (2015, February). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399-408).

San-Miguel-Ayanz, J., E. Schulte, G. Schmuck, A. Camia, P. Strobl, G. Liberta, C. Giovando, R. Boca, F. Sedano, P. Kempeneers, D. McInerney, C. Whitmore, S. Santos de Oliveira, M. Rodrigues, T. Durrant, P. Corti, F. Oehler, L. Vilar, and G. Amatulli, 2012. Comprehensive monitoring of wildfisres in Europe: The European Forest Fire Information System (EFFIS), Approaches to Managing Disaster - Assessing Hazards, Emergencies and Disaster Impacts (J. Tiefenbacher, editor), pp. 87–105, InTech, ISBN 978-953-51-0294-6.

Sodoge, Jan, Christian Kuhlicke, and Mariana Madruga de Brito. 2023. "Automatized Spatio-Temporal Detection of Drought Impacts from Newspaper Articles Using Natural Language Processing and Machine Learning." *Weather and Climate Extremes* 41(March):100574. doi: 10.1016/j.wace.2023.100574.

Stein, Lina, S. Karthik Mukkavilli, Birgit M. Pfitzmann, Peter W. J. Staar, Ugur Ozturk, Cesar Berrospi, Thomas Brunschwiler, and Thorsten Wagener. 2024. "Wealth Over Woe: Global Biases in Hydro-Hazard Research." *Earth's Future* 12(10). doi: 10.1029/2024EF004590.

Syed, S., & Spruit, M. (2017, October). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)* (pp. 165-174). Ieee.

Tanyaş, Hakan, Tolga Görüm, Islam Fadel, Cengiz Yıldırım, and Luigi Lombardo. 2022. "An Open Dataset for Landslides Triggered by the 2016 Mw 7.8 Kaikōura Earthquake, New Zealand." *Landslides* 19(6):1405–20. doi: 10.1007/s10346-022-01869-9.

Tanyaş, Hakan, Cees J. van Westen, Kate E. Allstadt, M. Anna Nowicki Jessee, Tolga Görüm, Randall W. Jibson, Jonathan W. Godt, Hiroshi P. Sato, Robert G. Schmitt, Odin Marc, and Niels Hovius. 2017. "Presentation and Analysis of a Worldwide Database of Earthquake-Induced Landslide Inventories." *Journal of Geophysical Research: Earth Surface* 122(10):1991–2015. doi: 10.1002/2017JF004236.

Tatli, Hasan, and Murat Türkeş. 2014. "Climatological Evaluation of Haines Forest Fire Weather Index over the Mediterranean Basin." *Meteorological Applications* 21(3):545–52. doi: 10.1002/met.1367.

Taylor, Faith E., Bruce D. Malamud, Katy Freeborough, and David Demeritt. 2015. "Enriching Great Britain's National Landslide Database by Searching Newspaper Archives." *Geomorphology* 249:52–68. doi: 10.1016/j.geomorph.2015.05.019.

Yetmen, Hurşit, and Ahmet Serdar Aytaç. 2017. "Journal of Current Researches on Social Sciences Influence of the Meteorological Conditions on Forest Fires in Winter and Spring in Eastern Black Sea Region : Case Study on Çamburnu (Sürmene) Forest Fire Meteorolojik Koşulların Doğu Karadeniz ' de Kış v." doi: 10.26579/jocress-7.2.26.

UNISDR, 2015. Sendai framework for disaster risk reduction 2015 -2030. Geneva. van Valkengoed, A.M., Steg, L., 2019. Meta-analyses of factors motivating climate change adaptation behaviour. Nat. Clim. Chang. 9, 158–163. https://doi.org/10.1038/s41558-018-0371-y

U.S. Geological Survey, 2008. Magnitude 7.9 — Eastern Sichuan, China, 2008 May 12 06:28:01 UTC. http://earthquake.usgs.gov/earthquakes/eqinthenews/2008/ us2008ryan/.

Waltham, A.C., Fookes, P.G., 2003. Engineering classification of karst ground conditions. Quarterly Journal of Engineering

Geology and Hydrogeology 36, 101–118.

685