Big Earth Data

ISSN: 2096-4471 (Print) 2574-5417 (Online) Journal homepage: www.tandfonline.com/journals/tbed20

A multimodal GeoAI approach to combining text with spatiotemporal features for enhanced relevance classification of social media posts in disaster response

David Hanny, Sebastian Schmidt, Shaily Gandhi, Michael Granitzer & Bernd Resch

To cite this article: David Hanny, Sebastian Schmidt, Shaily Gandhi, Michael Granitzer & Bernd Resch (23 Oct 2025): A multimodal GeoAI approach to combining text with spatiotemporal features for enhanced relevance classification of social media posts in disaster response, Big Earth Data, DOI: 10.1080/20964471.2025.2572140

To link to this article: https://doi.org/10.1080/20964471.2025.2572140

BIG EARTH DATA
https://doi.org/10.1080/20964471.2025.2572140

RESEARCH ARTICLE

OPEN ACCESS Check for updates

A multimodal GeoAI approach to combining text with spatiotemporal features for enhanced relevance classification of social media posts in disaster response

David Hanny[a], Sebastian Schmidt[a,b], Shaily Gandhi[a], Michael Granitzer[a,c] and Bernd Resch[a,d]

[a]GeoSocial Artificial Intelligence, IT: U Interdisciplinary Transformation University Austria, Linz, Austria; [b]Department of Geoinformatics - Z_GIS, University of Salzburg, Salzburg, Austria; [c]Chair of Data Science, University of Passau, Passau, Germany; [d]Center for Geographic Analysis, Harvard University, Cambridge, MA, USA

ABSTRACT

Geo-referenced social media data supports disaster management by offering real-time insights through user-generated content. To identify critical information amid high volumes of noise, classifying the relevance of posts is essential. Most existing methods primarily use textual features, neglecting spatial and temporal context despite its importance in determining relevance. This study proposes a multimodal approach that integrates text with spatiotemporal features for relevance classification of geo-referenced social media posts. We evaluate our method on 4,574 manually labelled posts from five disasters: the 2020 California wildfires, 2021 Ahr Valley floods, 2023 Chile wildfires, 2023 Turkey earthquake and 2023 Emilia-Romagna floods. Labels were assigned based on text, geographic location and time. Our spatiotemporal features include proximity to disaster impact sites, local co-occurrences with disaster-related posts, event type and geographic context. When utilised on their own, they achieved a macro F1 score of 0.713 with a random forest classifier. A fine-tuned TwHIN-BERT-base model using only text scored 0.779. For multimodal classification, we tested feature concatenation, in-context learning, stacking and partial stacking. Partial stacking produced the highest macro F1 score (0.814). Our multilingual, context-aware classification approach lays the groundwork for more integrated GeoAI applications in disaster management, the social sciences and beyond.

ARTICLE HISTORY
Received 28 April 2025
Accepted 17 September 2025

KEYWORDS
Machine learning; GeoAI; social media; disaster management; multimodal learning; relevance classification

1. Introduction

Natural disasters such as floods, wildfires or earthquakes have become increasingly frequent and severe due to factors like climate change, urbanisation and environmental degradation (Sauerborn & Ebi, 2012). These events cause significant disruptions, leading

CONTACT David Hanny david.hanny@it-u.at GeoSocial Artificial Intelligence, IT:U Interdisciplinary Transformation University Austria, Altenberger Straße 66c, Science Park 4, OG 2, Linz 4040, Austria
Supplemental data for this article can be accessed online at https://doi.org/10.1080/20964471.2025.2572140

to loss of life, economic damage and humanitarian crises. Effective disaster response relies on timely and accurate information to assess the situation and coordinate relief efforts. However, traditional information-gathering methods, such as official reports and satellite imagery, often face delays, making real-time data sources essential for improving situational awareness (Havas et al., 2017).

Social media platforms offer a massive stream of real-time, crowd-sourced information during natural disasters (Wang & Ye, 2018). However, extracting useful information from social media remains a significant challenge due to the high volume, noise and multimodal nature of posts. Relevance classification plays a crucial role in filtering these large amounts of data, helping to distinguish between critical information and less useful content for disaster management purposes. Existing studies on relevance classification have mainly focused on text classification (Kaufhold et al., 2020) or on integrating text with images (Koshy & Elango, 2023a). However, social media content is frequently also geographically or temporally referenced (Serere et al., 2023), which allows for categorisation based on spatiotemporal attributes.

In this context, the integration of spatial and temporal features into text classification has hardly been addressed by previous studies. This is especially notable since multimodal fusion has become an important research topic in computer science (Jiao et al., 2024). Kaufhold et al. (2020) incorporated the geographical and temporal distance of posts to disaster events for relevance classification by concatenating them with (tf-idf) vectors. However, as their labelling procedure was primarily content-based, the integration resulted in only a modest improvement in classification accuracy of 0.0043. Scheele et al. (2021) used a Convolutional Neural Network (CNN) to integrate text with spatiotemporal features in the case of Hurricane Sandy, classifying social media posts into five categories of disaster-related information such as safety advice, reports of damage or requests for aid. The integration of spatiotemporal features resulted in an increase in macro F1 score of up to 0.06 over the text-only model. Despite these earlier efforts, there is still limited research on how to effectively combine spatial and temporal features with modern transformer-based models. BERT, in particular, has been shown to significantly outperform traditional text classification methods such as tf-idf-based approaches and CNNs (Blomeier et al., 2024). However, how to incorporate spatiotemporal features into BERT-based models, especially for relevance classification, remains largely unexplored.

Building on these findings, we present a multimodal relevance classification approach that extends prior work by integrating textual, spatial and temporal information both during labelling and model training. In particular, we evaluate methods for early (feature-level), late (decision-level) and hybrid multimodal fusion. Our methodology is evaluated across five diverse natural disaster scenarios, varying in type, language, covered region and spatial extent: the 2020 California wildfires (USA), the 2021 Ahr Valley floods (Germany), the 2023 Chile wildfires, the 2023 Turkey earthquake, and the 2023 Emilia-Romagna floods (Italy). To represent non-textual context, we engineered a 13-dimensional feature set capturing geographic and temporal proximity to disaster events, local co-occurrence patterns of disaster-related posts across multiple radii and contextual encodings of event type and location. Some of these features were based on Earth Observation (EO) data, such as flood outlines or burnt areas, which we used to calculate distances in geographic space and time. The resulting feature set was statistically analysed and served as input to a range of machine learning classifiers, including logistic

regression, random forest, gradient boosting, Support Vector Machine (SVM), *k*-Nearest Neighbours (kNN), and naive Bayes, covering diverse learning paradigms. For text classification, we fine-tuned TwHIN-BERT-base, a multilingual transformer model pre-trained on 7 billion tweets. To integrate text and non-text features, we evaluated four strategies: direct feature concatenation with classification heads of varying complexity, in-context learning with structured inputs, and both full and partial stacking based on ensemble learning with a meta-learner. The resulting multimodal framework allows for the robust classification of social media relevance across diverse disaster scenarios.

Based on this setup, our study is guided by the following research questions:

- **RQ1**: Which spatial and temporal properties of geo-referenced social media data can most effectively enhance relevance classification for disaster response?
- **RQ2**: How can text be integrated with non-text features for effective multilingual relevance classification?
- **RQ3**: How does classification performance differ across use cases with various disaster types, languages and geographic regions?

The contributions of this study consequently are as follows:

(1) We present a dataset of 4,574 geo-referenced social media posts in more than eight languages from five natural disasters, manually annotated for relevance to disaster response using a multimodal labelling approach that incorporates textual content, geographic location and time.
(2) We engineer and statistically evaluate a 13-dimensional feature set that captures geographic and temporal proximity to disaster events, local co-occurrence patterns with disaster-related posts and contextual information like the disaster type and region.
(3) We propose and compare four distinct strategies for integrating text and spatiotemporal features for relevance classification: feature concatenation, in-context learning, full stacking and partial stacking.
(4) We conduct a comprehensive evaluation of classification performance across the five disaster scenarios and apply SHapley Additive exPlanations (SHAP) to quantify the influence of the engineered spatiotemporal features on model predictions.

2. Related work

Our work builds on previous efforts regarding social media analysis for disaster management, relevance classification and the integration of spatiotemporal features in text classification.

2.1. Geo-social media analysis for disaster management

Geo-social media has been shown to be an essential data source to accompany disaster management, providing real-time, crowd-sourced information to aid in emergency responses and recovery efforts (Wang & Ye, 2018). Since data from social media platforms is very diverse and noisy, a plethora of methods have been

proposed to identify and leverage disaster-related information. In this context, identifying social media posts relevant to a natural disaster constitutes a critical preliminary step, as it forms the foundation for reliable event detection and situational awareness. There are simple approaches that use keyword filtering combined with thresholds (Shah et al., 2021) or rely on disaster-related hashtags (Chowdhury et al., 2020). A keyword-based query strategy that iteratively updates the keyword list by ranking tf-idf weights from previously identified posts was introduced by Chen and Lim (2018). Similarly, Yigitcanlar et al. (2022) examined disaster-related Tweets in Australia through an analysis of word frequency and co-occurrence patterns. Other studies employ unsupervised methods to group similar texts: Havas and Resch (2021) used Latent Dirichlet Allocation (LDA) to cluster Tweets into semantic topics for real-time natural disaster monitoring. Zhou and Chen (2014) introduced an enhanced LDA-based method that incorporates the location and time of a post, utilising similarity joins from a social media stream. Hanny and Resch (2024) leveraged a Geo-GSOM and multimodal embeddings, which included semantics, sentiments, space and time for multimodal topic modelling, demonstrating the utility of their method for flood-related tweets.

There are also many supervised approaches that fine-tune Natural Language Processing (NLP) models to specifidisaster-related topics or problems. Huang et al. (2022) utilised BERT embeddings in a Bidirectional LSTM (Bi-LSTM) network with an attention mechanism to detect emergency-related posts on Sina Weibo. To further refine event classification, they applied unsupervised dynamical clustering based on textual similarity. Hanny et al. (2024) emphasised the advantage of AL combined with generic fine-tuning to enhance the accuracy of binary disaster-relatedness classification. Similarly, Powers et al. (2023) and Paul et al. (2023) showcased the effectiveness of pre-trained transformer models and hybrid deep neural networks in classifying disaster-related tweets and extracting fine-grained humanitarian information, outperforming classical models. Tekumalla and Banda (2022) showed that noisy learning can improve performance for both balanced and imbalanced disaster-related social media data, using different machine learning models and achieving the best performance with RoBERTa. There are also several studies that combine textual and imagery information. Koshy and Elango (2023a) fine-tuned a RoBERTa model with the CrisisMMD dataset and combined it with a Vision Transformer model to classify multimodal Twitter data. Similarly, Madichetty and Madisetty (2023) employed RoBERTa alongside VGG-16, integrating both textual and visual content related to disasters by multiplying the output class probabilities. Adwaith et al. (2022) evaluated various multimodal architectures for processing disaster-related text and imagery, finding that a combination of RoBERTa with ResNet or DenseNet yielded

Contribution beyond State of the Art (SotA)

Prior work has demonstrated the suitability of textual content for information extraction from social media posts during and after natural disasters. However, spatiotemporal context such as geographic and temporal proximity to disaster has mostly been incorporated as a separate or post hoc analysis step. This study addresses this limitation by integrating spatiotemporal context directly with textual information for social media content classification. Specifically, spatiotemporal features are fused with text in a multimodal and spatially explicit machine learning setting.

the best results. Li et al. (2022) integrated textual and visual information in a Graph Neural Network (GNN) to enhance classification performance.

2.2. Relevance classification of social media data in disaster response

Within the broader scope of geo-social media analysis for disaster management, a critical subtask is the classification of relevance, that is, identifying which social media posts actually contribute valuable information for emergency responders and decision-makers. Accurately classifying relevance is essential for filtering vast amounts of crowd-sourced data and ensuring that responders are not overwhelmed by noise (Vieweg et al., 2010). A central challenge in this area lies in determining what constitutes a relevant post for disaster management. Early work has adopted binary formulations of relevance, largely focusing on distinguishing posts that are useful for crisis response from those that are not. For example, Kaufhold et al. (2020) conceptualised relevance as a combination of abstract criteria (topicality, impersonality, necessity) and factual criteria (geographic and temporal proximity). Similarly, Agarwal and Yiliyasi (2010) framed *contextual relevance* as the degree of usefulness of the data for a certain task. Naturally, this definition can be extended to crisis situations where Jensen (2012) considered relevance as an aggregated indicator of usefulness, adequacy and value of information. More recent efforts have moved toward fine-grained and multi-class relevance classification, arguing that a binary framework fails to capture the nuances in post informativeness. de Albuquerque et al. (2015) proposed a three-class taxonomy: *off-topic*, *on-topic and relevant*, and *on-topic but irrelevant*, with relevance defined through a post's ability to enhance situational awareness. In a similar vein, Olteanu et al. (2015) introduced the categories *related and informative*, *related but not informative*, and *not related*, using a content-based framework that emphasised the presence of actionable or contextual information—such as mentions of affected populations, damage reports or advice. Extending this logic, Blomeier et al. (2024) employed an ordinal scale with four levels of relevance ranging from *very relevant* to *not relevant*, redefining relevance in terms of the practical helpfulness of posts for crisis responders.

Methodologically, various techniques have been applied to classify individual posts into relevance categories. Kaufhold et al. (2020) applied traditional machine learning methods, including a random forest and naive Bayes classifier, to categorise posts using tf and tf-idf vectors. More recent works increasingly rely on deep learning. CNN-based architectures (Huang et al., 2020), transformer models such as BERT (Blomeier et al., 2024; Madichetty et al., 2021) and GNNs (Li et al., 2022) have been employed to capture both syntactic and semantic aspects of social media posts. Multimodal approaches that incorporate both textual and visual information have also been used to classify relevance and informativeness (Adwaith et al., 2022; Koshy & Elango, 2023a).

A comprehensive overview of previous works, including the utilised relevance categories and methodologies, is provided in Table 1. Notably, the definition of relevance often varies between studies, reflecting differing operational goals. Some approaches prioritise situational awareness and actionability, while others emphasise thematic or contextual alignment with a disaster event. This variability complicates comparability across datasets and models. In addition, most previous studies focus on English-language data and, apart from Kaufhold et al. (2020), do not integrate contextual spatiotemporal features.

Table 1. Overview of literature on relevance classification of social media posts in disaster response, sorted descendingly by publishing year.

Paper	Classification methodology	Categories
Blomeier et al. (2024)	BERT, CNN, SVM, RF, NB	very relevant, rather relevant, barely relevant, not relevant
Koshy and Elango (2023b)	CNN, Bi-LSTM	need, availability, other
Powers et al. (2023)	NB, AdaBoost, SVM, MLP, logistic regression, ridge classifier, CNN, BERT, XLNet	relevant, not relevant
Adwaith et al. (2022)	XLNet, BERT, RoBERTa	informative, not informative
De Bruijn et al. (2020)	CNN	relevant, irrelevant
Kaufhold et al. (2020)	RF, NB	relevant, irrelevant
Madichetty and Sridevi (2019)	CNN, ANN	informative, non-informative
Derczynski et al. (2018)	SVM	informative, somewhat informative, not informative
Habdank et al. (2017)	NB, decision tree, RF, SVM, neural network	relevant, irrelevant
Nguyen et al. (2017)	CNN	affected individual, donations and volunteering, infrastructure and utilities, sympathy and support, other useful information, not related or irrelevant
Caragea et al. (2016)	CNN	informative, not informative
Imran et al. (2016)	NB, SVM, RF	several subcategories of relevance (e.g., "missing, trapped, or found people"), irrelevant
de Albuquerque et al. (2015)	keyword filtering	on-topic and relevant, on-topic but irrelevant, off-topic
Olteanu et al. (2015)	manual labelling	related and informative, related but not informative, not related to the crisis
Imran et al. (2013)	NB	informative, personal, other
Verma et al. (2011)	NB, maximum entropy	contributing to situational awareness, not contributing to situational awareness

Contribution beyond SotA

Most existing studies on relevance classification focus primarily on textual or visual content and are limited to individual case studies. To address these limitations, we adopt a multimodal labelling approach that incorporates textual, spatial and temporal information and systematically evaluate the predictive power of (1) spatiotemporal features, (2) textual content and (3) their combination. Our analysis spans across five disaster scenarios in different countries, covering various languages and disaster types, providing a broader and more generalisable assessment.

2.3. Multimodal fusion strategies

The fusion of data sets with different characteristics, e.g., spatial scales, and modalities, e.g., space and time, based on machine learning approaches has received increased attention in recent years. In general, a distinction is made between three types of fusion: First, there is *early fusion*, which merges data sources and modalities at the beginning of the processing pipeline. Typically, this is done by concatenating extracted unimodal features into a joint representation (Snoek et al., 2005). Second, *late fusion* combines the results of separate unimodal processing streams by merging the results in the very end. In the simplest case, this could be done using summation or averaging of prediction values (Gadzicki et al., 2020). In between these two extremes, there also exist *hybrid fusion* approaches, which combine the properties

of the two methods (D'mello & Kory, 2015), or *halfway/middle fusion*, where data or modalities are fused in the middle of the network (Damer et al., 2019).

Fusion methods are especially important in scientific fields where data with differing characteristics must be integrated, such as robotics (Duan et al., 2022), autonomous driving (Shahian Jahromi et al., 2019), healthcare (Jiao et al., 2024), and geoinformatics, where spatial data is inherently multimodal (Nikparvar & Thill, 2021). In this context, hybrid fusion has received growing attention in recent years as it enables more complex integration strategies. For instance, in Deep Neural Networks (DNNs), modalities can be fused by combining intermediate representations within the network architecture (e.g., Scheele et al. (2021)). Many studies use attention mechanisms, in particular cross-attention, to achieve this (Li & Tang, 2024; Tang et al., 2024). There are also GNN-based approaches which capture multimodal datasets as graphs, where nodes can represent unimodal or cross-modal information, and edges encode intra- and inter-modal relationships (Chen et al., 2023; Li et al., 2023). Additionally, contrastive learning has been used to learn joint representations of multimodal data by aligning information from different modalities in a shared embedding space (Radford et al., 2021). Fusion approaches have also been developed to handle missing modalities, often by leveraging cross-attention or contrastive learning techniques to maintain performance despite incomplete data (Liu et al., 2023).

With regard to spatial data, fusion approaches are popular in remote sensing, where different modalities include temporal, spectral and spatial resolutions. Proposed fusion methods include pansharpening through CNNs (Li et al., 2022), the fusion of visible and infrared imagery through cross-modal transformers (Park et al., 2024), or the late fusion of LiDAR data with optimal imagery (Bultmann et al., 2023). The late fusion of remote sensing data and contents from social media has also been proposed using log-linear pooling (Wieland et al., 2025).

2.4. Integration of spatiotemporal features in text classification

In social media analysis, spatial and temporal information is frequently employed to gain localised insights into online content. However, in most cases, it is applied in pre-processing as a filtering criterion or in post-processing stages as a means of contextualising model outputs (e.g., Parimala et al. (2021); Sodoge et al. (2023)). This sequential treatment limits the potential for models to learn joint patterns between textual and contextual signals.

A small number of previous works have attempted to integrate spatial and temporal features directly into the classification process. Scheele et al. (2021)

Contribution beyond SotA

Prior research has introduced various fusion techniques across domains. However, there remains limited work that systematically evaluates how spatiotemporal features can be integrated with text using different variants of multimodal fusion. Our study addresses this gap by explicitly comparing early, late, and hybrid fusion strategies for disaster relevance classification of geo-referenced social media posts.

proposed a CNN-based text mining approach that incorporates information about disaster phase and spatiotemporal proximity in the context of Hurricane Sandy. Their model classified social media posts into five humanitarian categories like caution and advice, damage reports or requests for aid. They specifically incorporated the geographic distance to the event, posting date, disaster status and contextual weather information. The text was first processed using a CNN, and the high-dimensional text embedding retrieved through max-pooling was concatenated with the non-text features. Subsequently, the concatenated feature vector was used as input to a classification head consisting of two fully connected layers. This architectural modification led to an increase in macro F1 score of up to 0.06 over the text-only baseline. Similarly, Kaufhold et al. (2020) integrated text with spatiotemporal non-text features for binary relevance classification by concatenating them tf-idf vectors and applying a random forest classification model on top. However, the integration of these contextual features only led to a very marginal increase in accuracy of up to 0.0043. Furthermore, Tian et al. (2023) investigated the extraction of spatiotemporal information from Wikipedia by classifying sentences into four types ranging from *strong* to *no* spatiotemporal relevance. In their best approach, input sentences were first processed through a pre-trained BERT model to generate contextualised word embeddings which were then passed to an Region-based Convolutional Neural Network (RCNN) layer and a fully connected layer for the final prediction. Their BERT-RCNN reached a macro F1 score of 0.74 compared to the text-only BERT baseline which yielded a macro F1 score of 0.72.

Other works have proposed spatiotemporal extensions to topic modelling or clustering tasks. Hanny and Resch (2024) introduced a topic modelling approach that embeds geographic location and time into a customised Geo-GSOM, enabling the identification of spatiotemporally coherent topic clusters. Likewise, Autelitano et al. (2019) developed a density-based clustering method for keyword extraction that explicitly incorporates both time and space to enhance the detection of disaster-related terms. There are also studies that explore semantic representations of spatiotemporal context without necessarily integrating them into the core classifier. Alfarrarjeh et al. (2017) estimated spatial sentiment distributions by partitioning the social media stream into space-time cells, though the spatiotemporal dimension remained external to the sentiment model itself. Popova et al. (2023) approached the problem from an ontological perspective, proposing a method to align and synchronise spatiotemporal content across platforms by mapping semantic relationships between places.

3. Materials and methods

The core methodology of our study consists of three parts:

(1) We collected geo-referenced posts from Twitter (now X) for five use case scenarios across the world. A subset of these posts was then categorised into three relevance classes by three human annotators, considering text, post location and posting time.

Contribution beyond SotA

Existing research on the fusion of text with spatiotemporal features lacks a systematic comparison of different information fusion strategies. Our study contributes to this gap by evaluating and comparing multiple such strategies for disaster relevance classification. The evaluated strategies include concatenation, in-context learning and ensemble learning (stacking). In addition, we assess the predictive power of spatiotemporal features using XAI, an aspect that remains largely unexplored in prior work.

(2) We engineered a 13-dimensional set of spatiotemporal (contextual) non-text features that capture geographic and temporal proximity, co-occurrence with disaster-related posts, the event type and the disaster region. We then analysed how these features vary across the three relevance classes.
(3) We trained and evaluated a series of machine learning models on our non-text features, the texts and both. For each feature configuration, we evaluated appropriate classification methods, including transformer neural networks, tree-based methods, logistic regression, SVM the model predictions, we additionally investigated the and a naive Bayes classifier. To better understand impact of features on the model output using SHAP.

A visual overview of this study workflow is available in Figure 1.

3.1. Data

To obtain a diverse dataset that covers different natural disasters across several regions of the world, we considered five use case scenarios: (1) the 2020 California wildfires (Safford et al., 2022) (2) the 2021 Ahr Valley floods (Koks et al., 2021) in Western Germany, (3) the 2023 Chile wildfires (Cordero et al., 2024), (4) the 2023 Emilia-Romagna floods in Italy (Arrighi & Domeneghetti, 2024) and (5) the 2023 Earthquake in Turkey (Dal Zilio & Ampuero, 2023). For each of the disaster events, we collected geo-referenced social media posts using the former Twitter v1.1 and v2 Recent Search and Filtered Stream API endpoints, filtering for posts with a geographic reference using a bounding box and temporal limits. The geographic reference can either be a point latitude/longitude coordinate or a bounding box. For the 2021 Ahr Valley floods, we additionally included data from (Blomeier et al., 2024) which covers the larger region of Germany and the entire year of 2021. This enabled a critical assessment of our labelling and classification approach relative to their methodology. Furthermore, the bounding boxes used for data collection across the selected use cases varied significantly in size, allowing for an evaluation of our method for different filter granularities. Table 2 provides an overview of the spatial and temporal properties of the obtained data, including the distribution of detected languages available in the post metadata. The language label *und* refers to posts where the language could not be identified by Twitter. This is often caused by content that consists only of media, hashtags or mentions. Figure 2 shows the geographic distribution of our Area Of Interests (AOIs).

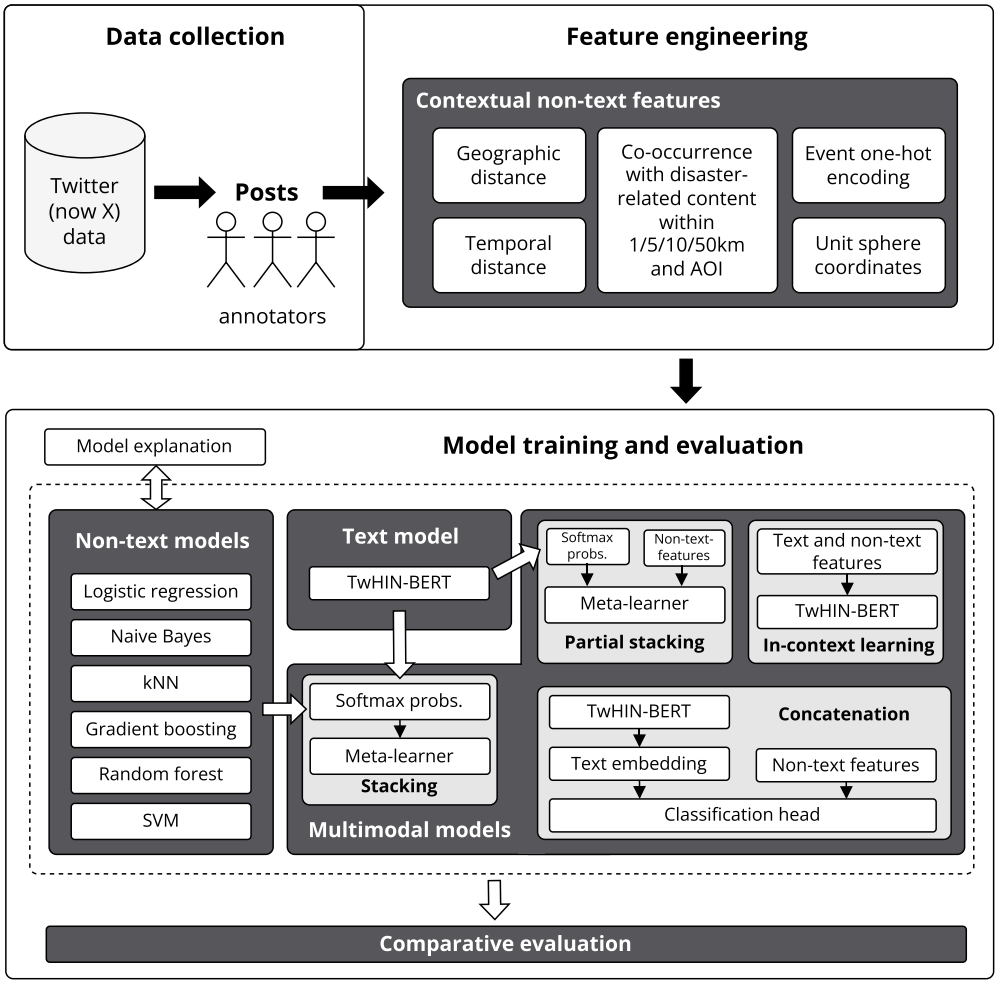

Figure 1. Methodological overview of our study, including all evaluated approaches. For operational use, only one method from the "model training and evaluation" section can be selected.

3.2. Labelling

To create a labelled dataset for our five use case scenarios, we used a hierarchical, multimodal three-class scheme inspired by Olteanu et al. (2015), consisting of the following relevance classes:

(1) **Related and relevant**: A post that is related to the respective natural disaster and relevant for emergency responders. It contains useful information for supporting disaster management (e.g., posts about destructions, in-situ information, critical infrastructure, affected individuals, affected areas, requests for help, caution or advice). Such posts typically originate from locations in close spatial proximity to the affected area and are posted during or shortly after the disaster (Vieweg et al., 2010).

Table 2. Spatial and temporal properties of the collected geo-social media posts along with the distribution of languages, sorted in descending order by their relative fraction, before further processing (*und* = undefined). The 2021 Ahr Valley floods dataset is made up of two parts: (1) posts collected via the former Twitter API and a regional bounding box and (2) the data from Blomeier et al. (2024) which spans across the entire country of Germany and the year 2021.

Event	#Posts	Languages	Temporal filter	Bounding box filter
2020 California wildfires	34,678,576	en (88.3%), es (2.6%), de (1.4%), it (0.7%), other (7%)	2020–04–01 to 2020–12–31	POLYGON ((−112.85 31.83, −112.85 42.58, −124.43 42.58, −124.43 31.83, −112.85 31.83))
2021 Ahr Valley floods (part 1)	9,719	de (53.1%), en (22.8%), und (5.0%), fi (2.2%), nl (2%), fr (1.2%), other (13.7%)	2021–07–01 to 2021–07–31	POLYGON ((6.00 50.01, 8.00 50.01, 8.00 50.75, 6.00 50.75, 6.00 50.01))
2021 Ahr Valley floods (part 2)	3,304	de (94.7%), und (2.8%), en (1.4%), nl (0.3%), other (0.8%)	2021–01–01 to 2021–12–31	POLYGON ((15.53 46.90, 15.53 55.06, 5.00 55.06, 5.00 46.90, 15.53 46.90))
2023 Chile wildfires	1,979,072	es (89.0%), und (3.0%), en (2.9%), pt (1.5%), other (3.6%)	2023–01–01 to 2023–06–29	POLYGON ((−65.23 -54.81, −65.23 -17.32, −74.26 -17.32, −74.26 -54.81, −65.23 -54.81))
2023 Emilia-Romagna floods	43,443	it (66.2%), und (10.5%), en (9.9%), es (3.3%), other (10.1%)	2023–04–24 to 2023–05–30	POLYGON ((13.00 43.66, 13.00 45.29, 9.75 45.29, 9.75 43.66, 13.00 43.66))
2023 Turkey earthquake	2,723,526	tr (80.1%), und (8.5%), en (2.6%), ar (0.7%), other (8.1%)	2023–01–01 to 2023–03–30	POLYGON ((43.95 33.80, 43.95 42.32, 29.18 42.32, 29.18 33.80, 43.95 33.80))

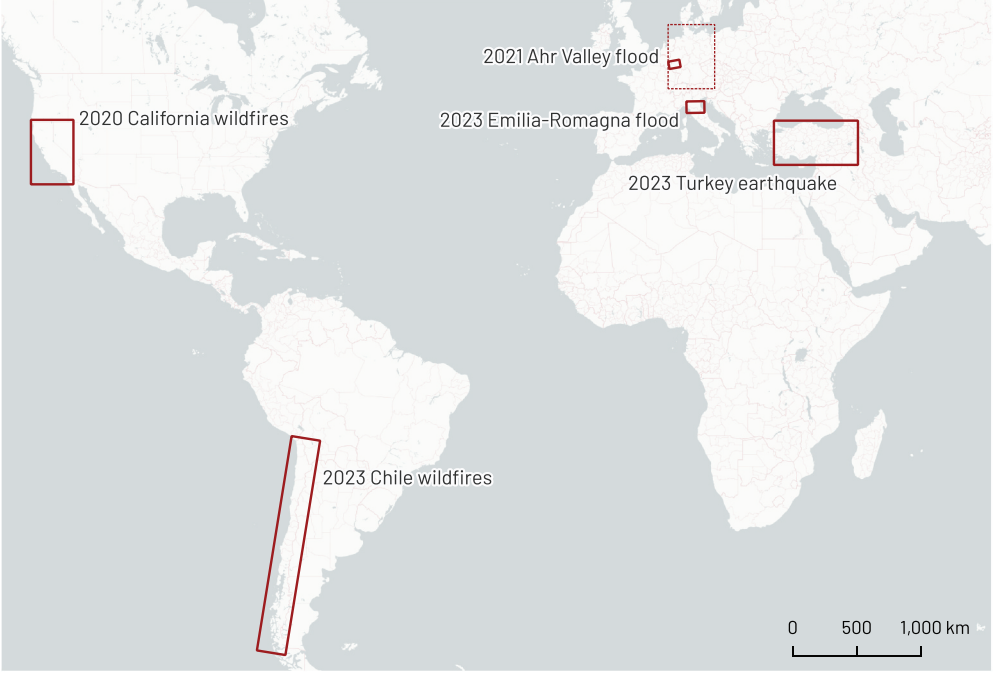

Figure 2. Geographic delineation for the data collection our five use case scenarios. The 2021 Ahr Valley floods dataset is made up of two parts: (1) posts collected via the former Twitter API and a regional bounding box and (2) the data from Blomeier et al. (2024) which spans across Germany.

(2) **Related but not relevant**: A post that refers to the respective disaster but does not contain helpful or valuable information for supporting disaster management (e.g., declarations of solidarity, volunteering initiatives, appeals for donations, political or religious statements, bot-generated content, comparisons to past events, shared news articles). These posts may originate from within or outside the affected area and may be posted during, after or shortly before the disaster.
(3) **Not related**: A post that has no relation to the disaster event in question, regardless of its location or time.

The hierarchical structure is depicted in Figure 3. We chose this class scheme because it allows for several binary sub-distinctions, which make the data useful for other purposes like disaster-relatedness classification as it enables a straightforward separation between disaster-related and unrelated content. Furthermore, posts can be separated regarding binary relevance. The proposed categories can also be viewed as ordinal where the ordering is based on the usefulness for emergency responders. That is, Not related < Related but not relevant < Related and relevant.

As we selected our raw datasets only based on spatial and temporal conditions, the vast majority of posts was unrelated to any disaster. Labelling those without further filtering would result in a highly imbalanced class distribution in the annotated data. Therefore, we pre-filtered the data using disaster-relatedness and few-shot relevance classification. First, we used the Disaster-Twitter-XLM-RoBERTa-

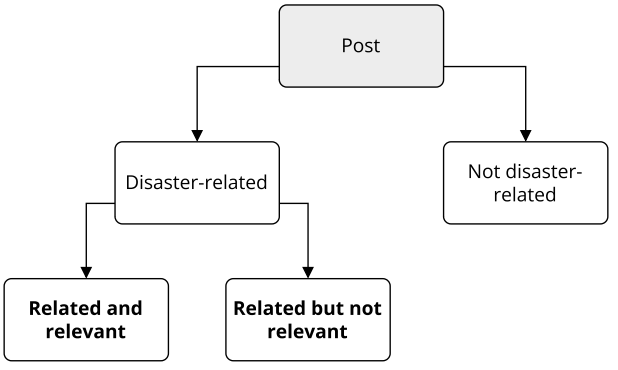

Figure 3. Hierarchical definition of relevance categories used throughout the study.

AL model by Hanny et al. (2024) to extract a subsample of $\leq 10,000$ posts per use case. This hard threshold was chosen as an upper limit to ensure we had a sufficient pool for downstream sampling. We randomly selected posts such that 67% were disaster-related and 33% were not, aligning with our three-class relevance scheme, in which two classes represent disaster-related content. In total, this preliminary subset consisted of 37,826 posts. Subsequently, we used GPT-4o-mini (OpenAI, 2024) with in-context learning to conduct an initial categorisation of each post into one of our three relevance categories. The model was provided with five labelled examples per use case and class descriptions similar to the ones defined above. To keep the annotation workload manageable, we then sampled ≤ 350 posts per class and use case based on the GPT-4o-mini predictions. This yielded a final dataset of 4,574 posts for human annotation. As part of this effort, each post was pre-processed by replacing user names and links with standardised tokens to avoid overfitting on specific users and external information from links. We also manually excluded all posts referring to a disaster event other than the one specified for the area of interest (e.g., mentions of an earthquake in Chile) to ensure a consistent analysis.

Next, each post was annotated based on textual content, geographic location and time, using a multi-step, consensus-based labelling approach with three human annotators. The annotators were all researchers in Geoinformatics with significant experience in disaster management and thus considered experts in the field. Only posts for which the categorisation of all human annotators matched, with discussion allowed, were considered further. Consequently, the Inter-Annotator Agreement (IAA) of the labelled dataset was 1.0. In the first step, posts were labelled purely based on their textual content. Subsequently, these initial labels were refined based on posting location and time. To ensure spatial relevance, the annotators agreed on a distance threshold of 100 kilometres from the disaster-affected area. Semantically relevant posts beyond this range were reclassified as "Related but not relevant," as the distance reduces their value for local emergency response. This threshold was informed by previous research: Sakaki et al. (2010) showed that the accuracy of tweet-based earthquake detection drops significantly beyond 100 kilometres, reflecting reduced situational awareness. Similarly, de Albuquerque et al. (2015) found that disaster-related tweets during floods were

concentrated within 100 kilometres of affected catchments. For wildfires, Sachdeva et al. (2017) reported that tweet locations correlated with observed particulate pollution, which can spread over distances of 100 kilometres or more (Joo et al., 2024). In addition, Schmidt et al. (2025) showed that locally posted wildfire-related tweets can be valuable for early warning and detection. Temporal relevance was defined based on emergency response timelines. After a natural disaster, the short-term response phase can last up to approximately 30 days (Federal Emergency Management Agency, 2024). Posts that were semantically relevant but published more than 30 days after the latest regional event were therefore labelled as "Related but not relevant," as they do not support immediate response efforts. Similarly, posts that occurred more than 7 days before the earliest regional event were considered as "Not related" because they lack the immediate temporal context required for disaster management. Previous studies have explored the use of social media for early warning and event detection (Wiegmann et al., 2021), and it has also been shown that such content is typically only indicative of an event a few days prior to impact (Schmidt et al., 2025). The hierarchical annotation guide used during labelling is visualised in Figure 4. Furthermore, the class distribution resulting from the multimodal labelling procedure is depicted in Table 3. Figure 5 additionally shows the language distribution within the labelled dataset. Notably, language distribution was strongly dominated by the official language of each affected country, with English as a frequent secondary language and other languages appearing marginally.

3.3. Feature engineering

To leverage the spatial and temporal properties of our data, we engineered five spatiotemporal non-text features, which cover a total of 13 dimensions. The features were designed to represent spatial and temporal proximity to the disaster event and contextual information about the event. A high-level overview of these features is available in Table 4.

3.3.1. Geographic and temporal distance

First, we considered the geographic distance of each post to the impact sites of each natural disaster, since the geo-location of social media posts may be useful for extracting in-situ information during an event (Vieweg et al., 2010). To effectively compute geographic distances, we collected event delineations for each use case. For the 2020 California wildfires, we downloaded the historical wildland fire perimeters from the California State Geoportal (Wallin, 2025). The flood delineations for the 2021 Ahr Valley floods in Germany were obtained from the Copernicus Emergency Management Service (EMS) (Copernicus EMS, 2021). Similarly, we collected burnt area delineations during the 2023 wildfires in Chile from the NASA Fire Information for Resource Management System (FIRMS) (NASA LANCE, 2023). For the 2023 Emilia-Romagna floods, we obtained flood delineations from the Geoportal of the Emilia-Romagna region (Agenzia per la Sicurezza Territoriale e Protezione Civile Regione Emilia-Romagna, 2024). Lastly, we collected the recorded earthquakes in Turkey and Syria in February 2023 from the United States Geological Survey (USGS) earthquake catalogue (USGS, 2023). This included the main earthquake and subsequent seismic activity in the affected regions of Turkey and Syria.

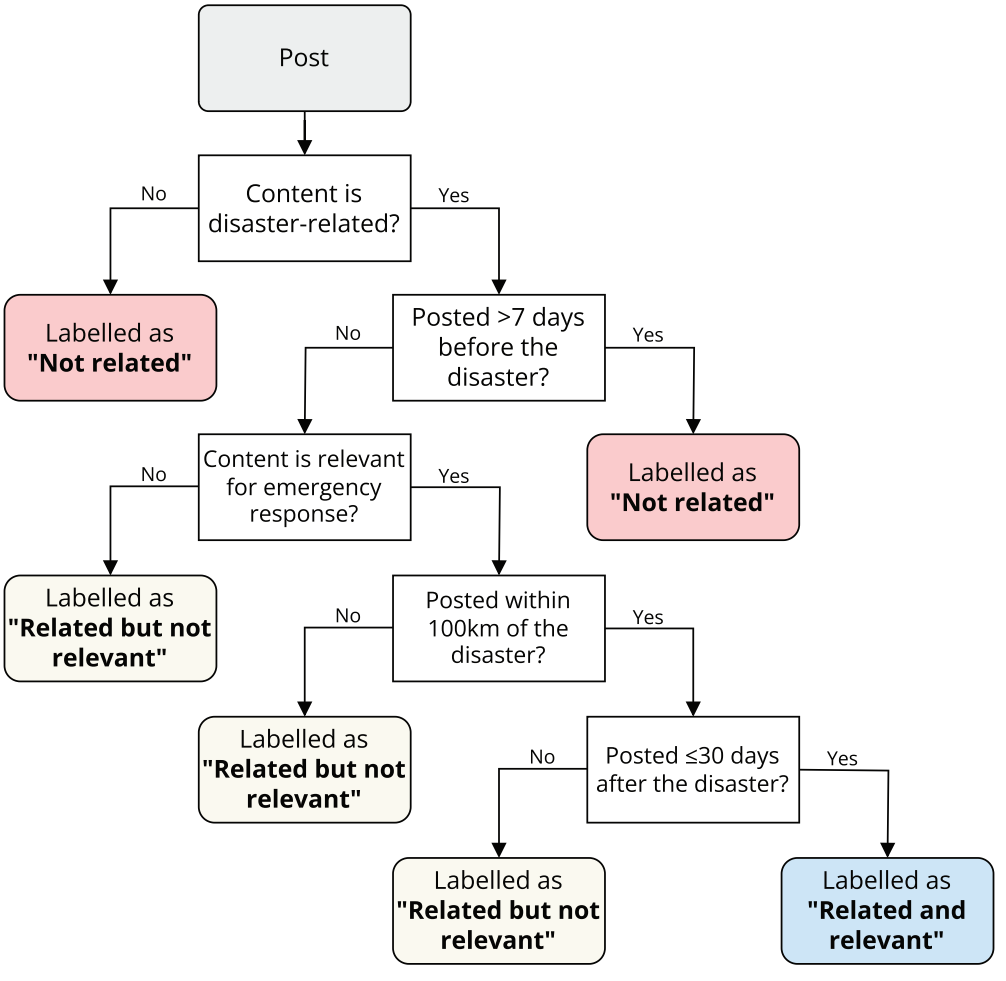

Figure 4. Visual annotation guide, depicting our hierarchical labelling procedure.

Table 3. Distribution of class labels in our annotated dataset before splitting it into training and test subsets.

Class	Labelled Posts
Not related	1,646
Related but not relevant	1,935
Related and relevant	993

Each dataset also contained temporal information about the respective event which we used to compute temporal distances. For this, we assumed that time is linear and continuous and used the timestamps provided by the Twitter API (without time zone). For the 2020 California wildfires, we took the alarm and containment date as an indicator for the start and end date of the fires. For the 2021 Ahr

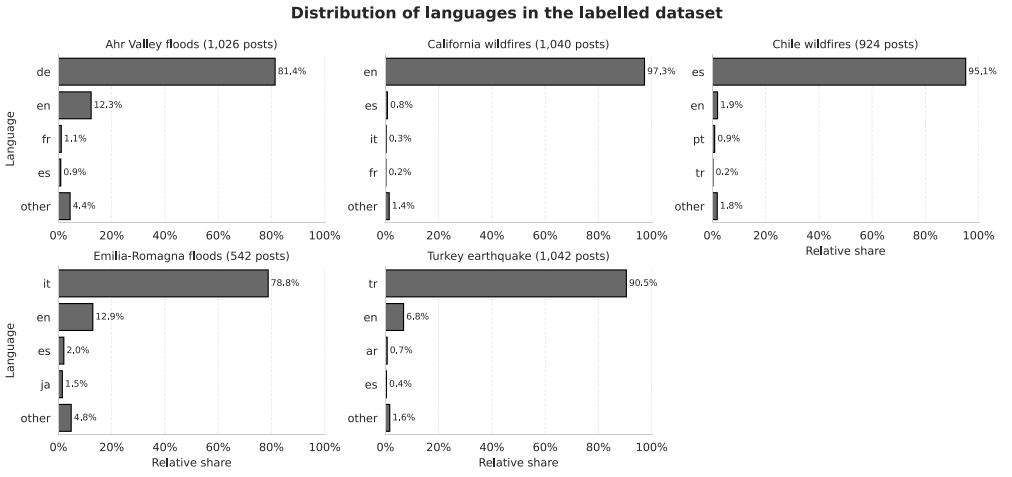

Figure 5. Relative distribution of languages in our labelled dataset of 4,574 posts. Languages are labelled according to their two-letter ISO 639 codes.

Table 4. Overview of the engineered non-text features used for relevance classification.

Name	Description	Dimensions
Geographic distance	Euclidean distance (in km) from the post location to the nearest disaster impact site	1
Temporal distance	Time difference (in hours) between the post timestamp and the nearest disaster event	1
Local co-occurrence counts	Number of disaster-related posts within the last 7 days, aggregated over radii of 1 km, 5 km, 10 km, 50 km, and across the entire AOI as a purely temporal aggregation	5
Event type encoding	One-hot encoding of disaster type: wildfire, flood, or earthquake	3
Location encoding	Spatial position of the AOI centroid projected onto a 3D unit sphere (x, y, z)	3

Valley data, we manually defined 2021–07–14 as a start date and 2021–07–17 as an end date (Koks et al., 2021) For the 2023 Chile wildfire data, burnt area and fire information were available for each individual day from 2023–01–01 to 2023–06–30. Because each burn was recorded per day, we did not need to define a general start or end period for the entire disaster. Instead, we treated the acquisition date of each burn as both its start and end date, which allowed us to compute spatial and temporal distances between posts and fire events on a per-day basis. The 2023 Emilia-Romagna floods occurred on two dates within May 2023, with separate flood delineations for each (Agenzia per la Sicurezza Territoriale e Protezione Civile Regione Emilia-Romagna, 2024). We therefore defined the start-to-end dates as 2023–05–02 to 2023–05–03 and 2023–05–16 to 2023–05–17, which were the two occurrence days, respectively. Lastly, we used the recording date of each seismic activity reported by the USGS as both start and end date for the 2023 Turkey earthquake data.

Based on this event information, we computed the geographic distance of each post to the nearest disaster impact site, i.e., flooded area, burnt area or earthquake epicentre. To achieve this, we first reprojected the post geometry (point coordinate or polygon) to match an equidistant coordinate system for the region of the use case (2020 California wildfires: *ESRI:102010*, 2021 Ahr Valley floods & 2023 Emilia-Romagna floods & 2023 Turkey earthquake: *ESRI:102031*, 2023 Chile wildfires: *ESRI:102032*), and then calculated the minimal Euclidean distance. To ensure that the selection of the nearest event was temporally coherent, we limited the candidate events to those occurring within 7 days of the post's timestamp and selected the geographically closest one. If no event was found within this time window, we used the nearest event regardless of time. The temporal distance in hours was determined as the time difference between the posting date (t_{post}) and the start (t_{start}) or end time (t_{end}) of the matched event as in Equation 1. This effectively captures whether a post was published before, during or after the event and the respective temporal distance.

$$\text{temporal distance} = \begin{cases} t_{\text{post}} - t_{\text{start}} & \text{if} \quad t_{\text{post}} < t_{\text{start}} \\ t_{\text{post}} - t_{\text{end}}, & \text{if} \quad t_{\text{post}} > t_{\text{end}} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

3.3.2. Co-occurrence with disaster-related posts

Another key feature in our methodology was the estimated density of disaster-related social media posts in the geographic vicinity of each post, which served as a proxy for the regional discourse. Prior research has shown that social media activity surges significantly during and after natural disasters (Resch et al., 2018). Therefore, we hypothesised that a post was more likely to be relevant if it was posted as part of an active discussion surrounding the respective disaster event. To quantify this, we used the preliminary subset of 37,826 posts from our labelling process, which was classified regarding disaster-relatedness using the Disaster-Twitter-XLM-RoBERTa-AL model by Hanny et al. (2024). We specifically computed the number of disaster-related posts within the preceding 7 days in the geospatial neighbourhood for each post. To evaluate different geographic vicinities, we calculated this value within a radius of 1 km, 5 km, 10 km and 50 km and for the total AOI for each event. An algorithmic description of this feature calculation is available in Algorithm 1.

3.3.3. Event type and location encoding

Lastly, incorporating contextual (geospatial) features into machine learning models has been shown to improve prediction performance by enabling the model to distinguish between different environmental and social settings (Lucas et al., 2023). To capture such contextual variations across our use cases, i.e. different events and regions of the world, we incorporated two additional features into our models. First, we encoded the event type (wildfire, flood or earthquake) as three-dimensional one-hot vectors. Second, we

Algorithm 1 Estimate co-occurrence of disaster-related posts

Require: A set of geo-referenced social media posts P
$D \leftarrow \{\}$ ▷ Set that will hold the disaster-related posts
$C \leftarrow$ `dict()` ▷ Dictionary mapping (radius, post) to co-occurrence count
▷ Gather all disaster-related posts
for each post $p \in P$ **do**
  if `relatedness_classfier.predict`(p) == `disaster-related` **then**
    $D \leftarrow D \cup \{p\}$
  end if
end for
▷ Compute co-occurrence counts
for each post $p \in P$ **do**
  $t_p \leftarrow$ timestamp of p
  $\text{loc}_p \leftarrow$ geolocation of p
  for each radius $r \in \{1, 5, 10, 50\}$km **do** ▷ Radius-based co-occurrence
    $C(r, p) \leftarrow |C_r(p) = |\{d \in D : t_d \in [t_p - 7\,\text{days}, t_p] \wedge \text{distance}(\text{loc}_p, \text{loc}_d) \leq r\}|$
  end for
  ▷ AOI-wide co-occurrence
  $C(\text{"AOI"}, p) \leftarrow |C_r(p) = |\{d \in D : t_d \in [t_p - 7\,\text{days}, t_p] \wedge \text{loc}_d \in \text{AOI of } p\}|$
end for
return C

computed the centroid (latitude/longitude) of each AOI and projected it onto a 3D unit sphere to obtain a normalised spatial representation. This three-dimensional location encoding was calculated using Equation 2, following the method described by Banchoff (1990).

$$\begin{aligned} x &= \cos(\text{latitude}) \cdot \cos(\text{longitude}) \\ y &= \cos(\text{latitude}) \cdot \sin(\text{longitude}) \\ z &= \sin(\text{latitude}) \end{aligned} \tag{2}$$

3.3.4. Feature evaluation

To evaluate the quality of our engineered features, we conducted a non-parametric Kruskal-Wallis *H*-test to statistically assess differences in geographic distance, temporal distance and co-occurrence features across our three relevance classes. It tests for the null hypothesis that the population median of all of the groups is equal (Kruskal & Wallis, 1952).

3.4. Classification

After feature engineering, we trained several classification models for our labelled data, which incorporated (1) our spatiotemporal non-text features, (2) the post texts, or (3) both. To realise this, we first created an 80%/20% train-test split of our labelled data, resulting in a training dataset of size 3,659 and a test dataset of size 915. To check for multicollinearity among distance and co-occurrence features, we calculated the VIF. The initial computation revealed that our co-occurrence feature at a 5 km radius had a VIF of 8.79, indicating high multicollinearity. To address this, we removed it from the dataset. After this adjustment, the highest VIF value was 3.47, indicating low multicollinearity, so all remaining 12 features were retained.

3.4.1. Non-text classifier

To assess the predictive power of our non-text features, we first defined a 7-dimensional base feature set consisting of geographic distance, temporal distance and the five co-occurrence features. We used all features at once to maximise predictive power as no strong redundancy was indicated by the VIF. We then evaluated four different feature configurations by concatenating this base representation with additional contextual information: (1) the 3-dimensional one-hot encoding of the event type, (2) the 3-dimensional location encoding from the event's centroid, (3) both the event type and location encoding and (4) the base features alone, without any additional contextual information.

The concatenated feature sets from these four configurations were used to train and evaluate six machine learning classifiers. Specifically, we considered logistic regression (Cox, 1958), a random forest (Ho, 1995), a SVM (Cortes & Vapnik, 1995), a gradient boosting decision tree model (Friedman, 2001), a kNN classifier (Fix & Hodges, 1989) and a Gaussian naive Bayes classifier (Lewis, 1998). These models were selected to represent a diverse set of learning paradigms, including linear models (logistic regression, SVM), tree-based ensemble models (random forest, gradient boosting), instance-based learning (kNN) and probabilistic classification (naive Bayes). For each model, we performed hyperparameter tuning via grid search over the parameter space defined in Appendix B, using 5-fold cross-validation on the training data (Agrawal, 2021). The best-performing configuration, based on the highest validation macro F1 score, was then used to train and evaluate each model on the test data.

3.4.2. Text classifier

Most previous research in relevance classification of social media data (e.g., Blomeier et al. (2024); Kaufhold et al. (2020)) focused primarily on text-based classification. Since we aim to combine text with contextual non-text features in our study, we also evaluated the performance of a purely text-based classification model on our data. To realise this, we utilised TwHIN-BERT-base, a multilingual language model that was pre-trained on 7 billion tweets from over 100 languages (Zhang et al., 2023). It has demonstrated superior performance to other specialised language models for social media data (e.g., Barbieri et al. (2022)) across several downstream tasks like text classification, hashtag prediction or social engagement prediction. We fine-tuned the model with our training data, using 20% of it for hyperparameter validation. The optimal hyperparameters were obtained through 20 rounds of optimisation with a TPE (Watanabe, 2023) and the search space detailed in Appendix B (based on Devlin et al. (2019)). TPE was chosen for its ability to efficiently explore high-dimensional and conditional hyperparameter spaces. Unlike grid or random search, TPE builds a probabilistic model of *good* and *bad* hyperparameter configurations, allowing for more informed sampling and faster convergence to optimal settings. This makes it particularly suitable for fine-tuning LLM where each training run is computationally expensive. After training, we selected the model with the highest validation macro F1 score across all epochs for final evaluation. The optimal parameter configuration identified throughout this process was a learning rate of 4.66×10^{-5}, batch size 16, a weight decay of 0.065 and 5 training epochs.

3.4.3. Multimodal classifier

For merging text with non-text features, we explored four different strategies: (1) Feature concatenation, where the 12-dimensional non-text feature vector was appended to the internal text representation of the BERT model before classification; (2) In-context learning, in which the non-text features were embedded directly into the input text as a structured prompt; (3) Stacking, where separate classifiers were trained on the text and non-text features, and their output probabilities were combined using a meta learner; and (4) Partial stacking, where we used the softmax output of the text classifier as an additional input feature alongside the non-text features, which were then passed to a meta learner. Based on the results in Section 3.4.1, which indicated a slight performance advantage of the 12-dimensional feature set over reduced variants, all multimodal classifiers were trained using the full set of non-text features in combination with text.

3.4.3.1. Concatenation. As a first strategy for combining textual and non-textual features, we extended the standard BERT architecture by appending numerical non-text features directly to the model's internal representation, followed by a classification head. A similar method was successfully applied by Scheele et al. (2021) within a CNN-based architecture, leading to an improvement of up to 6% points in macro F1 score compared to a text-only baseline. In our setup, we used TwHIN-BERT-base (Zhang et al., 2023) as the backbone model (see justification in Section 3.4.2). We extracted the pooled output from the BERT encoder (a 768-dimensional vector) and concatenated it with the 12-dimensional non-text feature vector. This combined representation was then passed to one of two classification head configurations:

(1) A **simple** head, where the concatenated 780-dimensional feature vector is passed directly to a linear output layer for classification.
(2) A **complex** head, which processes the non-text features separately using a feed-forward subnetwork consisting of two fully connected layers with 32 hidden units and ReLU activation. The output of this subnetwork is then concatenated with the pooled BERT output, resulting in an 800-dimensional vector. This vector is passed through a classifier composed of a fully connected layer with 256 hidden units, a ReLU activation, dropout ($p = 0.1$), and a final output layer projecting to the number of class labels.

For both variants, 20% of the training data was used for hyperparameter optimisation. We employed a TPE over 20 rounds, using the search space defined in Appendix B for the same reason as detailed in Section 3.4.2.

3.4.3.2. In-context learning. Brown et al. (2020) have shown that LLM can effectively learn from contextual information presented in the input text. Since then, few-shot prompting of generative language models has become a powerful technique for few-shot classification. However, in-context learning approaches have not been evaluated much for encoder-based models. Recent work by Samuel (2024) demonstrated that encoder-only models like BERT are also capable of performing in-context learning. Inspired by these findings, we explored in-context learning as an alternative, model-agnostic feature integration method by embedding the non-text features directly into

the input text. To provide a structured and interpretable format for this integration, we adopted a JavaScript Object Notation (JSON) schema, as illustrated in Figure 6. While no prior work has explicitly examined the impact of JSON structuring for in-context learning, we found it to slightly improve model performance while also being more readable for humans. Based on this enriched input data, we then fine-tuned a TwHIN-BERT-base (Zhang et al., 2023) model, using 20% of the training data and a TPE for hyperparameter optimisation, as in the previous approaches.

3.4.3.3. (Partial) stacking. Another widely used approach for learning from heterogeneous data is stacked generalisation or stacking—an ensemble learning approach. In this paradigm, several base models (referred to as level-0 learners) are first trained independently on heterogeneous data. Their predictions are then used to construct a new dataset, which serves as input for a meta-learner (or level-1 learner). The meta-learner is trained to learn an optimal combination of the base models' outputs, effectively assigning weights or rules for how much to rely on each base model's prediction. To prevent information leakage and ensure unbiased training of the meta-learner, base models are typically trained using *k*-fold cross-validation, where out-of-fold predictions for each training instance are collected and used to train the meta-learner (Wolpert, 1992).

To implement the stacked generalisation approach, we generated out-of-fold predictions using 5-fold cross-validation on the training data. For each fold split (4 training folds and 1 validation fold), we fine-tuned a TwHIN-BERT-base model (Zhang et al., 2023) for text classification using the same hyperparameter configuration described in Section 3.4.2. In parallel, we trained a random forest classifier on the full set of non-text features, following the setup in Section 3.4.1 To ensure consistency, we used the optimal hyperparameters identified in that section, which included 100 estimators, no maximum tree depth, and a minimum of 5 samples required to split an internal node.

Both the text and non-text models were then used for out-of-fold softmax probabilities predictions, which served as inputs to a meta-learner in two configurations:

```
{
  "text": "I've just seen the pictures of #Esch Eifel. It's crazy how much driftwood has
collected there. #Flood.",
  "Event Type": "flood",
  "Distance from Disaster (km)": 56.131,
  "Time Gap from Disaster (hours)": 131.861,
  "Disaster Posts within 1 km": 0,
  "Disaster Posts within 10 km": 17,
  "Disaster Posts within 50 km": 59,
  "Disaster Posts in Area of Interest": 104,
  "Central Latitude": 7.12,
  "Central Longitude": 50.56
}
```

Figure 6. Exemplary input when training TwHIN-BERT-base with in-context learning. For visualisation purposes, we translated the post to the English language.

(1) A **full stacking** setup, where we concatenated the softmax probabilities from both the text and non-text models (3 + 3 dimensions) as input to the meta-learner.
(2) A **partial stacking** setup, where we combined the softmax probabilities of the text model (3 dimensions) with the 12-dimensional non-text feature vectors, resulting in a 15-dimensional input.

For each stacking configuration, we evaluated the same six classifiers as in Section 4.2.1 as meta-learners. Hyperparameter optimisation was conducted analogously using 5-fold cross-validation on the training data, following a grid search over the parameter space defined in Appendix B.

3.5. Model evaluation

For the evaluation of our classification approaches, we computed standard classification metrics for our test data for each model and feature configuration. This includes the macro-averaged F1 score, accuracy and the average ROC-AUC score using a one-vs-rest strategy for multi-class classification. In addition to these conventional classification metrics, we also considered the Root Mean Squared Error (RMSE) between the predicted and true class labels, treating them as ordinal values encoded as integers (0, 1, 2). This offers additional insight into the severity of misclassifications by quantifying how far off the model's predictions were from the true labels (Gaudette & Japkowicz, 2009).

To furthermore gain a deeper understanding of the decision-making process in our non-text models, we employed SHAP (Lundberg & Lee, 2017), a game-theoretic approach for interpreting machine learning predictions. SHAP assigns an importance value to each input feature based on its contribution to the model output, as determined by Shapley values (Shapley, 1953). Originally developed to distribute gains fairly among players in cooperative games, Shapley values in machine learning represent the average marginal contribution of each feature across all possible feature subsets. We applied SHAP to analyse feature importance and contributions for the non-text classifier, which yielded the highest macro F1 score described in Section 3.4.1. This analysis enabled us to identify which contextual features most strongly influenced model predictions and to assess how effectively the models leveraged spatiotemporal information.

4. Results

Our results consist of three main parts: First, we analyse the properties of the engineered contextual non-text features to assess their potential for relevance classification. Second, we present the results of our model evaluation across the different input modalities. Third, we examine the three best-performing models based on non-text features, text features and both in more detail.

4.1. Feature engineering

The Kruskal-Wallis H test revealed statistically significant differences ($p < 0.05$) across most distance and co-occurrence features for all use cases, with two exceptions: geographic distance for the California wildfires ($p = 0.09$) and temporal distance for

the Chile wildfires ($p = 0.56$). These findings are reflected in the distribution of feature values shown in Figure 7. For the California wildfires, the mean geographic distance to the disaster event varied only slightly across relevance classes (5.7 to 6.9 km), which aligns with the non-significant test result. In contrast, the Chile wildfires showed greater differences in geographic distance (8.2 to 16.7 km) but no significant variation in temporal distance. For the Turkey earthquake, a clear gradient was observable in geographic distance to the epicentre. Posts labelled "Not related" were furthest from the event (270 km), followed by "Related but not relevant" (263 km), while "Related and relevant" posts were considerably closer (35 km). The Ahr Valley floods showed a similar pattern. "Related and relevant" posts were closest to the affected area (8.3 km) and also had the shortest average time difference to the disaster (27 hours). These were followed by the "Related but not relevant" class, and then the "Not related" class, which had the highest average distances in both dimensions. A similar trend was found for the Emilia-Romagna floods, where "Related and relevant" posts were also geographically closest to the event (3.7 km), clearly separating them from the other categories.

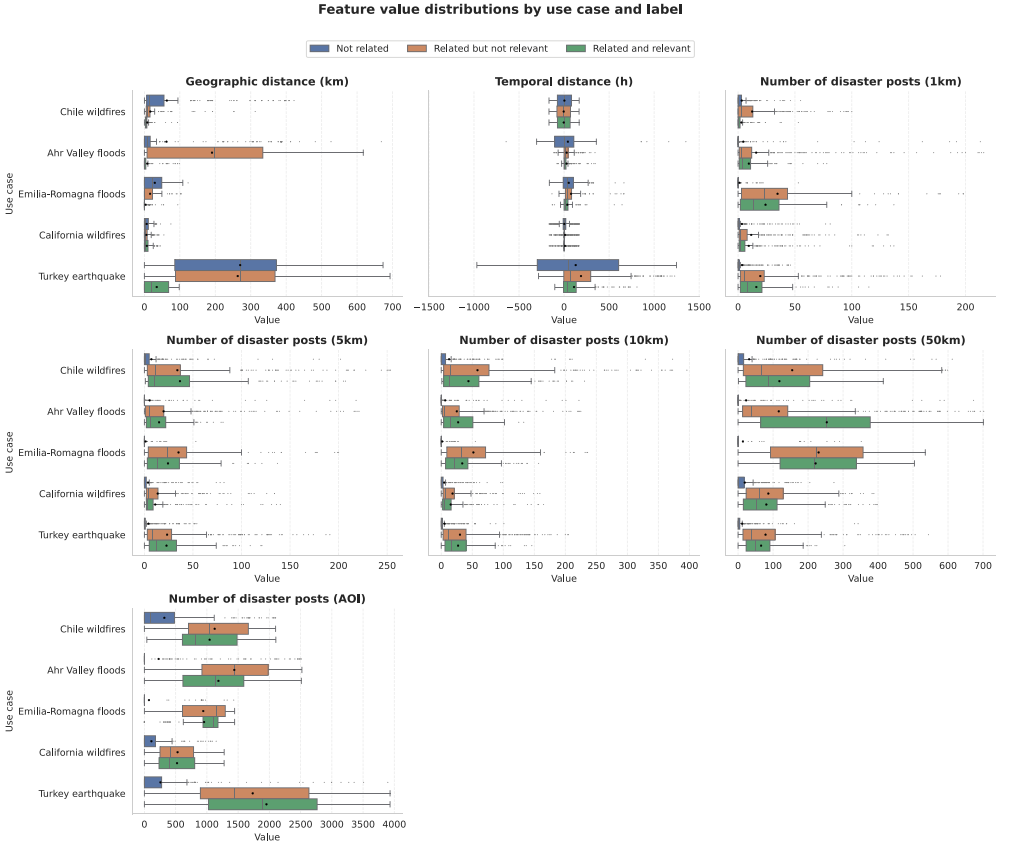

Figure 7. Distribution of feature values for each of our five use cases per human-labelled class, depicted using boxplot diagrams. The black dots denote the mean value and the gray dots depict outliers.

Across all events, the number of co-occurring disaster-related posts generally increased with relevance. For example, in the case of the California wildfires, the average number of disaster-related posts within 1 km rose from 3.4 in the "Not related" class to 11.6 and 9.4 in the "Related" classes. Similar trends were observed at larger spatial scales, where both "Related" classes consistently showed markedly higher co-occurrence counts than "Not related" posts. In some cases, the "Related but not relevant" class had slightly higher co-occurrence values than the "Related and relevant" class, though this was not consistent across all use cases.

Notably, the feature value distributions displayed a considerable number of outliers across all relevance classes and use cases. This effect was particularly evident in the Chile wildfires and Ahr Valley floods datasets for geographic and temporal distance to the disaster, and across all use cases for the co-occurrence features. These outliers can be partly attributed to limitations in data quality. Since 2019, Twitter (now X) has removed the option for precise geo-tagging, replacing it with coarser place-based tags that often correspond to large geographic areas such as cities or regions (Khalid, 2019). As a result, spatial references may be imprecise, and temporal references may also vary depending on user behaviour. Overall, the results indicate that posts in the "Not related" class were typically both geographically and temporally more distant from the disaster events and were surrounded by fewer nearby disaster-related posts. In contrast, posts labelled as "Related and relevant" or "Related but not relevant" were generally closer in time and space to the disaster and exhibited significantly higher levels of local disaster-related activity, supporting the utility of these features for relevance classification.

4.2. Classification

4.2.1. Non-text classifier

Table 5 compares the average cross-validation macro F1 scores across all evaluated models for different combinations of non-text features. The best-performing configuration combined all features: distance and co-occurrence-based features, event encoding and location encoding, yielding an average macro F1 score of 0.635. This configuration was therefore selected for use in all further experiments. The inclusion of contextual information, such as event type and geographic location, provided a slight performance advantage over using only distance and co-occurrence features. This finding aligns with prior research showing that social media activity during disasters varies by region, culture, and event characteristics (Lin et al., 2016; Sarmiento & Poblete, 2021).

Table 6 presents the evaluation metrics for all non-text classifiers using the 12-dimensional feature vector. This set includes all geographic and temporal distance features as well as co-occurrence features, excluding the 5 km radius feature due to high multicollinearity. The best-performing models were tree-based. The random forest achieved the highest macro F1 score on the test set (0.713), the highest accuracy (0.731), the best cross-validation macro F1 (0.704), a strong ROC-AUC of 0.887 and a low RMSE (0.627). Gradient boosting followed closely with a macro F1 of 0.710, tied for the highest accuracy (0.731) and scored the best ROC-AUC (0.891). The kNN classifier ranked third, with a macro F1 of 0.675 and an accuracy of 0.701. Among the remaining models, the SVM showed moderate performance (macro F1 0.573) but the lowest RMSE (0.615). Logistic

Table 5. Average cross-validation (CV) macro F1 score across all models for the evaluated feature configurations. The base features are our geographic/temporal distance and the co-occurrence features (excluding the feature computed at a 5 km radius due to high multicollinearity). Event encoding is a 3-dimensional one-hot vector for the disaster type. Location encoding is a 3-dimensional vector resulting from projecting latitude/longitude coordinates onto a unit sphere. The upward arrow (↑) indicates that higher values are better.

Base features	Event encoding	Location encoding	CV macro F1 ↑
✓	✗	✗	0.620
✓	✓	✗	0.630
✓	✗	✓	0.633
✓	✓	✓	**0.635**

Table 6. Evaluation metrics for our non-text classifiers with the 12-dimensional non-text feature vector as input. CV macro F1 is the average macro F1 score achieved during cross-validation. All other scores were computed on the unseen test data. The upward arrow (↑) indicates that higher values are better. The downward arrow (↓) indicates that higher values are better. The best values are marked in bold.

Model	CV macro F1 ↑	Macro F1 ↑	Accuracy ↑	ROC-AUC ↑	RMSE ↓
Logistic regression	0.622	0.624	0.674	0.861	0.707
Naive Bayes	0.562	0.601	0.622	0.811	0.777
kNN	0.659	0.675	0.701	0.857	0.664
Gradient boosting	0.695	0.710	**0.731**	**0.891**	0.640
Random forest	**0.704**	**0.713**	**0.731**	0.887	0.627
SVM	0.566	0.573	0.681	0.855	**0.615**

regression and naive Bayes performed significantly less effectively, particularly the latter, which yielded the worst scores across all metrics.

Next, we selected the random forest for feature interpretation with SHAP because it achieved the highest cross-validation and test macro F1 score. Figure 8 shows the associated SHAP beeswarm plots. The SHAP values indicate the extent to which each feature influenced the model's predictions, with positive or negative values reflecting a feature's contribution toward or against a given class. Temporal and geographic distance emerged as particularly important for distinguishing the "Related and relevant" class. Low values for both features had a strong positive impact, meaning the model was more likely to classify a tweet as relevant when it was posted close in time and space to the disaster event. In contrast, predictions for the "Related but not relevant" class were most impacted by greater geographic distance, suggesting that these posts may be contextually connected to the disaster but originate from locations farther away, making them less relevant in the model's view. Co-occurrence features, especially the number of disaster-related posts within the AOI, 50 km, and 10 km radii, were consistently among the most influential factors across all three classes. High values for these features tended to increase the likelihood of a tweet being classified as "Related and relevant" or "Related but not relevant". Conversely, low co-occurrence values had a strong influence on the "Not related" class, indicating that isolated posts not surrounded by other disaster-related content were more likely to be considered as unrelated. Finally, event type encodings and spherical coordinates contributed less overall but still played a minor role in the model's predictions when combined with the other features.

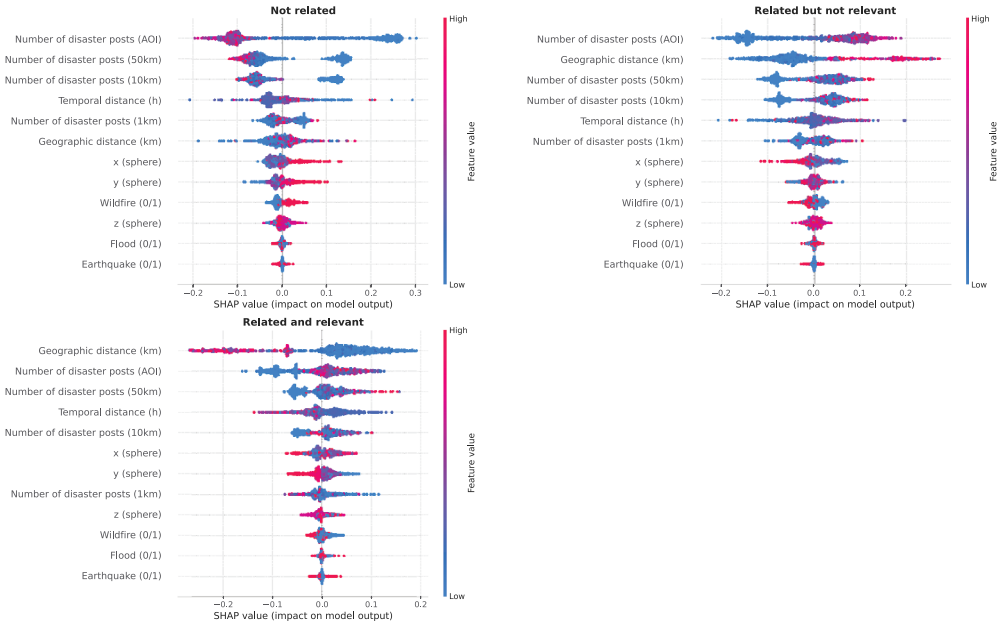

Figure 8. SHAP beeswarm plot for each of the three relevance classes for the random forest non-text model. It shows the impact of individual features on model output for each class.

4.2.2. Text classifier

The TwHIN-BERT-base model fine-tuned on only the post texts achieved a macro F1 score of 0.779, an accuracy of 0.802, a high ROC-AUC of 0.928, and a low RMSE of 0.516 on the test set. These results show a notable advantage over the non-text models presented in the previous section. This confirms that textual content is of particularly high importance for relevance classification. The findings are consistent with prior work by Kaufhold et al. (2020), which also identified text as the strongest predictor in crisis-related relevance assessments.

4.2.3. Multimodal classifier

Table 7 presents the evaluation results for all multimodal classifier configurations that combine textual with non-textual information. Among the tested strategies, the partial stacking approach achieved the strongest overall performance. In this setup, the gradient boosting meta-learner yielded the highest macro F1 score (0.814), the best accuracy (0.832), the lowest RMSE (0.484), and a strong ROC-AUC of 0.927. The random forest meta-learner performed similarly, also reaching an accuracy of 0.832 and achieving a macro F1 of 0.813. The full stacking approach, which combines probability outputs from independent text and non-text classifiers, also produced competitive results. Here, logistic regression achieved a macro F1 of 0.811 and the same top accuracy of 0.832, with an RMSE of 0.487. The stacking approaches therefore consistently outperformed both the text-only and non-text baselines, which achieved macro F1 scores of 0.779 and 0.713, respectively.

In-context learning and feature concatenation performed slightly below the best ensemble models. In-context learning only achieved a macro F1 of 0.781, while the

Table 7. Evaluation metrics for different multimodal classifier configurations. The best values are marked in bold, and the best scores within each category are underlined. Upward arrows (↑) indicate that higher values are better, the downward arrow (↓) indicates that higher values are better. The best non-text model and the fine-tuned text classifier are shown as baselines.

Method	Variant	Macro F1 ↑	Accuracy ↑	ROC-AUC ↑	RMSE ↓
Concatenation	Simple head	0.807	0.822	0.926	0.516
	Complex head	0.812	0.825	**0.940**	0.523
In-context learning	TwHIN-BERT-base	0.781	0.799	0.923	0.513
Probabilities (Stacking)	Logistic regression	0.811	**0.832**	0.938	0.487
	Naive Bayes	0.797	0.817	0.932	0.505
	kNN	0.790	0.813	0.916	0.512
	Gradient boosting	0.699	0.727	0.905	0.718
	Random forest	0.793	0.810	0.921	0.509
	SVM	0.800	0.824	0.910	0.495
Partial (Stacking)	Logistic regression	0.800	0.822	0.931	0.494
	Naive Bayes	0.762	0.778	0.910	0.580
	kNN	0.676	0.702	0.858	0.663
	Gradient boosting	**0.814**	**0.832**	0.927	**0.484**
	Random forest	0.813	**0.832**	0.932	0.487
	SVM	0.575	0.683	0.855	0.613
Non-text	Random forest	0.713	0.731	0.887	0.627
Text-only	TwHIN-BERT-base	0.779	0.802	0.928	0.516

concatenation approach yielded scores of 0.812 (simple head) and 0.807 (complex head). Feature concatenation within a neural network can therefore also be a powerful strategy for combining text with non-text features. This goes in line with the findings of Scheele et al. (2021) who achieved similar results with a CNN-based architecture.

Compared to the unimodal baselines, all multimodal classifier configurations substantially outperformed the best non-text model (macro F1: 0.713, accuracy: 0.731). Improvements over the text-only classifier (macro F1: 0.779, accuracy: 0.802) were more moderate but also consistent. The stacking approaches delivered the most reliable performance gains over the text-only model across all evaluation metrics, while concatenation also offered competitive performance. In-context learning yielded no meaningful improvement over the text-only baseline. These findings suggest that while the textual modality remains the strongest single predictor for relevance, spatiotemporal non-text features can significantly enhance classification performance when integrated with suitable strategies.

4.3. Comparative evaluation

To assess the added value of combining text with non-text modalities, we conducted a more detailed comparative analysis of the best-performing models across three classification approaches: (1) the non-text random forest model using our 12-dimensional feature vector, (2) the text-only TwHIN-BERT-base model, and (3) the best-performing multimodal model, which is based on partial stacking with a gradient boosting meta-learner. Specifically, we analysed model performance per use case, per relevance class and for binary classification tasks.

4.3.1. Performance per use case

As shown in Figure 9, the multimodal approach notably outperformed both unimodal classifiers in three out of five use cases. The largest overall improvement by combining text with non-text features was observed for the Ahr Valley floods, where the multimodal model achieved a macro F1 of 0.855, compared to 0.801 for the text-only and 0.748 for the non-text model. Similarly, for the Chile wildfires (0.800) and Turkey earthquake (0.833), the multimodal classifier also achieved the highest macro F1 scores, outperforming both the text and non-text alternatives. For the California wildfires, performance between the multimodal approach (0.722) and text classifier (0.724) was nearly identical, though both substantially outperformed the non-text model (0.526). The only case where the text model (0.870) outperformed the multimodal model (0.850) was the Emilia-Romagna floods. This may be attributed to the extensive pre-filtering of the dataset, which likely reduced the added value of spatial and temporal features. Moreover, the Emilia-Romagna dataset was relatively small, with only 542 posts remaining after pre-filtering, of which 100 were used for testing. For the Turkey earthquake, the non-text model achieved a higher macro F1 score (0.825) than the text classifier (0.713) and was only outperformed by the multimodal model. In this case, the dataset had been filtered only coarsely and the posts covered a broad geographic area across Turkey and Syria. These conditions appear to have increased the utility of contextual spatiotemporal features, enabling more effective predictions based on non-textual information. In contrast, for the California wildfires, spatiotemporal features added little value. This was likely caused by the widespread nature of the fires across the state of California (Wallin, 2025), which resulted in generally low and homogeneous distances between posts and event locations (as also evident in Section 4.1).

Figure 10 presents a spatial view of model performance for our multimodal classifier. It visualises whether a post was classified in line with our human annotation, considering semantics, space, and time. In the case of the 2023 Emilia-Romagna floods, almost all posts in close proximity to the flooded areas were correctly classified. For the 2023 Chile wildfires, it was noticeable that posts in the unaffected capital Santiago de Chile were also correctly classified. For the 2023 Turkey earthquake, we found that posts far away from the epicentre were classified correctly, with a few exceptions. However, there was a noticeable cluster of

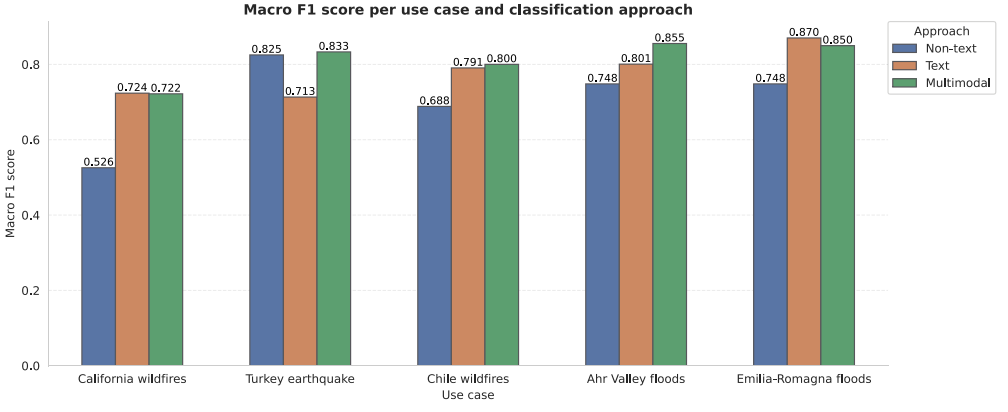

Figure 9. Macro F1 score for each classification approach across our five use case scenarios.

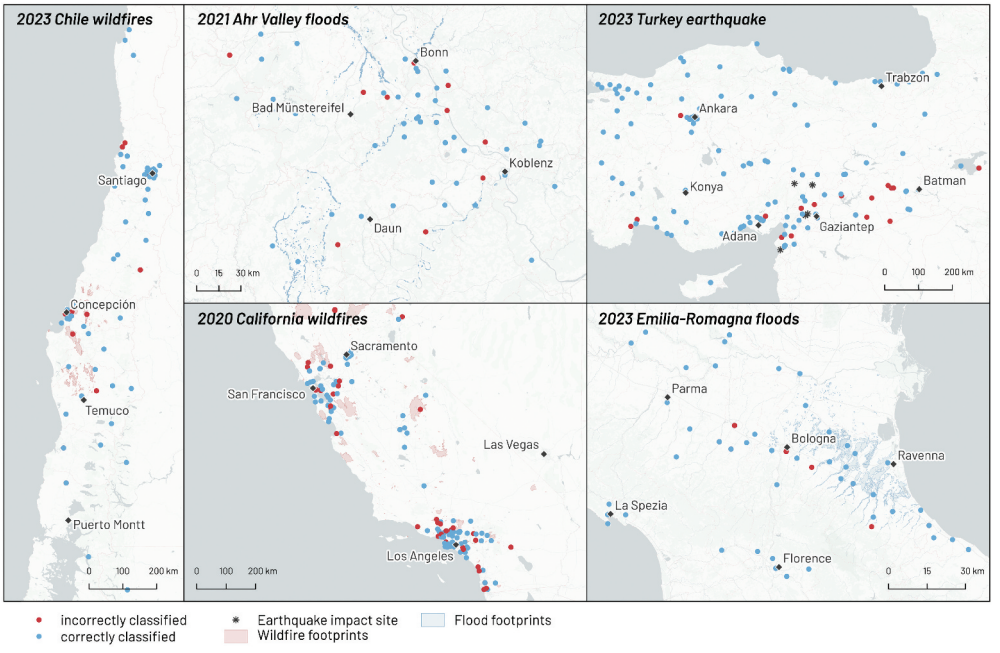

Figure 10. Evaluation of our multimodal predictions from a spatial perspective. The colour scheme indicates the agreement between our human annotation and the model results.

incorrectly classified posts south-west of Batman. No clear spatial pattern was recognisable for the 2020 California wildfires, where correctly and incorrectly classified posts were generally located in spatial vicinity. Figures D1 and D2 in the appendix visualise the same spatial view for prediction based on solely text and non-text features, respectively. While the results for text-based and multimodal classification were very similar, the predictions based just on our non-text features were more dispersed, especially for the California and Turkey use cases.

4.3.2. Performance per relevance class

Figure 11 shows the F1 scores for each relevance class. The multimodal classifier consistently outperformed the other models across all three classes. The most substantial improvement was observed for the "Related and relevant" class, where the multimodal model achieved an F1 score of 0.712, compared to 0.638 for the text-only model and 0.595 for the non-text model. Similarly, for the "Related but not relevant" class, the multimodal classifier reached an F1 score of 0.837, while the text-only model only achieved a value of 0.808 and the non-text model 0.739. The overall highest F1 scores were observed for the "Not related" class across all models, with the multimodal model scoring 0.894, the text-only 0.890, and the non-text model 0.804. These results indicate that posts unrelated to any natural disaster were the easiest to classify, while "Related and relevant" posts were more difficult to distinguish, particularly when relying on non-text features alone.

4.3.3. Performance for binary classification tasks

Based on our hierarchical class scheme, we further evaluated the performance of each classification approach on two binary classification tasks: disaster-relatedness and

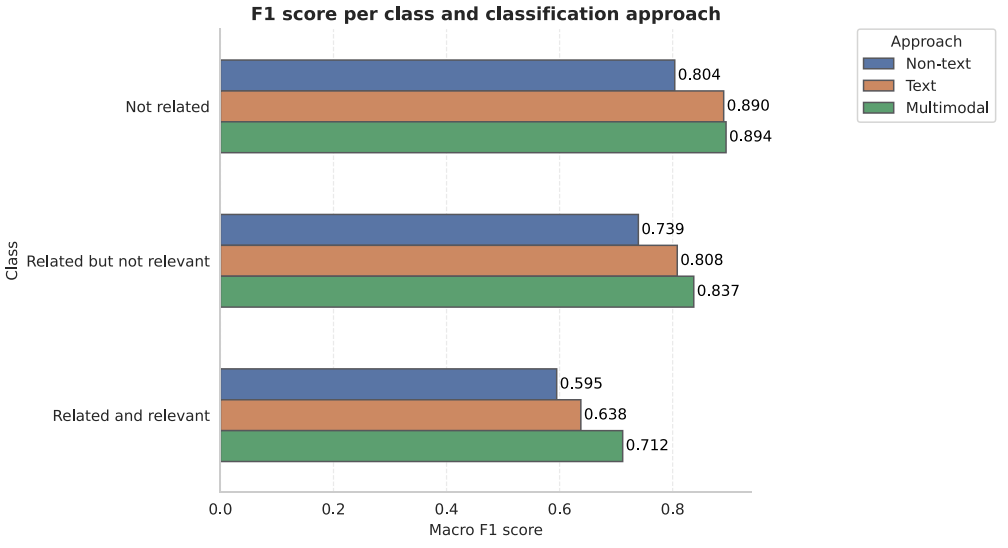

Figure 11. F1 score for each classification approach across the three relevance classes.

relevance/irrelevance. For disaster-relatedness, the classes "Related but not relevant" and "Related and relevant" were grouped as class 1, while "Not related" was assigned to class 0. For binary relevance classification, only "Related and relevant" was considered class 1, with the other two classes grouped as class 0. The results are presented in Table 8. Across both tasks, the multimodal model combining text and non-text features achieved the highest macro F1 scores, with 0.919 for disaster-relatedness and 0.819 for relevance. The text-only model followed closely, reaching 0.916 and 0.773, respectively. The non-text model showed lower performance in both cases, with macro F1 scores of 0.849 for disaster-relatedness and 0.743 for relevance. Notably, the greatest improvement from combining modalities was observed in the binary relevance classification task, where the multimodal model outperformed the text-only model by a difference of 0.046 in macro F1.

5. Discussion

Our discussion is split into two parts: First, we discuss the results of Section 4 and subsequently, we debate the limitations of our methodology.

Table 8. Evaluation metrics for binary classification of disaster-relatedness and relevance. The upward arrow (↑) indicates that higher values are better. The best values are marked in bold.

	Disaster-relatedness		Binary Relevance	
Method	Macro F1 ↑	Accuracy ↑	Macro F1 ↑	Accuracy ↑
Non-text	0.849	0.862	0.743	0.827
Text	0.916	0.925	0.773	0.855
Multimodal	**0.919**	**0.927**	**0.819**	**0.883**

5.1. Discussion of the results

Overall, the benefit of integrating text with non-text features was most pronounced for the Ahr Valley floods and Turkey earthquake datasets (macro F1 0.855 and 0.833), where the input datasets covered comparatively large spatial and temporal extents (c.f. Table 2). In these cases, proximity and co-occurrence features provided strong complementary signals. In contrast, for the Emilia-Romagna floods, where spatial and temporal pre-filtering was more granular, the inclusion of non-text features slightly reduced performance (macro F1 0.850). This suggests that non-text features are most useful when the input data spans broader geographic or temporal scopes. Across four out of five use cases and for the combined test set, multimodal models consistently outperformed or matched the performance of unimodal models. This confirms the value of multimodal fusion in heterogeneous disaster contexts. However, spatial evaluations of classification outcomes revealed that the best-performing multimodal classifier (partial stacking with gradient boosting) and the text-only classifier produced highly similar spatial patterns of correct and incorrect predictions. This indicates that the semantic content of the posts remained the dominant factor in determining classification accuracy, even in multimodal configurations.

To better understand the impact of non-text features on predictions, we conducted a SHAP analysis on the best-performing non-text classifier. This revealed that the model learned to rely heavily on spatiotemporal proximity and co-occurrence features, associating low geographic and temporal distances as well as high local post density with relevance. However, the SHAP value distributions showed that these relationships were not always clear-cut. For many features, high and low values appeared on both sides of the impact axis, suggesting that the model captured non-linear and context-dependent interactions rather than applying simple thresholds. This complexity reflects the nuanced nature of relevance in disaster-related posts, where proximity can be helpful but not definitive.

Prediction quality differed notably across disaster types. Text-based classification performed best for the flood datasets (macro F1 0.870 and 0.801), but worst for the Turkey earthquake (0.724). In contrast, non-text features achieved their highest performance on the Turkey earthquake (macro F1 0.825), likely because the event was spatially and temporally concentrated, while posts were more dispersed. The California wildfires were the most challenging use case overall. Non-text classification yielded the lowest macro F1 score (0.526), and both text-based (0.724) and multimodal models (0.722) only achieved intermediate performance, despite this being the only English-language dataset. This suggests that language was not the primary limiting factor. English-language posts were generally classified with high accuracy across all models. Our multimodal approach (partial stacking with gradient boosting) achieved perfect accuracy (1.00) for English posts in Chile and Turkey, and high accuracy in Germany (0.93), Italy (0.75), and California (0.71). Notably, non-text classification (random forest) also performed strongly for English posts outside of California, reaching 1.00 in Chile and Italy, 0.89 in Germany, and 0.88 in Turkey. This suggests that English-language content often aligned with the location and timing of the respective disasters. For non-English national languages, the text-based and multimodal classification methods also achieved high accuracy, typically above 0.85. Multimodal classification accuracy was 0.85 for Spanish in Chile, 0.87 for

German in Germany, 0.86 for Italian in Italy, and also 0.86 for Turkish in Turkey. Text-only classification yielded similar results, while non-text classification showed substantial performance drops for these languages (−0.03 to −0.14 compared to multimodal), indicating that relevant content in local languages was harder to classify without textual information. This may be caused by higher semantic diversity in non-English posts, whereas English content tended to be more direct or event-specific. Finally, the comparatively low performance across all models for the California wildfires appears to be driven by event characteristics. The wildfires were geographically widespread, meaning most posts were close to some affected area, which reduced the discriminative power of spatial and temporal proximity features. Moreover, distinguishing wildfire-related posts from unrelated house fire content proved challenging during annotation, introducing additional ambiguity in the textual content and potentially confusing the classifiers.

Compared to the four-class scheme of Blomeier et al. (2024), our simplified three-class scheme achieved significantly better results (macro F1 0.855 versus 0.715) for the same Ahr Valley floods use case. For real-world scenarios, where fast decisions are needed, this supports the use of a simplified scheme. Other studies (e.g., Kaufhold et al. (2020)) are not directly comparable to ours due to different definitions of relevance, varying use cases and other filtering methods. Regardless, our results suggest a step forward in relevance classification for disaster-related social media content. Our work is among the first to combine transformer-based models specialised for social media (TwHIN-BERT) with engineered contextual features, yielding consistent performance improvements.

In a real-world data stream, the vast majority of posts within the AOI would be unrelated to the disaster. Our evaluation, however, is slightly biased due to the fact that the test dataset was somewhat balanced, with 36% of posts labelled as "Not related", 42% as "Related but not relevant", and 22% as "Related and relevant". We therefore expect our approach to perform even better in real-world deployment, as posts unrelated to any disaster could be identified best (macro F1 up to 0.895). If required, the performance can be improved even further by binarising the class scheme as explained in Section 4.3.3

Lastly, our results highlight the advantages of tree-based models for non-text feature classification and ensemble approaches for multimodal classification. Random forest and gradient boosting consistently outperformed linear models and SVMs in the non-text setting, likely due to their ability to capture non-linear dependencies between distance and co-occurrence features. For the multimodal classification task, ensemble-based stacking yielded the best overall performance, suggesting that treating text and non-text features as complementary modalities and combining them at the decision level is an effective strategy for relevance classification.

5.2. Discussion of the methodology

Even though many considerations were made, our methodology still encounters some limitations. First, the quality of geo-references in our social media data remains uncertain. Since Twitter (now X) disabled precise geotagging in 2019 (Khalid, 2019), most posts include coarse location tags represented by bounding boxes instead of point coordinates. These vary in size from a few metres to several kilometres in diameter. What also remains unclear is whether users were physically present at the tagged location, as the georeference must be attached manually to each post (Serere & Resch, 2025). This spatial

imprecision may propagate through the feature engineering process, potentially affecting proximity-based features such as distance and co-occurrence. Furthermore, social media content can include false or misleading textual information, which complicates disaster-related analysis (Buntain & Golbeck, 2017). Future work should therefore address content plausibility more explicitly.

Second, the human annotation process posed significant challenges, even with our consensus-based approach. One particularly difficult case involved posts referring to past disasters that were similar to the event of interest. While such posts may help in disaster management by offering historical context or raising general awareness, they typically lack actionable or event-specific information relevant to the disaster under study. We therefore classified them as "Related but not relevant". Moreover, we encountered posts mentioning other types of natural disasters like smaller earthquakes in Chile. To reduce ambiguity and maintain consistency in the data, we excluded these posts from our final sample, though we acknowledge that real-world systems must learn to handle such content. Labelling decisions involving geographic space and time also proved complex, as relevance often depends on the interplay between event location, time and content. To address this, we relied on criteria from prior literature and expert guidance. We suggest that future research should investigate the multimodal definition of "relevance" in disaster-related social media content more thoroughly, ideally combining qualitative perspectives with quantitative evidence.

The set of non-text features introduced in Section 3.3 is furthermore not exhaustive. While it is possible to engineer additional features using links, likes or reposts, prior studies have shown limited added value of such interaction metrics compared to purely text (Buntain & Golbeck, 2017). However, follow-up work could explore the integration of local weather and climate information. Visual content, such as images or videos, was also beyond the scope of this study but represents a promising direction for integration in future work (Koshy & Elango, 2023a). Additionally, our co-occurrence features were calculated on a sample, not a full data stream, due to limited computing resources. Despite this, they still served as reliable indicators of local disaster-related activity. It is also worth noting that our distance-based features were derived from EO data, which may not always be available in real time, posing potential limitations for rapid response applications.

We also explored the effects of different normalisation strategies for our non-text features, including global and per-use-case z-score normalisation. However, this consistently resulted in equivalent or worse performance across all non-text and multimodal classification approaches and partly limited the generalisability of our data. One possible explanation for this is that tree-based models and neural networks are inherently capable of handling heterogeneous features of different scales (Singh & Singh, 2020), making normalisation unnecessary or even counterproductive. We therefore abstained from normalisation in our final methodology.

Our study is additionally limited from a model perspective. For instance, we excluded neural networks from the non-text classification and stacking approaches. Although we initially tested several small CNN and MLP variants, their performance was consistently worse than simpler models like random forests and gradient boosting, making them unsuitable given their added complexity. For merging text with non-text features, we furthermore investigated concatenating the 768-dimensional BERT representation of the

text directly with our non-text features and passing the resulting vector to a tree-based classification model. However, this approach consistently yielded macro F1 scores below the text-only baseline. We also explored alternative ensemble learning techniques like bagging with a weighted aggregation of softmax probabilities from a text and non-text classifier. However, these were consistently outperformed by stacking with a meta-learner on both the validation and test data. Similar effects have been observed in previous studies (Džeroski & Ženko, 2004). Future research should also investigate other types of information fusion approaches to solve the multimodal classification task at hand, e.g., through graph-based approaches or cross-modal attention. While we considered these techniques during the initial conception of this work, their effective implementation introduces substantial methodological complexity, such as how to encode spatial and temporal information within the attention mechanism, or how to construct graph structures that accurately capture multimodal relationships. Addressing these challenges requires systematic evaluation, which falls beyond the scope of this study. Further research could also explore the use of domain-adapted models such as CrisisBERT (Liu et al., 2021) for multimodal relevance classification, although its lack of multilingual support currently limits its applicability for global disaster scenarios.

One major limitation of all evaluated classification approaches is that they still require a large number of training data points. While Few-Shot Learning (FSL) has been widely explored for text classification, the integration of non-text features in such settings remains an open challenge (Tunstall et al., 2022). Trying to address this, we conducted preliminary evaluations of end-to-end inference using recent generative LLM, including GPT-4o (OpenAI, 2024), LLaMA-3 (Dubey et al., 2024), Qwen-2.5 (Yang et al., 2025), Mistral (Jiang et al., 2023), Gemma-2 (Riviere et al., 2024), and Phi-3 (Abdin et al., 2024). Notably, GPT-4o yielded a macro F1 score of 0.77 for purely semantic classification, when ignoring geographic space and time. However, it was still outperformed by our fine-tuned BERT model (TwHIN-BERT-base) for which the macro F1 score was 0.79 on the same train/test/validation split. Moreover, none of the tested LLM was effective in considering additional spatial or temporal modalities. Despite experimenting with various input formats (JSON, text) and prompting strategies (direct inference, chain-of-thought), model predictions remained the same or worsened, and classification performance generally dropped. In addition, we observed that zero-shot and few-shot prompting was highly sensitive to prompt engineering, structure and the provided examples, making the results difficult to reproduce and not directly comparable to our training-based approach. Most importantly, every LLM-based inference pipeline underperformed compared to our BERT-based text-only baseline. These preliminary findings suggest that multimodal classification with generative LLM, integrating geographic space and time, remains an open research problem.

For our evaluation, we used a dataset consisting of 4,574 samples, 3,659 of which were used for training and 915 for testing. It is the first annotated multimodal dataset for relevance classification of social media posts in disaster response, considering semantic, spatial and temporal aspects. Due to limited resources, we focused on developing a well-curated dataset of manageable size. Previous studies have shown that BERT-based models can achieve empirically robust performance with fewer than 5,000 training samples (Blomeier et al., 2024; Majdik et al., 2024; Phang et al., 2019), indicating that our dataset size should be sufficient for meaningful evaluation. Future work could explore the

benefits of scaling up the dataset and compare how different modelling approaches perform with varying dataset sizes. Follow-up research could also examine model behaviour using XAI techniques, particularly to understand the influence of toponyms, as task-specific training risks overfitting to location names.

Given access to a real-time social media stream, such as the X filtered stream (X Developer Platform, n.d.) or the Bluesky firehose (Bluesky, n.d.), all proposed classification approaches are suitable for real-time deployment once trained. Latency benchmarks (c.f. Table C1 in the appendix) show that BERT-based inference takes around 13 ms to 17 ms per post on an entry-level NVIDIA RTX A500 Laptop Graphics Processing Unit (GPU). All non-text and meta-classifier models, such as random forest or gradient boosting, consistently had negligible per-post latency (<0.1 ms). As a result, the expected throughput of our best-performing multimodal model (partial stacking with gradient boosting) exceeds 50 posts per second, even on an entry-level machine. Performance may vary depending on whether co-occurrence features are computed dynamically or pre-aggregated. In operational settings, they might be periodically updated in batches to maintain consistent throughput. Another consideration for real-time deployment concerns the dynamic nature of disaster zones. Our current approach treats disaster boundaries as static, using footprints from public official data sources (c.f. Section 3.3.1). However, in real-time scenarios, disaster zones are often fluid and only partially known as events unfold. To address this, our method can be adapted to update spatial footprints as they become available.

6. Conclusion

In this study, we presented a multimodal approach to relevance classification of geo-referenced social media posts in the context of disaster response. By integrating textual content with spatial and temporal features, including proximity to disaster impact locations, co-occurrence patterns with disaster-related content and contextual information about the disaster type and region, we addressed key limitations of purely text-based methods. Our approach was evaluated across five use case scenarios, spanning across multiple languages, geographic regions and disaster types.

The results demonstrate that spatiotemporal context offers valuable attributes for identifying relevant social media posts in disaster scenarios, both from a labelling and model perspective. Among the engineered features, the most effective were geographic distance, temporal distance and co-occurrence counts within the AOI, 50 km, and 10 km radii. The best-performing non-text classifier (a random forest model) achieved a macro F1 score of 0.713. These properties consistently helped distinguish relevant from irrelevant content, especially in datasets with broader geographic and temporal pre-filtering. SHAP analysis confirmed that the model learned meaningful, though partly non-linear, relationships between these features and post relevance. This directly addresses RQ1, demonstrating that spatial and temporal distance, together with co-occurrence patterns at moderate to broad spatial scales, are the most effective contextual properties for enhancing relevance classification in disaster response settings.

With regard to RQ2, we compared four strategies for integrating text and non-text features. These included feature concatenation using internal BERT

representations, in-context learning, full stacking of separate classifiers and partial stacking. In the partial stacking approach, we combined the softmax probabilities of a TwHIN-BERT-base text classifier with our spatiotemporal non-text features. The two ensemble-based methods consistently performed best. In particular, partial stacking with tree-based models achieved the highest macro F1 score of 0.814. In comparison, the best text-only classifier (TwHIN-BERT-base) reached a macro F1 score of 0.779. Decision-level fusion using ensemble models therefore was the most effective strategy for integrating text with contextual non-text features for multilingual relevance classification.

RQ3 focused on how classification performance differed by disaster type, language and region. Overall, our multimodal approach performed robustly across all use cases and achieved consistently high macro F1 scores ranging from 0.722 (California wildfires) to 0.855 (Ahr Valley floods). The largest improvement from combining text with non-text features was observed for the Ahr Valley floods, followed by the Chile wildfires and Turkey earthquake. These events were characterised by broader spatial coverage and more heterogeneous post distributions, which allowed the model to leverage spatiotemporal context effectively. In contrast, the best performance for the Emilia-Romagna floods dataset was achieved by the text-only classifier, likely due to the dataset's small size and narrow geographic focus. For the California wildfires, the multimodal and text models performed similarly, while the non-text model lagged behind. This was likely due to the widespread nature of the fires, which resulted in uniformly short distances between posts and fire locations, limiting the discriminative power of spatiotemporal features. Event-specific characteristics therefore play a crucial role for classification performance, while language and geographic region appear to be secondary factors.

Overall, this study demonstrates that combining text with spatiotemporal features can significantly improve the reliability of social media relevance classification in disaster contexts. Our approach is flexible, multilingual and applicable across different disaster types. The spatial and temporal distance features such as proximity to disaster impact zones, were derived from EO data, highlighting the value of EO for enhancing social sensing. Future work should explore online and few-shot learning scenarios, integrate additional modalities such as imagery and further refine the multimodal definition of relevance in collaboration with emergency responders. Beyond disaster management, our findings contribute to advancing GeoAI as a tool for the social sciences, where spatially explicit machine learning is increasingly used to analyse and understand human behaviour in space and time. By bridging language with spatiotemporal context in a unified framework, our approach offers new opportunities for extracting insights from social sensing data for public health, urban planning or communication science.

Acknowledgements

We would like to thank our colleague Andreas Kramer from IT:U, who helped us clarify an important inconsistency in the manuscript at the revision stage.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This project has received funding from the European Commission - European Union under HORIZON EUROPE (HORIZON Research and Innovation Actions) under grant agreement 101093003 [HORIZON-CL4-2022-DATA-01-01]. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union - European Commission. Neither the European Commission nor the European Union can be held responsible for them.

Notes on contributors

David Hanny is a PhD student with the GeoSocial AI group at IT:U. He holds a master's degree in data science from the University of Salzburg and a bachelor's degree in journalism from the University of Applied Sciences Vienna. His research focuses on spatially explicit machine learning of multi-modal social sensing data. He is particularly interested in social media and mobile phone data related to natural disasters and digital health.

Sebastian Schmidt is a research assistant with the GeoSocial AI group at IT:U and PhD student at the Department of Geoinformatics – Z GIS at the University of Salzburg. He holds a master's degree in Geography and a bachelor's degree in Romance philology from the University of Heidelberg. His research focuses on the analysis of geo-social media and corporate website data for environmental questions.

Shaily Gandhi is a Senior PostDoc with the GeoSocial AI Research group at IT:U, and a GIS expert with over 12 years of experience. She holds a PhD in Geospatial Technology from CEPT University, India, with a focus on bridging GIS and governance. Previously, she served as Deputy Center Head for the Center for Applied Geomatics, CRDF, and Program Chair for the M.Tech in Geomatics. Her research explores GIS, data science, and GeoAI models for the urban domain.

Michael Granitzer holds the Chair of Data Science at University of Passau since 2017 and is a Fellow Professor of Information Retrieval & Data Science at IT:U. His research interest is on applied machine learning, web information retrieval and natural language processing. He published over 250 mostly peer-reviewed publications including journal publications, book chapters and books in the above-mentioned fields.

Bernd Resch is Full Professor of GeoSocial AI at IT:U and a Visiting Scholar at Harvard University (USA). His research focuses on understanding cities as complex systems by analysing a wide range of digital data sources. In particular, he is involved in developing machine learning algorithms to analyse human-generated data such as social media posts and physiological measurements from wearable sensors. His findings contribute to various fields, including urban research, disaster management, epidemiology and beyond.

ORCID

David Hanny http://orcid.org/0009-0004-8017-0786

Data availability statement

Due to compliance with Twitter's (now X) API usage agreement, sharing the full tweet data used in this study is not feasible. Instead, we offer a list of tweet IDs along with our assigned ground truth relevance labels and the computed non-text features. Users can use the Twitter (X) v2 API to rehydrate the tweets based on the ID. This dataset is available on the Harvard Dataverse under https://doi.org/10.7910/DVN/0DBK04. The source code for reproducing the methodology, experiments, and results is available on GitHub: https://github.com/IT-U/GSAI_PUBLIC_Multimodal_Disaster_Relevance_Classification.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. arXiv: 2404.14219 [cs]. https://doi.org/10.48550/arXiv.2404.14219

Adwaith, D., Abishake, A. K., Raghul, S. V., & Sivasankar, E. (2022). Enhancing multi-modal disaster tweet classification using state-of-the-art deep learning networks. *Multimedia Tools & Applications*, *81*(13), 18483–18501. https://doi.org/10.1007/s11042-022-12217-3

Agarwal, N., & Yiliyasi, Y. (2010). Information quality challenges in social media. In *Proceedings of the International Conference on Information Quality (ICIQ 2010)*, Little Rock, Arkansas. ICIQ. https://www.researchgate.net/publication/260337476_Information_quality_challenges_in_social_media

Agenzia per la Sicurezza Territoriale e Protezione Civile Regione Emilia-Romagna. (2024. November 15). *Alluvione in Emilia-Romagna di maggio 2023, servizi cartografici a supporto delle attività di gestione dell'emergenza e della ricostruzione* (version 5). Geoportale regione Emilia-Romagna. Retrieved March 4, 2025, from https://geoportale.regione.emilia-romagna.it/approfondimenti/emergenza-maggio-23/emergenza-rer-maggio-2023-servizi

Agrawal, T. (2021). *Hyperparameter optimization in machine learning: Make your machine learning and deep learning models more efficient*. Apress. https://doi.org/10.1007/978-1-4842-6579-6

Alfarrarjeh, A., Agrawal, S., Kim, S. H., & Shahabi, C. (2017). Geo-spatial multimedia sentiment analysis in disasters. 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Tokyo, Japan (pp. 193–202). https://doi.org/10.1109/DSAA.2017.77

Arrighi, C., & Domeneghetti, A. (2024). Brief communication: On the environmental impacts of the 2023 floods in Emilia-Romagna (Italy). *Natural Hazards and Earth System Sciences*, *24*(2), 673–679. https://doi.org/10.5194/nhess-24-673-2024

Autelitano, A., Pernici, B., & Scalia, G. (2019). Spatio-temporal mining of keywords for social media cross-social crawling of emergency events. *Geoinformatica*, *23*(3), 425–447. https://doi.org/10.1007/s10707-019-00354-1

Banchoff, T. (1990). *Beyond the third dimension: Geometry, computer graphics, and higher dimensions*. Scientific American Library.

Barbieri, F., Espinosa Anke, L., & Camacho-Collados, J. (2022). XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 258–266). European Language Resources Association. https://aclanthology.org/2022.lrec-1.27

Blomeier, E., Schmidt, S., & Resch, B. (2024). Drowning in the information flood: Machine-learning-based relevance classification of flood-related tweets for disaster management. *Information*, *15*(3), 149. https://doi.org/10.3390/info15030149

Bluesky. (n.d.). *Firehose | Bluesky*. Retrieved July 30, 2025, from, https://docs.bsky.app/docs/advanced-guides/firehose

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Nee-Lakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bultmann, S., Quenzel, J., & Behnke, S. (2023). Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation. *Robotics and Autonomous Systems*, *159*, 104286. https://doi.org/10.1016/j.robot.2022.104286

Buntain, C., & Golbeck, J. (2017). Automatically identifying fake news in popular Twitter threads. In *2017 IEEE International Conference on Smart Cloud (Smart-Cloud)*, New York, NY, USA (pp. 208–215). https://doi.org/10.1109/SmartCloud.2017.40

Caragea, C., Silvescu, A., & Tapia, A. H. (2016). Identifying informative messages in disaster events using convolutional neural networks. In A. Tapia, P. Antunes, V. A. Bañuls, K. Moore, & J. Porto (Eds.), *Proceedings of the IS-CRAM, 2016 conference*. Federal University of Rio de Janeiro. https://idl.iscram.org/search.php?sqlQuery=SELECT%20author%2C%20title%2C%20type%2C%20year%2C%20publication%2C%20abbrev_journal%2C%20volume%2C%20issue%2C%20pages%2C%20keywords%2C%20abstract%2C%20address%2C%20corporate_author%2C%20thesis%2C%20publisher%2C%20place%2C%20editor%2C%20language%2C%20summary_language%2C%20orig_title%2C%20series_editor%2C%20series_title%2C%20abbrev_series_title%2C%20series_volume%2C%20series_issue%2C%20edition%2C%20issn%2C%20isbn%2C%20medium%2C%20area%2C%20expedition%2C%20conference%2C%20notes%2C%20approved%2C%20call_number%2C%20serial%20FROM%20refs%20WHERE%20serial%20%3D%201397%20ORDER%20BY%20author%2C%20year%20DESC%2C%20publication&client=&formType=sqlSearch&submit=Display&viewType=&showQuery=0&showLinks=1&showRows=25&rowOffset=&wrapResults=1&citeOrder=&citeStyle=APA&exportFormat=RIS&exportType=html&exportStylesheet=&citeType=html&headerMsg=

Chen, T., Hong, R., Guo, Y., Hao, S., & Hu, B. (2023). Ms^2-GNN: Exploring GNN- based multimodal fusion network for depression detection. *IEEE Transactions on Cybernetics*, *53*(12), 7749–7759. https://doi.org/10.1109/TCYB.2022.3197127

Chen, Z., & Lim, S. (2018). Collecting typhoon disaster information from Twitter based on query expansion. *ISPRS International Journal of Geo-Information*, *7*(4), 139. https://doi.org/10.3390/ijgi7040139

Chowdhury, J. R., Caragea, C., & Caragea, D. (2020). On identifying hashtags in disaster Twitter data. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(1), 498–506. https://doi.org/10.1609/aaai.v34i01.5387

Copernicus EMS. (2021. July 23). *Flood in western Germany* (EMSR517). Retrieved November 23, 2023, from https://mapping.emergency.copernicus.eu/activations/EMSR517/

Cordero, R. R., Feron, S., Damiani, A., Carrasco, J., Karas, C., Wang, C., Kraamwinkel, C. T., & Beaulieu, A. (2024). Extreme fire weather in Chile driven by climate change and El Niño-Southern oscillation (ENSO). *Scientific Reports*, *14*(1), 1974. https://doi.org/10.1038/s41598-024-52481-x

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *20*(2), 215–232. https://www.jstor.org/stable/2983890

Dal Zilio, L., & Ampuero, J.-P. (2023). Earthquake doublet in Turkey and Syria. *Communications Earth & Environment*, *4*(1), 1–4. https://doi.org/10.1038/s43247-023-00747-z

Damer, N., Dimitrov, K., Braun, A., & Kuijper, A. (2019). On learning joint multi-biometric representations by deep fusion. *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Tampa, FL, USA (pp. 1–8). https://doi.org/10.1109/BTAS46853.2019.9186011

de Albuquerque, J. P., Herfort, B., Brenning, A., & Zipf, A. (2015). A geographic approach for combining social media and authoritative data towards identifying useful information for disaster management. *International Journal of Geographical Information Science*, *29*(4), 667–689. https://doi.org/10.1080/13658816.2014.996567

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

Derczynski, L., Meesters, K., Bontcheva, K., & Maynard, D. (2018). Helping crisis responders find the informative needle in the tweet haystack. In K. Boersma & B. Tomaszewski (Eds.), *Proceedings of the 15th ISCRAM conference*. ISCRAM Association. https://idl.iscram.org/files/leonderczynski/2018/2139_LeonDerczynski_etal2018.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

D'mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), 1–36. https://doi.org/10.1145/2682899

Duan, S., Shi, Q., & Wu, J. (2022). Multimodal sensors and ML-based data fusion for advanced robots. *Advanced Intelligent Systems*, *4*(12), 2200213. https://doi.org/10.1002/aisy.202200213

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., & Zhao, Z. (2024. August 15). The llama 3 herd of models. *arXiv: 2407.21783 [cs]*. https://doi.org/10.48550/arXiv.2407.21783

Džeroski, S., & Ženko, B. (2004). Is combining classification with stacking better than selecting the best one? *Machine Learning*, *54*(3), 255–273. https://doi.org/10.1023/B:MACH.0000015881.36452.6e

Federal Emergency Management Agency. (2024. January 4). *2.3. Incident response and recovery generic operation phases | FEMA.gov*. Retrieved April 2, 2025, from https://www.fema.gov/oet-tools/chemical-incident-consequence-management/2/3

Fix, E., & Hodges, J. L. (1989). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *International Statistical Review/Revue Internationale de Statistique*, *57*(3), 238–247. https://doi.org/10.2307/1403797

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020). Early vs late fusion in multi-modal convolutional neural networks. *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, Rustenburg, South Africa (pp. 1–6). https://doi.org/10.23919/FUSION45008.2020.9190246

Gaudette, L., & Japkowicz, N. (2009). Evaluation methods for ordinal classification. In Y. Gao & N. Japkowicz (Eds.), *Advances in Artificial Intelligence* (pp. 207–210). Springer. https://doi.org/10.1007/978-3-642-01818-3_25

Habdank, M., Rodehutskors, N., & Koch, R. (2017). Relevancy assessment of tweets using supervised learning techniques: Mining emergency related tweets for automated relevancy classification 2017. *4th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM),* Münster, Germany (pp. 1–8). https://doi.org/10.1109/ICT-DM.2017.8275670

Hanny, D., & Resch, B. (2024). Multimodal geo-information extraction from social media for supporting decision-making in disaster management. *AGILE: GIScience Series*, *5*, 1–8. https://doi.org/10.5194/agile-giss-5-28-2024

Hanny, D., Schmidt, S., & Resch, B. (2024). Active learning for identifying disaster related tweets: A comparison with keyword filtering and generic fine-tuning. In K. Arai (Ed.), *Intelligent Systems and Applications* (pp. 126–142). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-66428-1_8

Havas, C., & Resch, B. (2021). Portability of semantic and spatial-temporal machine learning methods to analyse social media for near-real-time disaster monitoring. *Natural Hazards*, *108*(3), 2939–2969. https://doi.org/10.1007/s11069-021-04808-4

Havas, C., Resch, B., Francalanci, C., Pernici, B., Scalia, G., Fernandez-Marquez, J. L., Van Achte, T., Zeug, G., Mondardini, M. R., Grandoni, D., Kirsch, B., Kalas, M., & Lorini, V., & Rüping, S. (2017). E2mC: Improving emergency management service practice through social media and crowdsourcing analysis in near real time. *Sensors*, *17*(12), 2766. https://doi.org/10.3390/s17122766

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* (Vol. 1, pp. 278–282). IEEE. https://doi.org/10.1109/ICDAR.1995.598994

Huang, L., Shi, P., Zhu, H., & Chen, T. (2022). Early detection of emergency events from social media: A new text clustering approach. *Natural Hazards*, *111*(1), 851–875. https://doi.org/10.1007/s11069-021-05081-1

Huang, X., Li, Z., Wang, C., & Ning, H. (2020). Identifying disaster related social media for rapid response: A visual-textual fused CNN architecture. *International Journal of Digital Earth*, *13*(9), 1017–1039. https://doi.org/10.1080/17538947.2019.1633425

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., & Meier, P. (2013). Practical extraction of disaster-relevant information from social media. *Proceedings of the 22nd International Conference on World Wide Web*, 1021–1024. https://doi.org/10.1145/2487788.2488109

Imran, M., Mitra, P., & Castillo, C. (2016). Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. *arXiv preprint arXiv:1605.05894*. https://doi.org/10.48550/ARXIV.1605.05894

Jensen, G. E. (2012. January 1). *Key criteria for information quality in the use of online social media for emergency management in New Zealand* [Doctoral dissertation]. Open Access Te Herenga Waka-Victoria University of Wellington. https://doi.org/10.26686/wgtn.17003638

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., & Sayed, W. E. (2023). Mistral 7B. *arXiv: 2310.06825*. https://doi.org/10.48550/arXiv.2310.06825

Jiao, T., Guo, C., Feng, X., Chen, Y., & Song, J. (2024). A comprehensive survey on deep learning multi-modal fusion: Methods, technologies and applications. *Computers, Materials & Continua*, *80*(1), 1–35. https://doi.org/10.32604/cmc.2024.053204

Joo, T., Rogers, M. J., Soong, C., Hass-Mitchell, T., Heo, S., Bell, M. L., Ng, N. L., & Gentner, D. R. (2024). Aged and obscured wildfi smoke associated with downwind health risks. *Environmental Science & Technology Letters*, *11*(12), 1340–1347. https://doi.org/10.1021/acs.estlett.4c00785

Kaufhold, M.-A., Bayer, M., & Reuter, C. (2020). Rapid relevance classification of social media posts in disasters and emergencies: A system and evaluation featuring active, incremental and online learning. *Information Processing & Management*, *57*(1), 102132. https://doi.org/10.1016/j.ipm.2019.102132

Khalid, A. (2019. June 19). *Twitter removes precise geo-tagging option from tweets*. Engadget. Retrieved March 25, 2025, from https://www.engadget.com/2019-06-19-twitter-removes-precise-geo-tagging.html

Koks, E., Van Ginkel, K., Van Marle, M., & Lemnitzer, A. (2021). Brief communication: Critical infrastructure impacts of the 2021 mid-July Western European flood event. *Risk Assessment, Mitigation and Adaptation Strategies, Socioeconomic and Management Aspects*. https://doi.org/10.5194/nhess-2021-394

Koshy, R., & Elango, S. (2023a). Multimodal tweet classification in disaster response systems using transformer-based bidirectional attention model. *Neural Computing & Applications*, *35*(2), 1607–1627. https://doi.org/10.1007/s00521-022-07790-5

Koshy, R., & Elango, S. (2023b). Utilizing social media for emergency response: A tweet classification system using attention-based BiLSTM and CNN for resource management. *Multimedia Tools & Applications*, *83*(14), 41405–41439. https://doi.org/10.1007/s11042-023-16766-z

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec & C. Rouveirol (Eds.), *Machine learning: ECML-98* (pp. 4–15). Springer. https://doi.org/10.1007/BFb0026666

Li, J., Wang, Y., & Li, W. (2022). Mgmp: Multimodal graph message propagation network for event detection. In B. Þ. Jónsson, C. Gurrin, M.-T. Tran, D.-T. Dang-Nguyen, A.-M.-C. Hu, B. H. T. Thanh, & B. Huet (Eds.), *Multi-media modeling* (pp. 141–153). Springer International Publishing. https://doi.org/10.1007/978-3-030-98358-1_12

Li, S., & Tang, H. (2024. November 26). Multimodal alignment and fusion: A survey. *arXiv: 2411.17040 [cs]*. https://doi.org/10.48550/arXiv.2411.17040

Li, W., Liu, Q., Fan, S., Xu, C., & Bai, H. (2023). Dual-stream GNN fusion network for hyperspectral classification. *Applied Intelligence*, *53*(22), 26542–26567. https://doi.org/10.1007/s10489-023-04960-3

Lin, X., Lachlan, K. A., & Spence, P. R. (2016). Exploring extreme events on social media: A comparison of user reposting/retweeting behaviors on Twitter and Weibo. *Computers in Human Behavior*, *65*, 576–581. https://doi.org/10.1016/j.chb.2016.04.032

Liu, J., Capurro, D., Nguyen, A., & Verspoor, K. (2023). Attention-based multimodal fusion with contrast for robust clinical prediction in the face of missing modalities. *Journal of Biomedical Informatics*, *145*, 104466. https://doi.org/10.1016/j.jbi.2023.104466

Liu, J., Singhal, T., Blessing, L. T., Wood, K. L., & Lim, K. H. (2021). CrisisBERT: A robust transformer for crisis classification and contextual crisis embedding. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media* (pp. 133–141). Association for Computing Machinery. https://doi.org/10.1145/3465336.3475117

Lucas, B., Vahedi, B., & Karimzadeh, M. (2023). A spatiotemporal machine learning approach to forecasting COVID-19 incidence at the county level in the USA. *International Journal of Data Science and Analytics*, *15*(3), 247–266. https://doi.org/10.1007/s41060-021-00295-9

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv: 1705.07874 [cs, stat]*. https://doi.org/10.48550/arXiv.1705.07874

Madichetty, S. M. S., & Madisetty, S. (2023). A RoBERTa based model for identifying the multi-modal informative tweets during disaster. *Multimedia Tools & Applications*, *82*(24), 37615–37633. https://doi.org/10.1007/s11042-023-14780-9

Madichetty, S., Muthukumarasamy, S., & Jayadev, P. (2021). Multi-modal classification of Twitter data during disasters for humanitarian response. *Journal of Ambient Intelligence and Humanized Computing*, *12*(11), 10223–10237. https://doi.org/10.1007/s12652-020-02791-5

Madichetty, S., & Sridevi, M. (2019). Detecting informative tweets during disaster using deep neural networks. *2019 11th International Conference on Communication Systems & Networks (COMSNETS)*, Bengaluru, India (pp. 709–713). https://doi.org/10.1109/COMSNETS.2019.8711095

Majdik, Z. P., Graham, S. S., Shiva Edward, J. C., Rodriguez, S. N., Karnes, M. S., Jensen, J. T., Barbour, J. B., & Rousseau, J. F. (2024). Sample size considerations for fine-tuning large language models for named entity recognition tasks: Methodological study. *JMIR AI*, *3*, e52095. https://doi.org/10.2196/52095

NASA LANCE. (2023). *NASA Fire Information for Resource Management System*. Fire information for resource management system. Retrieved March 30, 2025, from https://firms.modaps.eosdis.nasa.gov/map/Land,AtmosphereNear-real-time CapabilityforEOS (LANCE)

Nguyen, D., Ali Al Mannai, K., Joty, S., Sajjad, H., Imran, M., & Mitra, P. (2017). Robust classification of crisis-related data on social networks using convolutional neural networks. *Proceedings of the International AAAI Conference on Web & Social Media*, *11*(1), 632–635. https://doi.org/10.1609/icwsm.v11i1.14950

Nikparvar, B., & Thill, J.-C. (2021). Machine learning of spatial data. *ISPRS International Journal of Geo-Information*, *10*(9), 600. https://doi.org/10.3390/ijgi10090600

Olteanu, A., Vieweg, S., & Castillo, C. (2015). What to expect when the unexpected happens: Social media communications across crises. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, Vancouver, BC, Canada (pp. 994–1009). https://doi.org/10.1145/2675133.2675242

OpenAI. (2024. July 18). *GPT-4o mini: Advancing cost-efficient intelligence*. Retrieved August 7, 2024, from https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Parimala, M., Swarna Priya, R. M., Praveen Kumar Reddy, M., Lal Chowdhary, C., Kumar Poluru, R., & Khan, S. (2021). Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. *Software: Practice & Experience*, *51*(3), 550–570. https://doi.org/10.1002/spe.2851

Park, S., Vien, A. G., & Lee, C. (2024). Cross-modal transformers for infrared and visible image fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(2), 770–785. https://doi.org/10.1109/TCSVT.2023.3289170

Paul, N. R., Balabantaray, R. C., & Sahoo, D. (2023). Fine-tuning transformer-based representations in active learning for labelling crisis dataset of tweets. *SN Computer Science*, *4*(5), 553. https://doi.org/10.1007/s42979-023-02061-z

Phang, J., Févry, T., & Bowman, S. R. (2019). Sentence encoders on STILTs: Supplementary training on intermediate labeled-data tasks. *arXiv: 1811.01088 [cs]*. https://doi.org/10.48550/arXiv.1811.01088

Popova, M., Siemens, E., & Karpov, K. (2023). The concept of text processing in an ontological approach to spatio-temporal social network analysis. *2023 30th International Conference on Systems, Signals and Image Processing (IWSSIP)*, Ohrid, North Macedonia (pp. 1–4). https://doi.org/10.1109/IWSSIP58668.2023.10180274

Powers, C. J., Devaraj, A., Ashqeen, K., Dontula, A., Joshi, A., Shenoy, J., & Murthy, D. (2023). Using artificial intelligence to identify emergency messages on social media during a natural disaster: A deep learning approach. *International Journal of Information Management Data Insights*, *3*(1), 100164. https://doi.org/10.1016/j.jjimei.2023.100164

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv: 2103.00020 [cs]. https://doi.org/10.48550/arXiv.2103.00020

Resch, B., Usländer, F., & Havas, C. (2018). Combining machine-learning topic models and spatio-temporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *45*(4), 362–376. https://doi.org/10.1080/15230406.2017.1356242

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., & Andreev, A. (2024). Gemma, 2: Improving open language models at a practical size. *arXiv: 2408.00118 [cs]*. https://doi.org/10.48550/arXiv.2408.00118

Sachdeva, S., McCaffrey, S., & Locke, D. (2017). Social media approaches to modeling wildfi smoke dispersion: Spatiotemporal and social scientific investigations. *Information Communication & Society*, *20*(8), 1146–1161. https://doi.org/10.1080/1369118X.2016.1218528

Safford, H. D., Paulson, A. K., Steel, Z. L., Young, D. J. N., & Wayman, R. B. (2022). The 2020 California fire season: A year like no other, a return to the past or a harbinger of the future? *Global Ecology & Biogeography*, *31*(10), 2005–2025. https://doi.org/10.1111/geb.13498

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide*

Web, Raleigh, North Carolina, USA (pp. 851–860). Association for Computing Machinery. https://doi.org/10.1145/1772690. 1772777.

Samuel, D. (2024). BERTs are generative in-context learners. Advances in Neural Information Processing Systems, Vancouver, BC, Canada (Vol. 37. pp. 2558–2589). Curran Associates Inc. Retrieved April 4, 2025, from https://proceedings.neurips.cc/paper_files/paper/2024/hash/04ea184dfb5f1babb78c093e850a83f9-Abstract-Conference.html

Sarmiento, H., & Poblete, B. (2021). Crisis communication: A comparative study of communication patterns across crisis events in social media. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*, New York, NY, USA (pp. 1711–1720). Association for Computing Machinery. https://doi.org/10.1145/3412841.3442044

Sauerborn, R., & Ebi, K. (2012). Climate change and natural disasters - integrating science and practice to protect health. *Global Health Action*, *5*(1), 19295. https://doi.org/10.3402/gha.v5i0.19295

Scheele, C., Yu, M., & Huang, Q. (2021). Geographic context-aware text mining: Enhance social media message classification for situational awareness by integrating spatial and temporal features. *International Journal of Digital Earth*, *14*(11), 1721–1743. https://doi.org/10.1080/17538947.2021.1968048

Schmidt, S., Friedemann, M., Hanny, D., Resch, B., Riedlinger, T., & Mühlbauer, M.. (2025). Enhancing satellite-based emergency mapping: {{identifying}} wildfires through Geo-social Media analysis. *Big Earth Data*, 1–23. https://doi.org/10.1080/20964471.2025.2454526

Serere, H. N., & Resch, B. (2025). Dynamic named entity recognition model to distinguish authors' positions relative to mentioned locations. *Geomatica*, *77*(1), 100055. https://doi.org/10.1016/j.geomat.2025.100055

Serere, H. N., Resch, B., & Havas, C. R. (2023). 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*(3), e0282942. https://doi.org/10.1371/journal.pone.0282942

Shah, S. A., Yahia, S. B., McBride, K., Jamil, A., & Draheim, D. (2021). Twitter streaming data analytics for disaster alerts. *2021 2nd International Informatics and Software Engineering Conference (IISEC)*, 1–6. https://doi.org/10.1109/IISEC54230.2021.9672370

Shahian Jahromi, B., Tulabandhula, T., & Cetin, S. (2019). Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles. *Sensors*, *19*(20), 4357. https://doi.org/10.3390/s19204357

Shapley, L. S. (1953. December 31). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games (AM-28), volume II* (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Snoek, C. G. M., Worring, M., & Smeulders, A. W. M. (2005). Early versus late fusion in semantic video analysis. In *Proceedings of the 13th Annual ACM International Conference on Multimedia*, Hilton, Singapore (pp. 399–402). Association for Computing Machinery. https://doi.org/10.1145/1101149.1101236

Sodoge, J., Kuhlicke, C., & De Brito, M. M. (2023). Automatized spatio-temporal detection of drought impacts from newspaper articles using natural language processing and machine learning. *Weather and Climate Extremes*, *41*, 100574. https://doi.org/10.1016/j.wace.2023.100574

Tang, L., Hu, Q., Wang, X., Liu, L., Zheng, H., Yu, W., Luo, N., Liu, J., & Song, C. (2024). A multimodal fusion network based on a cross-attention mechanism for the classification of parkinsonian tremor and essential tremor. *Scientific Reports*, *14*(1), 28050. https://doi.org/10.1038/s41598-024-79111-w

Tekumalla, R., & Banda, J. M. (2022). An empirical study on characterizing natural disasters in class imbalanced social media data using weak supervision. *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan (pp. 4824–4832). https://doi.org/10.1109/BigData55660.2022.10020594

Tian, M., Hu, X., Huang, J., Ma, K., Li, H., Zheng, S., Tao, L., & Qiu, Q. (2023). Spatio-temporal relevance classification from geographic texts using deep learning. *ISPRS International Journal of Geo-Information*, *12*(9), 359. https://doi.org/10.3390/ijgi12090359

Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). *Efficient few-shot learning without prompts*. *arXiv: 2209.11055 [cs]*. https://doi.org/10.48550/arXiv.2209.11055

USGS. (2023). *Earthquake catalog*. Latest earthquakes. Retrieved March 30, 2025, from https://earthquake.usgs.gov/earthquakes/search/

Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J., Palmer, M., Schram, A., & Anderson, K. (2011). Natural language processing to the rescue? Extracting "situational awareness" tweets during mass emergency. *Proceedings of the International AAAI Conference on Web & Social Media*, *5*(1), 385–392. https://doi.org/10.1609/icwsm.v5i1.14119

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1079–1088). https://doi.org/10.1145/1753326.1753486

Wallin, K. (2025). *California fire perimeters (all)*. California, USA, California State Geoportal. Retrieved March 4, 2025, from https://gis.data.ca.gov/datasets/CALFIRE-Forestry:california-fire-perimeters-all/explore

Wang, Z., & Ye, X. (2018). Social media analytics for natural disaster management. *International Journal of Geographical Information Science*, *32*(1), 49–72. https://doi.org/10.1080/13658816.2017.1367003

Watanabe, S. (2023. May 26). Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv: 2304.11127 [cs]*. https://doi.org/10.48550/arXiv.2304.11127

Wiegmann, M., Kersten, J., Senaratne, H., Potthast, M., Klan, F., & Stein, B. (2021). Opportunities and risks of disaster data from social media: A systematic review of incident information. *Natural Hazards and Earth System Sciences*, *21*(5), 1431–1444. https://doi.org/10.5194/nhess-21-1431-2021

Wieland, M., Schmidt, S., Resch, B., Abecker, A., & Martinis, S. (2025). Fusion of geospatial information from remote sensing and social media to prioritise rapid response actions in case of floods. *Natural Hazards*, *121*(7), 8061–8088. https://doi.org/10.1007/s11069-025-07120-7

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

X Developer Platform. (n.d.). *Filtered stream introduction*. X. Retrieved July 30, 2025, from https://docs.x.com/x-api/posts/filtered-stream/introduction

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., & Qiu, Z. (2025). Qwen2.5 Technical Report. arXiv: 2412.15115 [cs]. https://doi.org/10.48550/arXiv.2412.15115

Yigitcanlar, T., Regona, M., Kankanamge, N., Mehmood, R., D'Costa, J., Lindsay, S., Nelson, S., & Brhane, A. (2022). Detecting natural hazard-related disaster impacts with social media analytics: The case of Australian states and territories. *Sustainability*, *14*(2), 810. https://doi.org/10.3390/su14020810

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., & El-Kishky, A. (2023). TwHIN-BERT: A socially-enriched pre-trained language model for multilingual tweet representations at Twitter. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach, CA, USA (pp. 5597–5607). Association for Computing Machinery. https://doi.org/10.1145/3580305.3599921

Zhou, X., & Chen, L. (2014). Event detection over Twitter social media streams. *The VLDB Journal*, *23*(3), 381–400. https://doi.org/10.1007/s00778-013-0320-3