AGILE: GIScience Series, 7, 4, 2026. https://doi.org/10.5194/agile-giss-7-4-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.
© Author(s) 2026. This work is distributed under the Creative Commons Attribution 4.0 License.

Towards Fine-grained Relevance Scoring of Social Media Posts in Disaster Response: A Decay-based Regression Approach

David Hanny [1], Andreas Kramer [1,2], Ehsaneddin Jalilian [1], Sebastian Schmidt [1,3], and Bernd Resch [1,4]

[1]GeoSocial Artificial Intelligence, Interdisciplinary Transformation University (IT:U), Linz, Austria
[2]Department of Computer Science, Idaho State University, ID, USA
[3]Department of Geoinformatics - Z_GIS, University of Salzburg, Salzburg, Austria
[4]Center for Geographic Analysis, Harvard University, Cambridge, MA, USA

Correspondence: David Hanny (david.hanny@it-u.at)

Abstract. Geo-social media posts can provide valuable real-time information during natural disasters. However, assessing their relevance for emergency response remains difficult. Most existing approaches simplify relevance into discrete classes. To incorporate space and time, they typically rely on manually engineered distance features, overlooking the non-linear effects of spatial and temporal proximity. This study introduces a more fine-grained approach for multimodal relevance assessment that integrates spatio-temporal decay transformations with ordinal regression. Using a multilingual, geo-referenced X (formerly Twitter) dataset spanning floods, wildfires, and earthquakes, we evaluated four decay transformations and reformulated relevance assessment as a regression task. A stretched exponential decay function best captured the non-linear decline of relevance with increasing spatial and temporal distance. Incorporating decay-transformed features into a multimodal meta-learning framework improved both prediction performance and stability. Reformulating the task as a regression problem reduced RMSE and increased R^2 compared to classification, with Support Vector Regression (SVR) achieving the strongest results (RMSE: 0.231, R^2: 0.617). Granularity and entropy analyses revealed that regression provided much finer relevance estimates than classification. Overall, our research presents a first step towards making insights from social media more actionable and decision-ready for disaster response.

Submission Type. model, analysis

BoK Concepts. [AM10] Data mining, [GS3] Use of geospatial information

Keywords. disaster management, geo-social media, spatio-temporal decay, regression analysis, ordinal modelling, GeoAI

1 Introduction

Social media data has become a valuable complementary data source for disaster management. During floods, wildfires, or earthquakes, citizens frequently post eyewitness observations, requests for help, or emotional reactions that can support traditional monitoring systems (Blomeier et al., 2024; Acikara et al., 2023). A subset of this user-generated content is geographically and temporally referenced (Serere et al., 2023), enabling the analysis of online communication patterns across space and time (Wang and Ye, 2018). However, the high volume, noise, and multimodal nature of social media data make it difficult to determine which posts are actually relevant for emergency responders. As a result, the automatic assessment of post relevance has become a key research problem in Geospatial Artificial Intelligence (GeoAI) (Kaufhold et al., 2020; Koshy and Elango, 2023; Hanny et al., 2025).

Existing relevance assessment methods typically treat the task as a supervised text classification problem, categorising posts into discrete classes such as *relevant* or *irrelevant* (e.g. de Albuquerque et al., 2015; Blomeier et al., 2024). Recent studies further incorporate image data (e.g. Koshy and Elango, 2023) or contextual features such as spatial and temporal proximity to the disaster impact site (Scheele et al., 2021; Hanny et al., 2025). The latter is particularly important in the context of geoinformatics, as post relevance is highly dependent on such distances (Vieweg et al., 2010). However, two limitations persist with existing approaches: First, spatio-temporal features are usually manually engineered, commonly as raw distances to the event epicentre or using fixed thresholds (Scheele et al., 2021). Yet, these representations do not leverage prior knowledge about the non-linear relationship between distance and post relevance (Yabe et al., 2020),

and instead force the model to learn this decay pattern from the data. As a result, performance becomes heavily dependent on the model type and its ability to learn these non-linear relationships (Yan et al., 2019). Second, relevance is commonly treated as a categorical variable. While this simplifies model training, it fails to capture the fact that human judgments of relevance are much more nuanced. Some posts are more useful or urgent than others, even within the same label category.

This research aims to address both limitations by introducing two methodological extensions for relevance assessment of social media content: (1) We quantify existing distance-decay relationships in annotated posts and integrate non-linear spatial and temporal decay transformations directly into relevance assessment methods. (2) We further propose a regression method that represents relevance as a continuous-valued number instead of discrete ordinal classes. Using the multimodal benchmark dataset of Hanny et al. (2025), which contains geo-referenced X (formerly Twitter) posts annotated based on text, spatial, and temporal attributes across multiple disaster events, we evaluate how our proposed extensions affect predictive performance and interpretability.

Overall, our work is guided by two central research questions:

- **RQ1:** To what extent can distance decay transformations enhance the predictive value of spatio-temporal features for relevance assessment of disaster-related social media data?
- **RQ2:** How effectively does regression on ordinal categories capture the graded nature of relevance compared to discrete classification of disaster-related social media content?

2 Related Work

Identifying which social media posts provide actionable information during natural disasters has long been a core challenge in disaster management and geoinformatics. Early studies defined relevance as a binary construct, distinguishing useful from non-useful posts for situational awareness. Vieweg et al. (2010) and Imran et al. (2013) demonstrated that keyword-filtered Twitter streams contained valuable on-the-ground information but required supervised filtering to separate such informative content. Subsequent work applied traditional machine-learning algorithms such as Naive Bayes classifiers, Support Vector Machines (SVMs), and random forests using term frequency (tf) or term frequency inverse document frequency (tf-idf) features (Verma et al., 2011; Kaufhold et al., 2020).

The rise of deep learning led to more sophisticated text representations and thereby improved filtering performance. Caragea et al. (2016) and Madichetty and Sridevi (2019) used Convolutional Neural Networks (CNNs) to capture semantic patterns within texts. Later, transformer models such as Bidirectional Encoder Representations from Transformers (BERT) and Robustly Optimised BERT Pre-training Approach (RoBERTa) became the de facto standard for relevance classification (Madichetty et al., 2021; Blomeier et al., 2024). Multimodal architectures that fuse textual and visual features have also emerged as a promising research direction. Koshy and Elango (2023) combined a fine-tuned RoBERTa model with a Vision Transformer, while Adwaith et al. (2022) evaluated hybrid text–image networks for identifying informative disaster imagery. Despite these advances, most approaches still treat relevance as a discrete label, ignoring gradual differences in information usefulness.

Annotation schemes beyond binary classification have also emerged. de Albuquerque et al. (2015) introduced three classes – *off-topic*, *on-topic but irrelevant*, and *on-topic and relevant* – to capture varying levels of content relevance. Olteanu et al. (2015) extended this taxonomy by distinguishing between *related and informative* versus *related but not informative* content. Blomeier et al. (2024) proposed an ordinal four-class relevance scheme from *very relevant* to *not relevant*, emphasising practical helpfulness for emergency responders. Hanny et al. (2025) took a similar approach with three classes and annotated posts as *not related* to the disaster, *related but not relevant*, and *related and relevant*. However, even in such ordinal settings, models are typically trained as multi-class classifiers, which restricts categorisation to a few discrete labels. This discretisation overlooks the fact that human judgements may reflect more fine-grained differences: two posts within the same class can still vary in their relevance to emergency responders.

This gap is notable given that transformer-based language models have been widely applied to regression tasks in recent years. A popular example is polarity-based sentiment analysis, where models predict a continuous score ranging from –1 (very negative) to +1 (very positive) (Rodríguez-Ibánez et al., 2023). Wang et al. (2022) furthermore demonstrated that transformer-based encoder models such as BERT can be explicitly calibrated to produce continuous-valued scores rather than discrete labels, and several studies have proposed specialised loss functions to improve continuous-valued predictions (e.g. Zhang and Li, 2024). Despite these developments, text regression in social media analysis has, so far, been used primarily for predicting sentiment polarity (Kanaparthi et al., 2023) or user attributes such as age (Chen et al., 2016) or engagement rates (Ngo-Ye and Sinha, 2014).

Beyond textual content, geographic information about where a post was created also provides valuable cues for assessing relevance in disaster response (Hanny et al., 2025). Posts located closer to the disaster epicentre tend to be more useful, as they are more likely to contain eyewitness observations or locally relevant information

(Vieweg et al., 2010; Kaufhold et al., 2020). Research suggests that this spatial dependence follows a non-linear distance decay pattern, where post relevance decreases rapidly close to the event and then more slowly at larger distances (Kottwitz et al., 2023; Wu et al., 2019). Similar spatial decay effects have also been observed in overall disaster-related posting activity (Resch et al., 2018).

Temporal proximity follows a comparable pattern. Posts created closer in time to the disaster event tend to be more relevant for emergency response, because they are more likely to reflect immediate impact conditions (Vieweg et al., 2010). Relevance in time has likewise been shown to follow a non-linear decay pattern, characterised by a rapid rise in relevant posting activity shortly after event onset, followed by a gradual decline (Resch et al., 2018; Kottwitz et al., 2023).

Previous work has attempted to incorporate these spatial and temporal cues by manually engineering proximity features, typically based on raw distances, fixed radii, or binary distance thresholds (Kaufhold et al., 2020; Scheele et al., 2021; Hanny et al., 2025). However, raw distances leave it entirely up to the model to infer the complex, non-linear decay between distance and relevance. The ability to learn these patterns can vary greatly across different model types, especially when training data are limited or imbalanced (Ouyang et al., 2019; Yan et al., 2019). Simultaneously, fixed radii treat all posts within a buffer as equally relevant, and threshold features naturally yield abrupt, step-wise changes.

Given how proximity features were used in earlier work, important gaps remain. Prior studies have not systematically analysed how relevance actually decays with increasing spatial or temporal distance, nor have they done so across multiple disaster events. Moreover, distance-decay functions have not been integrated into the relevance classification task itself. Instead, they are typically applied only after classification, for example, when interpolating flood extents from labelled tweets (Huang et al., 2019). Furthermore, even though transformer models can produce continuous-valued outputs, relevance in a disaster context has not yet been modelled as a regression problem. In this work, we address these gaps by (i) quantifying distance-decay effects in relevance annotations, (ii) incorporating decay-based proximity features directly into the model, and (iii) predicting relevance as a continuous-valued score.

3 Materials and Methods

Building on the identified research gaps, we first focus on integrating spatio-temporal decay into relevance assessment, and subsequently investigate the effects of ordinal regression compared to classification. An overview of the study design is provided in Figure 1.

3.1 Data

This study builds on an existing multimodal dataset introduced by Hanny et al. (2025), which contains 4,574 geo-referenced social media posts collected from five major natural disasters: the 2020 California wildfires, the 2021 Ahr Valley floods in Germany, the 2023 Chile wildfires, the 2023 Emilia-Romagna floods in Italy, and the 2023 Turkey earthquake. Each tweet in the dataset was annotated based on textual content, timestamp, and geolocation (either coordinates or bounding box) using a three-class labelling scheme, covering the classes *Not related* (36%), *Related but not relevant* (42%), and *Related and relevant* (22%). Although these classes are discrete, we argue that they reflect increasing degrees of relevance, ranging from content that is not disaster-related to disaster-related and directly relevant for emergency response. On this basis, we treated the class labels as ordinal in their respective order. We mapped them to continuous relevance scores in the normalised range $[0, 1]$, assigning *Not related* = 0, *Related but not relevant* = 0.5, and *Related and relevant* = 1. While the exact numeric encoding could be arbitrary, using a normalised range of $[0, 1]$ ensured that all scores remained intuitive and easy to interpret. Likewise, we used equal intervals to avoid additional assumptions about unequal class spacing. Accordingly, the chosen numeric encoding represents a pragmatic choice that preserves ordinal structure and enables the learning of continuous-valued relevance scores in a regression setting. In line with the experimental setup of Hanny et al. (2025), we adopted the same data split, using 3,659 posts for training and 915 posts for testing to ensure comparability of results. Table 1 provides an overview of the size of each disaster-specific subset.

Table 1. Overview of our evaluation data.

Use case	Train size	Test size
2020 California wildfires	836	204
2021 Ahr Valley floods	802	224
2023 Chile wildfires	767	157
2023 Emilia Romagna floods	442	100
2023 Turkey earthquake	812	230

3.2 Baseline Method

As a baseline, we adopted the partial stacking methodology proposed by Hanny et al. (2025). In their study, the authors manually engineered a set of spatio-temporal non-text features which captured spatial and temporal proximity to disaster impact sites, local co-occurrences with disaster-related posts, event type and geographic context. After removing highly correlated variables, the final feature set comprised 12 dimensions. The authors showed that these features differed statistically across relevance classes and could be used for predicting post relevance, achieving macro F1

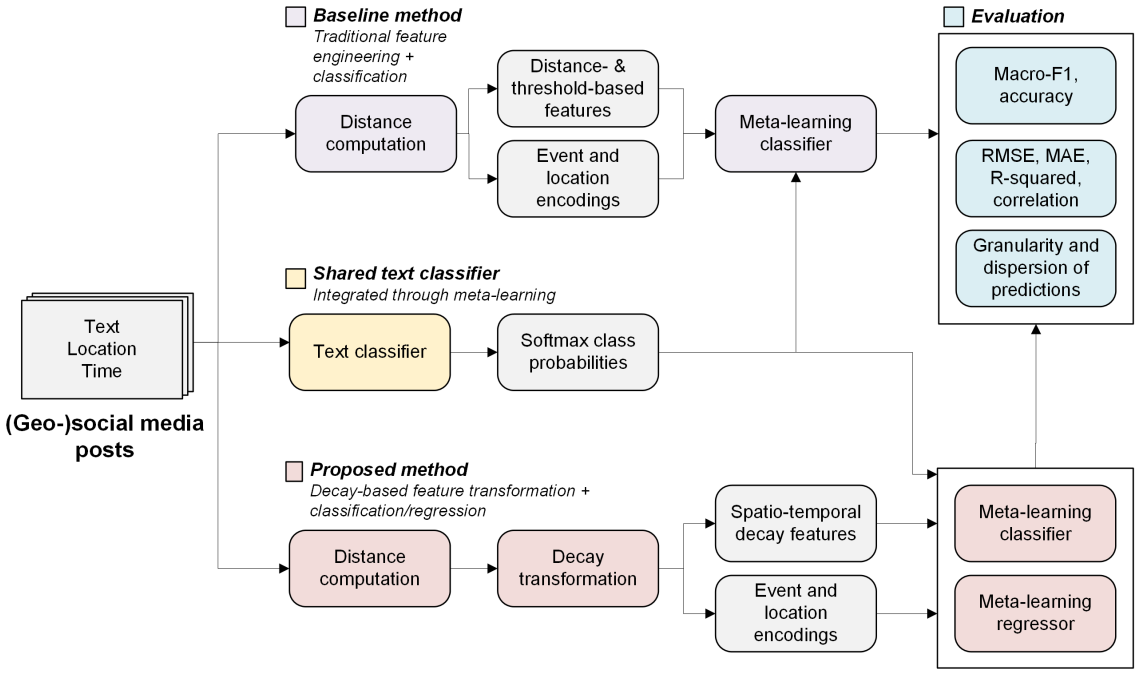

Figure 1. Overview of our methodology.

scores of up to 0.713. A detailed overview of the feature set is provided in Table 2. For multimodal classification, the authors applied late-stage multimodal fusion using partial stacking, where the softmax probabilities of a fine-tuned TwHIN-BERT-base text classifier[1] were concatenated with the engineered spatio-temporal features and used as input to a meta-learner. TwHIN-BERT (Zhang et al., 2023) is a multilingual transformer model pre-trained on 7 billion tweets, and has been shown to outperform models such as BERTweet (Nguyen et al., 2020), making it well-suited for classification tasks on social media posts. In their study, Hanny et al. (2025) evaluated multiple machine learning models as meta-learners, finding that tree-based gradient boosting and random forest models performed best (macro F1 0.814 and 0.813), whereas SVM and k-Nearest Neighbour (kNN) classification performed worst (macro F1 0.575 and 0.676). For reference, applying the fine-tuned TwHIN-BERT classifier on text alone, without any spatio-temporal features, achieved a macro-F1 score of 0.779.

3.3 Spatio-temporal Decay

To obtain distance features that generalise across different model types, we transformed raw spatial and temporal distances using a selection of non-linear decay functions derived from previous research (Kottwitz et al., 2023; Wu et al., 2019), whose parameters were fitted on our training data. These decay functions map distance values to normalised scores in the range $[0, 1]$, replacing the distance and co-occurrence features from the baseline method (c.f. Sec. 3.2). This allows relevance to decrease gradually and non-linearly with increasing distance or time, even when used in linear models. The resulting representation, therefore, was expected to provide more robust model performance. We evaluated the decay-based representations both with and without event-type and location encodings, which serve as domain-independent contextual variables that may capture complementary information not directly reflected by the decay functions.

Because the original relevance labels were discrete and took on only three possible values, the raw distance–relevance data showed abrupt jumps rather than a continuous trend, making it unsuitable for direct curve fitting. To obtain a smoother relationship, we binned the training data into quantile-based intervals of 70 samples each (approximately 50 bins) for geographic and temporal distances separately and computed the mean relevance score within each bin. The choice of bin number and size was motivated as a compromise between the square-root rule, which recommends setting the number of intervals to approximately $k = \sqrt{n}$, based on the total sample size n (Pearson, 1892), and the Rice University Rule, which recommends $k = 2 \cdot \lceil n^{\frac{1}{3}} \rceil$ (Rubia, 2024). For our training set of $n = 3,659$ samples, these rules would result in roughly 60 and 30 bins, respectively. We eventually selected a bin number towards the upper end of this range to better capture non-linear decay patterns. Finally, we fitted four decay functions to this smoothed data using non-linear least squares optimisation (Dennis Jr. and Welsch, 1978):

[1] The fine-tuned TwHIN-BERT model was downloaded from HuggingFace: https://huggingface.co/hannybal/multilingual-disaster-relevance-twhin-bert

Table 2. Overview of the non-text features used for relevance classification in the baseline method (Hanny et al., 2025).

Name	Description	Dimensions
Geographic distance	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, 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

- **Linear:** $f(d) = a - b \cdot d$
- **Exponential:** $f(d) = a \cdot e^{-b \cdot d}$
- **Inverse:** $f(d) = \frac{a}{b+d}$
- **Stretched exponential:** $f(d) = a \cdot e^{-(b \cdot d)^c}$

where $f(d)$ is the normalised score at distance d, and a, b, and c are free parameters that determine the scale and shape of the decay.

Goodness of fit was measured by the coefficient of determination R^2 (Fisher, 1992), which quantifies the proportion of the variance in the target variable that is explained by the model. The best-fitting decay functions for geographic and temporal distances were then used to transform distance values into decay-based proximity features. To evaluate the effect of the decay-based representation, we used the same meta-learning setup as in the baseline, replacing the manually engineered variables with our decay features. For the meta-classifier, we evaluated logistic regression (Cox, 1958), a random forest (Ho, 1995), a SVM (Cortes and Vapnik, 1995), a gradient boosting decision tree model (Friedman, 2001), a kNN classifier (Fix and 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, tree-based ensembles, instance-based learning, and probabilistic classification. The hyperparameters of each model were optimised for the validation macro F1 score via grid search and 5-fold cross-validation on the training data (Agrawal, 2021). The parameter space used for the models is summarised in Table B1 in the appendix.

For evaluation purposes, we mainly focused on the macro F1 score and accuracy on the test data. As Chicco et al. (2021) noted, justifying the choice of evaluation metrics is essential for meaningful interpretation of results. The macro F1 score reflects the balanced average of precision and recall across all classes (Lewis, 1991). It penalises large discrepancies and rewards consistency across the two (Rainio et al., 2024). Accuracy represents a more intuitive metric and indicates the proportion of correct predictions made (Mitchell, 1997). Since the learning task at hand was multi-class classification with discrete labels, and the evaluation dataset was somewhat balanced, we expected the two metrics to provide a reliable estimate of model performance (Opitz, 2024). To further assess the robustness of each feature configuration, we quantified the variation of these metrics across meta-learners, using standard deviation, min–max range, and the Coefficient of Variation (CV) (Brown, 1998).

3.4 Ordinal Regression

We understand post relevance as an inherently continuous-valued concept. For instance, the posts *"Heavy rain all night, some nearby villages are flooded"* and *"Our house is under water, we need rescue"* may both be considered relevant, but the latter conveys a higher degree of urgency, and it is consequently more important that responsible disaster management entities receive this information. Therefore, instead of treating relevance assessment as a discrete classification problem, we reformulated it as a regression task to better capture these gradual differences.

Accordingly, we used our numerically encoded ordinal relevance labels $\{0, 0.5, 1\}$ and applied the same partial stacking setup as in the classification experiments, adapted for regression. In this configuration, we combined the softmax outputs of the text classifier with the best-performing decay-based proximity features and the contextual event and location encodings. The resulting fused representation served as input to a selection of regression models: Ridge regression (Hoerl and Kennard, 1970), Support Vector Regression (SVR) (Drucker et al., 1996), a kNN regressor (Fix and Hodges, 1989), a random forest regressor (Ho, 1995) and a gradient boosting regressor (Friedman, 2001). Again, we selected this range of models to represent a diverse set of learning paradigms. Regression performance was assessed using the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R^2, and Spearman's rank correlation (ρ). The RMSE measures the average magnitude of prediction errors, penalising larger deviations more strongly (Chai and Draxler, 2014). The MAE captures the absolute average error magnitude (Willmott and Matsuura, 2005). R^2 provides an interpretable measure of model fit, as it is upper-bounded by the value 1, for perfect model fit, whereas 0 corresponds to predicting the mean value of all points (Chicco et al., 2021). Lastly, Spearman's rank correlation (Spearman, 1904) assesses the monotonic relationship between predicted and true relevance scores, capturing how well the

model preserves the ordinal ranking of the data. All model hyperparameters were optimised for validation RMSE using grid search (c.f. Table B1) and 5-fold cross-validation. To assess whether ordinal regression also retains meaningful classification performance, we additionally rounded the continuous-valued predictions to the nearest discrete relevance label $\{0, 0.5, 1\}$ and computed the corresponding macro-F1 and accuracy scores.

Notably, ordinal regression cannot be evaluated solely based on such performance metrics, as this may obscure the benefits of producing fine-grained relevance scores. In disaster scenarios, it can be highly beneficial for emergency responders to have granular filtering options for reducing the amount of information displayed in an operations control system (Zukunftsforum Öffentliche Sicherheit, 2023). We therefore also assessed the granularity of model outputs, i.e. whether the model produces a diverse range of relevance scores that support such filtering in practice. Specifically, we measured the number of unique predicted values, differential entropy, and the Gini index. The number of unique predicted values captures how many distinct levels of relevance a model produces. Differential entropy quantifies how widely the predicted scores are spread across the range of possible values (Shannon, 1948). It is not bounded in the positive or negative direction, but with input values in a fixed range, higher differential entropy indicates that the model produces dispersed predictions and covers more of the available scale (Cover and Thomas, 2006). Conversely, the Gini index G measures how concentrated the predictions are, ranging from 0 (perfectly uniform distribution) to 1 (all predictions concentrated at a single value) (Gini, 1912). It can be computed directly from continuous predictions by constructing a Lorenz curve over their sorted values (Gastwirth, 1972).

3.5 Data and Software Availability

All experiments were implemented in Python (van Rossum, 1995). The evaluation data is based on the dataset introduced by Hanny et al. (2025), which was collected via the X v1.1 and v2 Application Programming Interface (API) endpoints. It can only be shared as a list of post IDs in accordance with X's developer policies[2]. Both the source code and datasets for the research presented in this paper are publicly available on Zenodo under 10.5281/zenodo.18739246 and on GitHub[3]. The shared data includes post IDs, corresponding relevance labels, and all derived features, which do not directly reflect raw content. To enable full reproducibility, we additionally provide the softmax outputs of the fine-tuned TwHIN-BERT classifier presented by Hanny et al. (2025).

[2] https://developer.x.com/en/developer-terms/policy

[3] https://github.com/IT-U/GSAI_PUBLIC_Multimodal_Relevance_Regression

4 Results

We begin by comparing the classification results obtained using decay-based proximity features against the manual baseline, followed by the presentation of the ordinal regression outcomes.

4.1 Spatio-temporal Decay

Figure 2 illustrates the mean post relevance score within each quantile bin of our training data, and how it decreased with increasing geographic and temporal distance from the disaster impact sites. For the spatial dimension, all four fitted functions (linear, exponential, inverse, and stretched exponential) demonstrated a negative relationship between geographic distance and mean relevance score, with R^2 values between 0.122 and 0.296. The stretched exponential function achieved the highest R^2, indicating a low–moderate fit to the quantile-binned data. For the temporal dimension, the overall relationship was also negative but stronger, with R^2 values ranging from 0.365 to 0.749. The stretched exponential function again achieved the highest R^2.

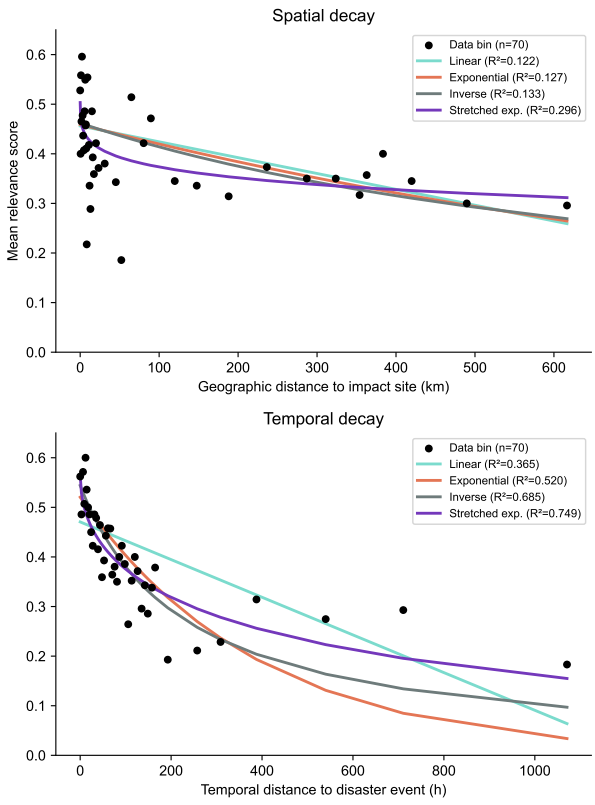

Figure 2. Spatial and temporal decay functions fitted on the binned training data using non-linear least squares optimisation.

Based on these results, we used the fitted stretched exponential decay functions to transform both spatial and temporal distance values into decay-based proximity features. These were subsequently integrated into the

partial stacking classification framework. Figure 3 summarises the resulting classification performance across different feature configurations and meta-learners. The manually engineered baseline achieved an average macro F1 score of 0.740 and an average accuracy of 0.774, with the gradient boosting model reaching the highest individual scores (macro F1: 0.814, accuracy: 0.832). The spatial or temporal decay features alone increased the mean scores but resulted in slightly lower best-case values compared to the baseline. When spatial and temporal decay features were combined, both metrics improved. The best overall performance was achieved when spatio-temporal decay features were complemented with contextual event and location encodings, resulting in a mean macro F1 of 0.796 and a mean accuracy of 0.812, with the random forest model achieving the highest individual scores (macro F1: 0.823, accuracy: 0.837). Across all decay-based configurations, the interquartile ranges were considerably narrower than for the manual baseline. Notably, the exclusively temporal decay configuration occupied largely non-overlapping score ranges for both accuracy and macro F1 compared to the configurations involving both spatial and temporal decay.

Table 3 provides an overview of the average macro F1 score and the respective variability for each feature configuration across the six meta-learners. The manually engineered baseline showed high instability, with a standard deviation of 0.096 and a wide performance range of 0.240, indicating strong sensitivity to the chosen classifier. Introducing spatial or temporal decay features reduced this variability. Spatial decay lowered the dispersion considerably (std. 0.034, range 0.093, CV 4.4%), while temporal decay yielded the most stable performance overall (std. 0.009, range 0.023, CV 1.1%). Combining both spatial and temporal decay achieved the highest mean macro F1 (0.796) while maintaining low variability. Adding event and location encodings preserved the high mean score, increased variability marginally, but resulted in a slight improvement in best-case macro F1, reaching up to 0.823 (c.f. Figure 3).

Table 3. Mean value, standard deviation, min-max range and CV of the macro F1 score across all six meta-learning classifiers, given different feature configurations. The best scores are marked in bold. Arrows denote whether lower (↓) or higher (↑) values are more desired.

Configuration	Mean ↑	Std. ↓	Range ↓	CV ↓
Manual baseline	0.740	0.096	0.240	13.0%
Spatial decay	0.776	0.034	0.093	4.4%
Temporal decay	0.769	**0.009**	**0.023**	**1.1%**
Spatio-temporal decay	**0.796**	0.010	0.030	1.3%
Spatio-temporal decay + event/loc encodings	**0.796**	0.015	0.038	1.8%

4.2 Ordinal Regression

Building on the previous experiments, we reformulated the relevance assessment task as an ordinal regression problem. All regression models were trained using the same feature configuration as in the previous section, combining spatio-temporal decay features with contextual event and location encodings, and the softmax outputs of the fine-tuned text classifier. Table 4 summarises the regression outcomes together with the top-performing classification results from prior experiments. Among the regression models, SVR and ridge regression achieved the lowest RMSE values (0.231 and 0.232) and the highest R^2 goodness-of-fit scores (0.617 and 0.615). kNN regression showed comparable performance, while the tree-based regressors (random forest and gradient boosting) reached notably higher RMSE (0.265 and 0.287) and lower R^2 scores (0.495 and 0.410). In contrast, the classification models reached the lowest MAE values (0.094 and 0.095) and the highest Spearman's ρ correlation (0.790 and 0.793). After rounding regression predictions to the nearest class, the macro F1 and accuracy scores of regression were close to those of the best-performing classification models, with the best scores being achieved by SVR and kNN.

Table 5 compares the granularity of the predicted relevance scores produced by the best-performing classifier (random forest) and regressor (SVR) from above, selected based on macro F1 score and RMSE. Naturally, classification yielded only three unique values, reflecting the three discrete class labels. The regressor, in contrast, produced 914 distinct prediction values. The regression outputs also exhibited a higher differential entropy (–0.515). For the classifier, differential entropy could not be computed directly because when many predictions take on exactly the same value, the underlying formula tries to evaluate the logarithm of zero (Cover and Thomas, 2006). To nonetheless obtain numerical estimates, we added small amounts of uniform jitter in the ranges $[0, 10^{-12}]$ or $[0, 10^{-14}]$ to the classifier's predictions. This resulted in entropy scores between –22.834 and –26.843. With smaller or no jitter, the estimate converged towards $-\infty$. The Gini index provided more stable results for both methods. Here, regression resulted in a lower concentration value (0.378) compared to classification (0.441).

Figure 4 depicts the difference in prediction granularity between the random forest classifier and the SVR regressor. The classifier produced three distinct peaks at 0, 0.5, and 1, corresponding directly to the three discrete class labels. In contrast, the SVR model generated a continuous range of scores across the value range $[0, 1]$. These regression predictions were slightly more concentrated in three regions – between 0 and 0.1, around 0.5, and between 0.9 and 0.95 -— but also included many intermediate values that were not reflected in the classification results.

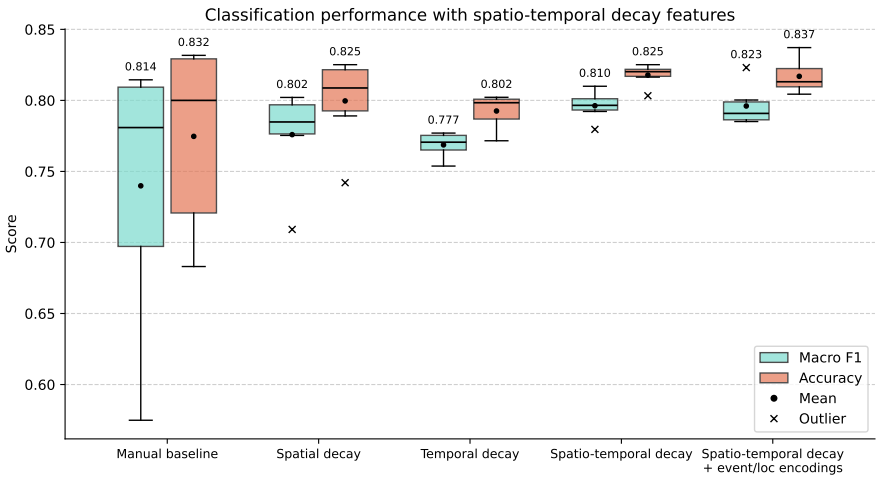

Figure 3. Distribution of classification performance for our six evaluated meta-learning classifiers on the text-based softmax probabilities concatenated with spatio-temporal decay features.

Table 4. Evaluation metrics for classification and regression meta-learners. The depicted classification methods are the top-performing configurations from Sec. 4.1. Regression methods are all based on decay features. Macro F1 and accuracy for regression were computed by rounding the predicted relevance scores. The best values are marked in bold, the second-best values are underlined. Arrows denote whether lower (↓) or higher (↑) values are better.

Model	Task	RMSE ↓	MAE ↓	R^2 ↑	ρ ↑	M-F1 ↑	ACC ↑
Gradient boosting	Classification (baseline)	0.242	<u>0.095</u>	0.580	**0.793**	<u>0.814</u>	<u>0.832</u>
Random forest	Classification (decay-based)	0.244	**0.094**	0.572	<u>0.790</u>	**0.823**	**0.837**
SVR	Regression	**0.231**	0.147	**0.617**	0.771	0.781	0.812
Ridge regression	Regression	<u>0.232</u>	0.146	<u>0.615</u>	0.765	0.767	0.797
kNN	Regression	0.234	0.130	0.605	0.765	0.782	0.810
Random forest	Regression	0.265	0.196	0.495	0.681	0.697	0.737
Gradient boosting	Regression	0.287	0.190	0.410	0.674	0.703	0.739

Table 5. The number of unique values, differential entropy and the Gini index for the best-performing meta-learning classifier (random forest) and regressor (SVR), both using spatio-temporal decay features, contextual encodings, and softmax outputs of the fine-tuned text classifier. The best values are marked in bold. Arrows denote whether lower (↓) or higher (↑) scores indicate higher prediction granularity.

Method	Unique values ↑	Entropy ↑	Gini index ↓
Classification	3	$-\infty$	0.441
Regression	**914**	**-0.515**	**0.378**

To assess the practical implications of regression for an operations control system, we further compared how many posts in our test dataset were assigned scores ≥ 0.9. The random forest classifier yielded 188 posts (21%) in this range, whereas the regressor produced only 96 such scores (10%). In qualitative terms, high regression scores typically corresponded to reports of damage, urgent rescue needs, or on-site observations, while lower scores captured peripheral commentary. Figure 5 illustrates these differences for the subset of posts related to the 2021 Ahr Valley floods in Germany. The regression map shows smooth, continuous variations in predicted relevance and highlights localised clusters of highly relevant posts along the Ahr River. In contrast, the classification map exhibits sharp category boundaries and cannot reflect gradual relevance differences between spatially adjacent posts. To depict this even more clearly, Table 6 shows some exemplary posts and their respective relevance scores.

5 Discussion

In the following, we interpret the experimental findings and critically examine the limitations of our methodology.

5.1 Discussion of the Results

Our results show that post relevance generally decreases with increasing spatial and temporal distance, following a stretched exponential decay pattern. This is consistent with observations by Kottwitz et al. (2023). The decay effect was substantially more pronounced for temporal distance ($R^2 = 0.749$) than for geographic distance ($R^2 =$

Table 6. Predicted relevance scores for selected posts from the test data regarding the 2021 Ahr Valley floods. The examples were translated from German to English for visualisation. Scores might be influenced by geographic and temporal factors that are not explicitly depicted.

Text	Class label	Relevance score
Please share!! The dam in #Wuppertal has collapsed! #Disaster http	Related and relevant	0.92
This is all that's left of my basement… #flood #ahr @ Bad Neuenahr-Ahrweiler http	Related and relevant	0.83
This is in #Sinzig where tragically 12 residents at this home for disabled people died as the flood waters rose http	Related and relevant	0.64
If you thought 2021 would be more relaxed and then the #flooding just screws everything up—strength to the victims	Related but not relevant	0.61
!! Please share and spread the word. !! Official and verified information about the situation in the crisis area #Ahrweiler is now also available at: http #Flooding #Ahr #InfoTweet	Related but not relevant	0.55
I can't even put my shock into words... #monumentalSystemFailure #floodDisaster	Related but not relevant	0.40

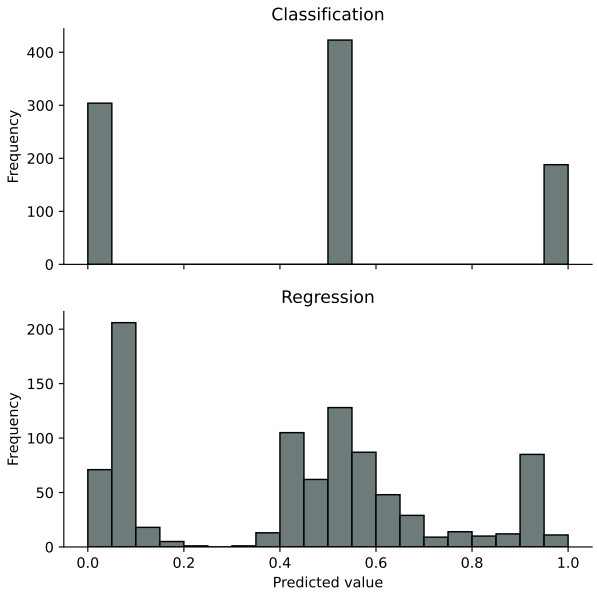

Figure 4. Histogram of relevance prediction best-performing meta-learning classifier (random forest) and regressor (SVR), using spatio-temporal decay features, contextual encodings, and softmax outputs of the fine-tuned text classifier. Each bin has an interval range of 0.05.

0.296). This is in line with prior work demonstrating that disaster-related posting activity is highly concentrated around the time of the event (Havas and Resch, 2021). Spatial patterns, by contrast, can be more diffuse because users often post about disasters from locations far removed from the affected areas (Schmidt et al., 2025). Moreover, spatial and temporal proximity alone do not determine whether a post is relevant. As emphasised by Vieweg et al. (2010), the semantic content of a post remains crucial. The observed variability in relevance across time and space is therefore expected, and the differing R^2 values reflect the realistic interplay between proximity and content.

Replacing the manually engineered features of Hanny et al. (2025) with spatio-temporal decay features led to significantly more stable classification performance across model types. This is likely because the decay features explicitly encode the spatio-temporal relevance patterns documented in prior work (e.g. Kottwitz et al., 2023), rather than requiring the models to learn these relationships from scratch. In other words, the decay functions introduce an inductive bias grounded in geography that guides the models toward realistic proximity–relevance relationships. Consequently, the decay-based proximity features potentially also provide a closer approximation of underlying post relevance. Our analysis further indicates that spatial decay was more critical for model performance than temporal decay. In particular, performance distributions of the temporal decay and spatio-temporal decay configurations were largely non-overlapping across all evaluated meta-learners. This suggests that incorporating spatial decay features provided the most discriminative power, enabling clearer separation between the three relevance classes. The additional contribution of temporal decay on top was comparatively small. The pattern was consistent with the geographically localised nature of disaster events, where location provided strong discriminative power, whereas time alone was insufficient as a filtering criterion.

When viewing relevance assessment as an ordinal regression problem, we observed that regression models outperformed classification in terms of RMSE and R^2, indicating that they captured the variability of the underlying relevance labels more effectively and produced smaller prediction errors, when weighing large deviations higher. At the same time, classification models achieved lower MAE and higher Spearman's ρ. This behaviour is caused by the fundamental difference of discrete versus continuous-valued predictions: Exact class matches yield an error of zero, and rank agreement is easier to attain when the target space consists of only three discrete levels. For the same reason, macro-F1 and accuracy, which are both defined for discrete class predictions, decreased slightly when continuous regression outputs were rounded to the nearest class label.

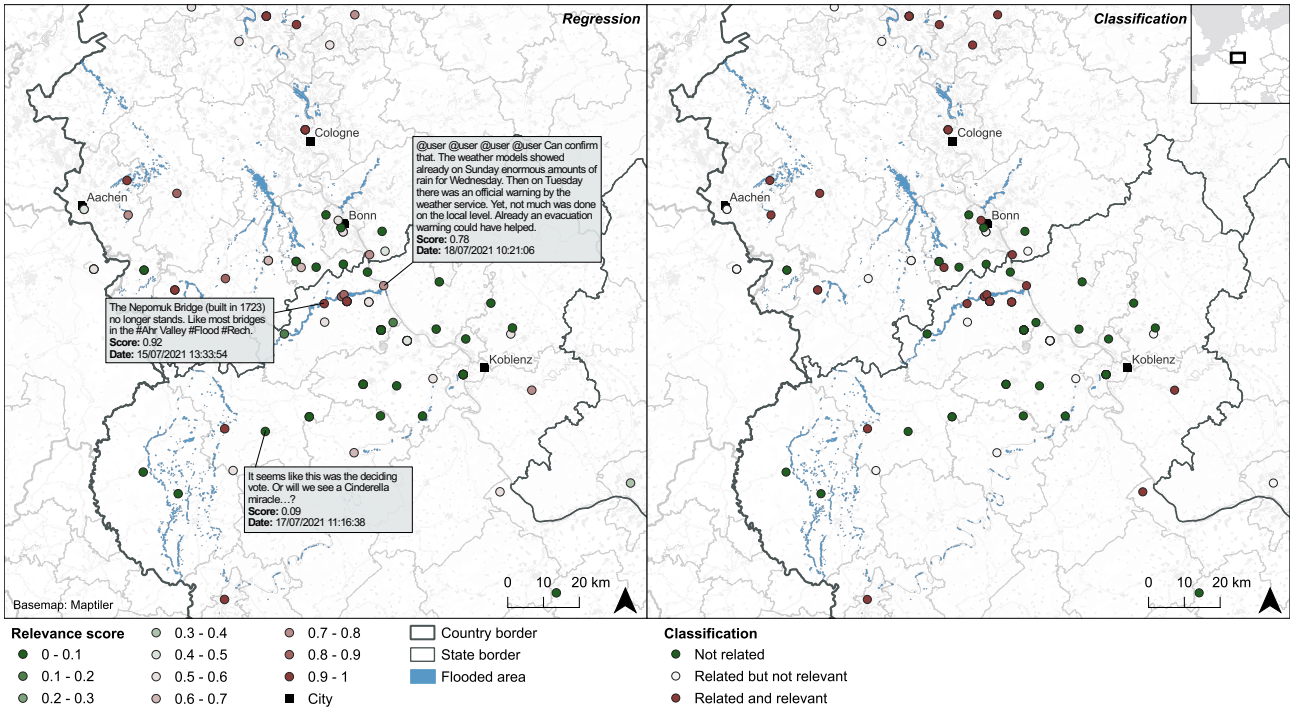

Figure 5. Qualitative comparison of relevance regression versus classification on the test data regarding the 2021 Ahr Valley floods. A closer inspection of posts along the flooded areas highlights the contrasting results of both approaches.

However, regression demonstrated a clear practical advantage, namely that its predictions were substantially more granular. For instance, the SVR regression outputs covered the entire $[0,1]$ range, with 914 unique values, and yielded higher differential entropy and a lower Gini index than classification. Nevertheless, the predicted values remained concentrated around the original class encodings. This was likely induced by the discretely annotated training data, which contains only limited information about the full continuous-valued spectrum of relevance. Despite this constraint, our regression approach enabled an effective distinction between posts that would otherwise fall into the same category. For operations control systems, this granularity can be beneficial, as continuous-valued scores allow emergency responders to flexibly set thresholds (e.g., ≥ 0.8 or ≥ 0.9) and to prioritise information based on nuanced relevance levels, rather than being restricted to three discrete classes. In this sense, ordinal regression complements traditional classification by offering a more nuanced representation of relevance that can support adaptive filtering and thereby aid decision-making in real-world disaster response.

5.2 Discussion of the Methodology

Despite the demonstrated advantages of spatio-temporal decay features and ordinal regression, several methodological limitations remain. First, our work builds directly on the dataset and partial stacking framework introduced by Hanny et al. (2025), as this is currently the only available multilingual, multimodal and ordinally labelled dataset for disaster-related relevance assessment of social media posts. However, both the training and test sets are relatively small, containing 3,659 and 915 samples, respectively. This limits the robustness of the learned decay functions, as geographic references in social media posts may be imprecise and posting times may be affected by unknown delays (Serere et al., 2023; Havas and Resch, 2021), introducing a degree of noise into our estimation of spatio-temporal decay. As a result, our ability to analyse decay effects across individual disaster types or geographic regions is constrained. While we were able to identify clear overall decay trends, a more granular investigation would require significantly larger datasets. Although our decay functions were derived from established formulations in previous research (Kottwitz et al., 2023), future work should explore how machine learning methods could be used to learn similarly robust decay representations. In addition, future research could investigate more sophisticated distance formulations beyond simple proximity by incorporating contextual factors such as accessibility, infrastructure, and characteristics of the physical environment.

At the same time, although recent studies suggest that adequate performance can still be achieved with relatively small training sets (Majdik et al., 2024), larger and more diverse datasets would likely also improve the stability and generalisability of our regression approach. As noted in Sec. 5.1, it notably exhibited pronounced density peaks around the original discrete relevance levels $\{0, 0.5, 1\}$. This suggests that the model still tends to reproduce these discrete boundaries to a degree, even when trained

in a continuous-valued regression setting. Incorporating datasets with alternative and more fine-grained ordinal labelling schemes (e.g. Blomeier et al., 2024) could help mitigate this effect. It would also alleviate the constraints imposed by the small size of our dataset. Future work may additionally explore continuous annotation schemes (e.g. Parde and Nielsen, 2017), which would enable models to capture subtle relevance differences much more effectively and reduce the dependency of regression outputs on predefined categories.

It is further worth noting that all models were trained and evaluated on data from the same set of events, which limits conclusions about the transferability of our approach. Follow-up research should examine its ability to generalise to unseen events and geographic regions. Simultaneously, our evaluation was inherently multi-dimensional, combining measures of prediction accuracy and output granularity. Future studies could formalise these aspects into a single operational ranking metric, allowing for a more user-oriented model comparison.

In terms of model architectures, we also experimented with small neural networks, but these consistently underperformed relative to simpler models such as ridge regression or logistic regression, while requiring substantially more hyperparameter tuning. Developing more effective neural architectures for multimodal relevance regression, therefore, remains an open research challenge. As noted by Hanny et al. (2025), cross-modal transformer models represent a promising direction for integrating text with spatio-temporal features. Finally, our study relied exclusively on text-based softmax outputs and spatio-temporal features. This allowed for a controlled investigation of decay effects and ordinal regression. However, integrating additional modalities such as imagery, user metadata, or network features could further enhance predictive performance. Exploring such multimodal extensions, therefore, represents an important avenue for future research.

6 Conclusions

This study examined how spatio-temporal decay transformations and ordinal regression can enhance the assessment of post relevance in disaster-related social media communication. Regarding **RQ1**, we found that distance-decay transformations substantially improved the robustness of spatio-temporal features for both classification and regression. The stretched exponential function best captured the observed non-linear decline in relevance with increasing spatial and temporal distance, tying in with the findings of Kottwitz et al. (2023). When integrated with text-based features and contextual encodings in a partial stacking framework, these decay-based proximity features improved both prediction stability as well as average and best-case performance compared to manually engineered baseline features. This suggests that decay-based representations provide a more realistic and generalisable way to model spatial and temporal influence on content relevance.

For **RQ2**, we reformulated relevance assessment as an ordinal regression task based on numerically encoded labels. Regression models, particularly SVR, ridge regression, and kNN, achieved lower RMSE and higher R^2 than classification models. Although classification retained advantages in MAE, Spearman's ρ, and macro-F1 after discretisation, regression produced substantially more granular outputs, with 914 unique values across the possible value range of $[0, 1]$. This granularity enables flexible thresholding and thereby more effective prioritisation of information derived from social media. A qualitative comparison for the 2021 Ahr Valley floods supports these findings: Regression outputs highlighted gradual variations in situational importance that discrete classification could not capture.

Overall, our results show that relevance in disaster-related social media data is both spatially dependent and continuous-valued in nature. Distance-decay transformations can significantly enhance the robustness of spatio-temporal features, while ordinal regression provides much finer-grained relevance estimates compared to discrete classification. Within a broader research context, the study advances the foundations for GeoAI-based disaster response by bridging the gap between categorical and human-like, graded assessments of relevance, contributing to the broader goal of turning geo-referenced social media data into actionable, decision-ready knowledge.

Acknowledgements

. All intellectual and creative work, including the analysis, interpretation of data, and writing the draft of the manuscript, is original and has been conducted by the authors without Artificial Intelligence (AI) assistance. However, the authors declare that they have used generative AI, specifically GPT-5, for language editing in the revision stage of this manuscript.

. This work 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). This research has also received funding from the Austrian Research Promotion Agency (FFG) through the project MOSAIK (Grant Number 926200).

. The authors declare no competing interests.

. Conceptualisation, D.H., A.K., and E.J.; methodology, D.H., A.K., and E.J.; software, D.H., and A.K.; validation, D.H., and A.K.; formal analysis, D.H., and A.K.; investigation, D.H., and A.K.; resources, B.R.; data curation, D.H., S.S., and B.R;

writing—original draft preparation, D.H.; writing—review and editing, D.H., A.K., E.J., S.S. and B.R.; visualisation, D.H., and S.S.; supervision, B.R.; project administration, B.R.; funding acquisition, B.R.

References

Acikara, T., Xia, B., Yigitcanlar, T., and Hon, C.: Contribution of Social Media Analytics to Disaster Response Effectiveness: A Systematic Review of the Literature, Sustainability, 15, 8860, https://doi.org/10.3390/su15118860, 2023.

Adwaith, D., Abishake, A. K., Raghul, S. V., and Sivasankar, E.: Enhancing Multimodal Disaster Tweet Classification Using State-of-the-Art Deep Learning Networks, Multimedia Tools and Applications, 81, 18 483–18 501, https://doi.org/10.1007/s11042-022-12217-3, 2022.

Agrawal, T.: Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient, Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4842-6579-6, 2021.

Blomeier, E., Schmidt, S., and Resch, B.: Drowning in the Information Flood: Machine-Learning-Based Relevance Classification of Flood-Related Tweets for Disaster Management, Information, 15, 149, https://doi.org/10.3390/info15030149, 2024.

Brown, C. E.: Coefficient of Variation, in: Applied Multivariate Statistics in Geohydrology and Related Sciences, edited by Brown, C. E., pp. 155–157, Springer, Berlin, Heidelberg, https://doi.org/10.1007/978-3-642-80328-4_13, 1998.

Caragea, C., Adrian Silvescu, and Tapia, A. H.: Identifying Informative Messages in Disaster Events Using Convolutional Neural Networks, in: Proceedings of the ISCRAM 2016 Conference, ISCRAM Association, Rio de Janeiro, Brazil, 2016.

Chai, T. and Draxler, R. R.: Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? – Arguments against Avoiding RMSE in the Literature, Geoscientific Model Development, 7, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014, 2014.

Chen, J., Li, S., Dai, B., and Zhou, G.: Active Learning for Age Regression in Social Media, in: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, edited by Sun, M., Huang, X., Lin, H., Liu, Z., and Liu, Y., pp. 351–362, Springer International Publishing, Cham, https://doi.org/10.1007/978-3-319-47674-2_29, 2016.

Chicco, D., Warrens, M. J., and Jurman, G.: The Coefficient of Determination R-squared Is More Informative than SMAPE, MAE, MAPE, MSE and RMSE in Regression Analysis Evaluation, PeerJ Computer Science, 7, e623, https://doi.org/10.7717/peerj-cs.623, 2021.

Cortes, C. and Vapnik, V.: Support-Vector Networks, Machine Learning, 20, 273–297, https://doi.org/10.1007/BF00994018, 1995.

Cover, T. M. and Thomas, J. A.: Elements of Information Theory, Wiley-Interscience, Hoboken, N.J, 2nd ed edn., 2006.

Cox, D. R.: The Regression Analysis of Binary Sequences, Journal of the Royal Statistical Society. Series B (Methodological), 20, 215–242, 1958.

de Albuquerque, J. P., Herfort, B., Brenning, A., and Zipf, A.: A Geographic Approach for Combining Social Media and Authoritative Data towards Identifying Useful Information for Disaster Management, International Journal of Geographical Information Science, 29, 667–689, https://doi.org/10.1080/13658816.2014.996567, 2015.

Dennis Jr., J. E. and Welsch, R. E.: Techniques for Nonlinear Least Squares and Robust Regression, Communications in Statistics - Simulation and Computation, 7, 345–359, https://doi.org/10.1080/03610917808812083, 1978.

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A., and Vapnik, V.: Support Vector Regression Machines, in: Advances in Neural Information Processing Systems, vol. 9, MIT Press, 1996.

Fisher, R. A.: Statistical Methods for Research Workers, in: Breakthroughs in Statistics: Methodology and Distribution, edited by Kotz, S. and Johnson, N. L., pp. 66–70, Springer, New York, NY, https://doi.org/10.1007/978-1-4612-4380-9_6, 1992.

Fix, E. and Hodges, J. L.: Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties, International Statistical Review / Revue Internationale de Statistique, 57, 238–247, https://doi.org/10.2307/1403797, 1989.

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine., The Annals of Statistics, 29, 1189–1232, https://doi.org/10.1214/aos/1013203451, 2001.

Gastwirth, J. L.: The Estimation of the Lorenz Curve and Gini Index, The Review of Economics and Statistics, 54, 306–316, https://doi.org/10.2307/1937992, 1972.

Gini, C.: Variabilità e Mutabilità: Contributo Allo Studio Delle Distribuzioni e Delle Relazioni Statistiche. [Fasc. I.], Studi Economico-Giuridici Pubblicati per Cura Della Facoltà Di Giurisprudenza Della R. Università Di Cagliari, Tipogr. di P. Cuppini, 1912.

Hanny, D., Schmidt, S., Gandhi, S., Granitzer, M., and Resch, B.: A Multimodal GeoAI Approach to Combining Text with Spatiotemporal Features for Enhanced Relevance Classification of Social Media Posts in Disaster Response, Big Earth Data, 0, 1–45, https://doi.org/10.1080/20964471.2025.2572140, 2025.

Havas, C. and Resch, B.: Portability of Semantic and Spatial–Temporal Machine Learning Methods to Analyse Social Media for near-Real-Time Disaster Monitoring, Natural Hazards, 108, 2939–2969, https://doi.org/10.1007/s11069-021-04808-4, 2021.

Ho, T. K.: Random Decision Forests, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 1, pp. 278–282, IEEE Comput. Soc. Press, Montreal, Que., Canada, https://doi.org/10.1109/ICDAR.1995.598994, 1995.

Hoerl, A. E. and Kennard, R. W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, 12, 55–67, https://doi.org/10.2307/1267351, 1970.

Huang, X., Wang, C., and Li, Z.: Linking Picture with Text: Tagging Flood Relevant Tweets for Rapid Flood Inundation Mapping, Proceedings of the ICA, 2, 1–6, https://doi.org/10.5194/ica-proc-2-45-2019, 2019.

Imran, M., Elbassuoni, S., Castillo, C., Diaz, F., and Meier, P.: Practical Extraction of Disaster-Relevant Information from Social Media, in: Proceedings of the 22nd International Conference on World Wide Web, pp. 1021–1024, ACM, Rio de Janeiro Brazil, https://doi.org/10.1145/2487788.2488109, 2013.

Kanaparthi, S. D., Patle, A., and Naik, K. J.: Prediction and Detection of Emotional Tone in Online Social Media Mental Disorder Groups Using Regression and Recurrent Neural Networks, Multimedia Tools and Applications, 82, 43 819–43 839, https://doi.org/10.1007/s11042-023-15316-x, 2023.

Kaufhold, M.-A., Bayer, M., and Reuter, C.: 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, 102 132, https://doi.org/10.1016/j.ipm.2019.102132, 2020.

Koshy, R. and Elango, S.: Multimodal Tweet Classification in Disaster Response Systems Using Transformer-Based Bidirectional Attention Model, Neural Computing and Applications, 35, 1607–1627, https://doi.org/10.1007/s00521-022-07790-5, 2023.

Kottwitz, M., Zhang, G., and Xu, J.: The Time- and Distance-Decay Effects of Hurricane Relevancy on Social Media: An Empirical Study of Three Hurricanes in the United States, Annals of GIS, 29, 469–484, https://doi.org/10.1080/19475683.2023.2236678, 2023.

Lewis, D. D.: Evaluating Text Categorization I, in: Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991, 1991.

Lewis, D. D.: Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval, in: Machine Learning: ECML-98, edited by Nédellec, C. and Rouveirol, C., pp. 4–15, Springer, Berlin, Heidelberg, https://doi.org/10.1007/BFb0026666, 1998.

Madichetty, S. and Sridevi, M.: Detecting Informative Tweets during Disaster Using Deep Neural Networks, in: 2019 11th International Conference on Communication Systems & Networks (COMSNETS), pp. 709–713, IEEE, Bengaluru, India, https://doi.org/10.1109/COMSNETS.2019.8711095, 2019.

Madichetty, S., Muthukumarasamy, S., and Jayadev, P.: Multi-Modal Classification of Twitter Data during Disasters for Humanitarian Response, Journal of Ambient Intelligence and Humanized Computing, 12, 10 223–10 237, https://doi.org/10.1007/s12652-020-02791-5, 2021.

Majdik, Z. P., Graham, S. S., Shiva Edward, J. C., Rodriguez, S. N., Karnes, M. S., Jensen, J. T., Barbour, J. B., and Rousseau, J. F.: Sample Size Considerations for Fine-Tuning Large Language Models for Named Entity Recognition Tasks: Methodological Study, JMIR AI, 3, e52 095, https://doi.org/10.2196/52095, 2024.

Mitchell, T. M.: Machine Learning, McGraw-Hill Series in Computer Science, McGraw-Hill, New York, 1997.

Ngo-Ye, T. L. and Sinha, A. P.: The Influence of Reviewer Engagement Characteristics on Online Review Helpfulness: A Text Regression Model, Decision Support Systems, 61, 47–58, https://doi.org/10.1016/j.dss.2014.01.011, 2014.

Nguyen, D. Q., Vu, T., and Tuan Nguyen, A.: BERTweet: A Pre-Trained Language Model for English Tweets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, edited by Liu, Q. and Schlangen, D., pp. 9–14, Association for Computational Linguistics, Online, https://doi.org/10.18653/v1/2020.emnlp-demos.2, 2020.

Olteanu, A., Vieweg, S., and Castillo, C.: 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, CSCW '15, pp. 994–1009, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/2675133.2675242, 2015.

Opitz, J.: A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice, Transactions of the Association for Computational Linguistics, 12, 820–836, https://doi.org/10.1162/tacl_a_00675, 2024.

Ouyang, F.-s., Guo, B.-l., Ouyang, L.-z., Liu, Z.-w., Lin, S.-j., Meng, W., Huang, X.-y., Chen, H.-x., Qiu-gen, H., and Yang, S.-m.: Comparison between Linear and Nonlinear Machine-Learning Algorithms for the Classification of Thyroid Nodules, European Journal of Radiology, 113, 251–257, https://doi.org/10.1016/j.ejrad.2019.02.029, 2019.

Parde, N. and Nielsen, R.: Finding Patterns in Noisy Crowds: Regression-based Annotation Aggregation for Crowdsourced Data, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, edited by Palmer, M., Hwa, R., and Riedel, S., pp. 1907–1912, Association for Computational Linguistics, Copenhagen, Denmark, https://doi.org/10.18653/v1/D17-1204, 2017.

Pearson, K.: The Grammar of Science, Nature, 46, 97–99, https://doi.org/10.1038/046097a0, 1892.

Rainio, O., Teuho, J., and Klén, R.: Evaluation Metrics and Statistical Tests for Machine Learning, Scientific Reports, 14, 6086, https://doi.org/10.1038/s41598-024-56706-x, 2024.

Resch, B., Usländer, F., and Havas, C.: Combining Machine-Learning Topic Models and Spatiotemporal Analysis of Social Media Data for Disaster Footprint and Damage Assessment, Cartography and Geographic Information Science, 45, 362–376, https://doi.org/10.1080/15230406.2017.1356242, 2018.

Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., and Cuenca-Jiménez, P.-M.: A Review on Sentiment Analysis from Social Media Platforms, Expert Systems with Applications, 223, 119 862, https://doi.org/10.1016/j.eswa.2023.119862, 2023.

Rubia, J. M. D. L.: Rice University Rule to Determine the Number of Bins, Open Journal of Statistics, 14, 119–149, https://doi.org/10.4236/ojs.2024.141006, 2024.

Scheele, C., Yu, M., and Huang, Q.: 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, 1721–1743, https://doi.org/10.1080/17538947.2021.1968048, 2021.

Schmidt, S., Friedemann, M., Hanny, D., Resch, B., Riedlinger, T., and Mühlbauer, M.: Enhancing Satellite-Based Emergency Mapping: Identifying Wildfires through

Geo-Social Media Analysis, Big Earth Data, 0, 1–23, https://doi.org/10.1080/20964471.2025.2454526, 2025.

Serere, H. N., Resch, B., and Havas, C. R.: Enhanced Geocoding Precision for Location Inference of Tweet Text Using spaCy, Nominatim and Google Maps. A Comparative Analysis of the Influence of Data Selection, PLOS ONE, 18, e0282 942, https://doi.org/10.1371/journal.pone.0282942, 2023.

Shannon, C. E.: A Mathematical Theory of Communication, The Bell System Technical Journal, 27, 379–423, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x, 1948.

Spearman, C.: The Proof and Measurement of Association between Two Things, The American Journal of Psychology, 15, 72–101, https://doi.org/10.2307/1412159, 1904.

van Rossum, G.: Python Tutorial, Tech. Rep. R 9526, CWI (National Research Institute for Mathematics and Computer Science), NLD, 1995.

Verma, S., Vieweg, S., Corvey, W., Palen, L., Martin, J., Palmer, M., Schram, A., and Anderson, K.: Natural Language Processing to the Rescue? Extracting "Situational Awareness" Tweets During Mass Emergency, Proceedings of the International AAAI Conference on Web and Social Media, 5, 385–392, https://doi.org/10.1609/icwsm.v5i1.14119, 2011.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L.: 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, CHI '10, pp. 1079–1088, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/1753326.1753486, 2010.

Wang, Y., Beck, D., Baldwin, T., and Verspoor, K.: Uncertainty Estimation and Reduction of Pre-trained Models for Text Regression, Transactions of the Association for Computational Linguistics, 10, 680–696, https://doi.org/10.1162/tacl_a_00483, 2022.

Wang, Z. and Ye, X.: Social Media Analytics for Natural Disaster Management, International Journal of Geographical Information Science, 32, 49–72, https://doi.org/10.1080/13658816.2017.1367003, 2018.

Willmott, C. J. and Matsuura, K.: Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance, Climate Research, 30, 79–82, https://doi.org/10.3354/cr030079, 2005.

Wu, Y., Pal, A., Wang, J., and Kant, K.: Incremental Spatial Clustering for Spatial Big Crowd Data in Evolving Disaster Scenario, in: 2019 16th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 1–8, https://doi.org/10.1109/CCNC.2019.8651840, 2019.

Yabe, T., Tsubouchi, K., Fujiwara, N., Sekimoto, Y., and Ukkusuri, S. V.: Understanding Post-Disaster Population Recovery Patterns, Journal of the Royal Society Interface, 17, 20190 532, https://doi.org/10.1098/rsif.2019.0532, 2020.

Yan, D., Zhou, X., Wang, X., and Wang, R.: An Off-Center Technique: Learning a Feature Transformation to Improve the Performance of Clustering and Classification, Information Sciences, 503, 635–651, https://doi.org/10.1016/j.ins.2019.06.068, 2019.

Zhang, B. and Li, C.: Advancing Semantic Textual Similarity Modeling: A Regression Framework with Translated ReLU and Smooth K2 Loss, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, edited by Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., pp. 11 882–11 893, Association for Computational Linguistics, Miami, Florida, USA, https://doi.org/10.18653/v1/2024.emnlp-main.663, 2024.

Zhang, X., Malkov, Y., Florez, O., Park, S., McWilliams, B., Han, J., and El-Kishky, A.: 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, KDD '23, pp. 5597–5607, Association for Computing Machinery, New York, NY, USA, https://doi.org/10.1145/3580305.3599921, 2023.

Zukunftsforum Öffentliche Sicherheit: GRÜNBUCH Lagebild. Interdisziplinäres Lagebild in Echtzeit. Erkenntnisse Und Handlungsempfehlungen Zur Verbesserung Der Lagefrüherkennung Und Der Lagebewältigung., Zukunftsforum Öffentliche Sicherheit e.V. (ZOES), Berlin, 2023.

Appendix B: Hyperparameter Space

Table B1. The hyperparameter space explored during the grid search for each model.

Model	Hyperparameters
Logistic regression	C: $\{0.1, 1, 10\}$, penalty: {l2}, solver: {lbfgs, saga}, max_iter: {1000, 2000}
Ridge regression	α: {0.01, 0.1, 1.0, 10.0}
SVM/SVR	C: $\{0.1, 1, 10\}$, gamma: {scale, auto}, kernel: {rbf}
Random forest	n_estimators: {50, 100, 200}, max_depth: {None, 10, 20}, min_samples_split: {2, 5}
Gradient boosting	n_estimators: {50, 100, 200}, max_depth: {3, 5, 7}, learning_rate: {0.01, 0.1, 0.2}
kNN	n_neighbors: {3, 5, 7, 9}, weights: {uniform, distance}
Gaussian Naive Bayes	var_smoothing: {1e-09, 1e-08, 1e-07}