Environmental and Ecological Statistics
https://doi.org/10.1007/s10651-025-00699-x

Mining social media data to track environmental disaster events

Emiliano del Gobbo[1] · **Luigi Ippoliti**[2] · **Lara Fontanella**[3] · **Barbara Cafarelli**[4]

Received: 23 January 2025 / Revised: 19 November 2025 / Accepted: 12 December 2025

Abstract
The increasing prevalence of social media usage has led to the emergence of mining social media data as a valuable resource for disaster response. Mining their textual data presents opportunities and challenges. Advanced techniques in natural language processing (NLP) and machine learning enable the extraction of relevant information while effectively filtering out noise and misinformation. Real-world cases, such as Hurricane Harvey (2017), Hurricane Ida (2021), Hurricane Milton (2024), and Hurricane Melissa (2025) highlight the important role of social media in coordinating disaster relief efforts and enhancing situational awareness. Challenges include unstructured and ambiguous data, diverse user credibility, and overwhelming data volume. The aim of this research is to develop a methodology that integrates textual classification of social media data, spatial and temporal analysis, and visual analytics to provide rapid responses during natural disasters.

Keywords Natural language processing · Environmental disasters · Social media · Machine learning · Spatial and temporal smoothing

1 Introduction

The advent of social media platforms has revolutionized interpersonal communication, resulting in a substantial surge in the production of textual data on a daily basis. Particularly during crises and natural disasters, social media serves as a crucial real-time information source, aiding disaster responders in making well-informed decisions and optimizing resource allocation. Nevertheless, the vast quantity and diversity of data generated by social media users pose challenges in extracting pertinent and actionable information.

Handling Editor: Luiz Duczmal.

Extended author information available on the last page of the article

Published online: 23 February 2026

The exploration of mining social media textual data for disaster response has emerged as a promising research area aimed at addressing the multifaceted challenges of disaster management. By leveraging advancements in Natural Language Processing (NLP), machine learning, and information retrieval, researchers and practitioners have developed techniques to extract critical information from social media in real-time. These methods can identify key details, such as the locations of individuals in need of assistance, the severity and extent of disaster impacts, and the progress of ongoing response efforts.

The potential of social media mining for disaster response has been demonstrated in several real-world scenarios (Young et al. 2020). For instance, during the 2015 Nepal earthquake, social media played a crucial role in facilitating support and assistance (Bossu et al. 2015; Subba 2016). Platforms such as Twitter and Facebook were widely used by affected individuals, local communities, and relief organizations to share information and organize relief efforts. Similarly, during Hurricane Harvey in 2017, social media users provided real-time updates on stranded individuals' locations, flood conditions, and the availability of shelters and supplies, underscoring the critical role of these platforms in disaster response.

However, using social media data presents several challenges. The data are often unstructured, noisy, and ambiguous, making the extraction of relevant information a complex task. Moreover, social media content originates from a diverse user base, encompassing individuals with varying levels of credibility, expertise, and biases, which raises concerns about the reliability and accuracy of the data. The sheer volume of textual data generated during disasters further complicates timely classification and analysis of crisis-related messages.

The challenge of extracting accurate location information and estimating event origins is particularly pressing for maintaining situational awareness (Suwaileh et al. 2023). The dynamic and unstructured nature of social media posts, coupled with user variability, complicates the task of pinpointing the geographical sources of reported incidents. Additionally, the abundance of noise, misinformation, and ambiguity in these posts exacerbates the difficulty of deriving reliable spatial information. Despite advancements in geotagging and geolocation techniques, achieving satisfactory situational awareness remains elusive due to persistent uncertainties in the granularity and accuracy of spatial data derived from social media.

Content classification presents another critical challenge. Contextual ambiguity often leads to misinterpretation, as identical phrases can convey vastly different meanings depending on the context, thereby hindering accurate categorization. Moreover, the phenomenon of concept drift, where the discourse around disasters evolves dynamically, poses a substantial challenge. This necessitates adaptive models capable of capturing emerging trends. Domain adaptation becomes essential to ensure model generalization for disaster-related texts, as these texts often exhibit unique language patterns and topic distributions specific to crises.

Furthermore, the language of social media differs significantly from other forms of communication, and each platform has distinct linguistic characteristics. For example, X (formerly Twitter) imposes constraints on message length, encouraging brevity and extensive use of hashtags. Addressing these challenges requires sophisticated machine learning methodologies tailored to the specific characteristics of social

media content. Given the urgency and time-sensitive nature of disaster events, these methods must be capable of real-time application without the need for extensive human intervention, manual tuning, or contextual adjustments. By overcoming these challenges, classifiers can efficiently discern and categorize disaster-related social media content, ultimately enhancing disaster response efforts and ensuring more effective crisis management.

This study specifically addresses the challenge of impact assessment during disasters, a critical task for emergency responders and government agencies in gauging the severity and scope of the disaster and efficiently allocating resources. Leveraging state-of-the-art machine learning techniques, the study develops a methodological framework for impact assessment using tweets messages related to Hurricane Ida (U.S. Environmental Protection Agency 2021). Specifically, this research addresses three different issues in tracking disaster events through social media data in the context of real-time disaster monitoring.

We first propose a classification model based on NLP techniques and deep neural networks to grasp the content of messages contained in the Hurricane Ida dataset. To accomplish this task, the model is trained on large textual datasets to enhance semantic awareness, and then fine-tuned using content from other labeled disaster datasets. Essential to this task is the human-annotated disaster-related dataset, IDRISI, which is used for transfer learning purposes.

The study also implements a methodology for detecting geolocation from the textual content of tweets to enhance spatial analysis. Given that geolocation information is directly available in the metadata of only a few tweets, a neural network-based model is adopted to infer this information from the text itself. This model leverages advanced NLP techniques to analyze the content and context of the tweets, effectively estimating their geographical origins.

Finally, we produce smoothed maps that allow us to track the diffusion of the disaster by analyzing the textual content of tweets related to Hurricane Ida. These maps provide an effective visualization of the spatio-temporal distribution and evolution of the hurricane's impacts. By interpreting the geographic references and context embedded in the tweets, the study enhances understanding of how the disaster unfolds across different regions. This approach not only aids in identifying areas most affected by the hurricane but also supports decision-making processes for resource allocation and emergency response planning.

The paper is organized as follows. Section 2 introduces the motivating example of the Hurricane Ida dataset, which consists of geolocated social media posts collected during the 2021 Hurricane Ida disaster. This dataset captures real-time public reactions, situational updates, and spatiotemporal patterns of information dissemination. Section 3 focuses on automatic labeling, presenting the transformer architecture DistilBERT (Sanh et al. 2019) - a smaller and faster variant of BERT (Devlin et al. 2018) - designed to learn contextual relationships between words in text. Section 3.1 details an application of multiclass classification, where DistilBERT is fine-tuned using the IDRISI dataset (Suwaileh et al. 2023), which contains labeled data from multiple disaster events. The model is then tested on a subset of the Hurricane Ida dataset in Section 4, demonstrating its adaptability and performance. The proposed framework integrates state-of-the-art NLP techniques, geolocates tweets using a combination

of location extraction from text and metadata, and facilitates disaster visualization across spatial and temporal dimensions. To further enhance these capabilities, Section 5 introduces for the first time an algorithm for spatial and temporal smoothing of data on maps, enabling effective tracking of disaster dynamics. Finally, Section 6 provides concluding remarks and outlines potential avenues for future research, highlighting opportunities to further develop the proposed framework.

2 Hurricane Ida as motivating example

Numerous literature reviews have highlighted significant advancements in utilizing social media data for disaster response (Landwehr and Carley 2014; Wang and Ye 2018). In particular, researchers have focused their attention towards primary aspects such as temporal and spatial information for accurately pinpointing events, as well as content analysis, which is essential for understanding specific occurrences (Wang and Ye 2018).

The present study aims to analyze the content of tweets posted during Hurricane Ida (Phillips 2021). Hurricane Ida, a powerful Category 4 Atlantic hurricane, ranks as the second-most damaging and intense hurricane to hit the U.S. state of Louisiana, following Hurricane Katrina in 2005. Specifically, Hurricane Ida formed on August 26, 2021, and dissipated on September 4, 2021, caused widespread destruction and had a significant impact on many regions.

The dataset comprises over 1,800,000 tweets, all in English, gathered through searches using keywords such as *#ida*, *'hurricane ida'*, and *hurricaneIda*. During the onset of the crisis, a diverse range of real-time information was shared by individuals directly impacted, those seeking assistance, and volunteers offering help. These messages contain crucial and actionable information that aid in understanding and responding to the situation effectively. Hence, the analysis of tweets related to Hurricane Ida is crucial for understanding the real-time impact and response needs during such a catastrophic event.

This study first involves employing classification techniques to categorize messages, thereby providing insights into the nature of events, their gravity, and the most affected areas (Alam et al. 2021; Castillo 2016; Kumar et al. 2019). Several researchers have introduced deep learning approaches to automatically categorize tweets concerning natural disasters based on their relevance and urgency (Burel and Alani 2018; Alam et al. 2020). These advancements facilitate the rapid identification of critical information, enabling efficient disaster response and resource allocation.

Another challenging aspects of Hurricane Ida data analysis is the extraction of location information, commonly referred to as Location Mention Recognition (LMR). Recognizing the indispensability of accurate location data in providing timely support, researchers have developed models specifically tailored to address this issue (see, for example, Starbird 2012, Imran et al. 2013 and Suwaileh et al. 2023).

In the following, we shall show that the application of NLP tools to tweets related to Hurricane Ida can provide important insights into the disaster's impact and the needs of affected individuals. By leveraging deep learning models for relevance and urgency classification, responders can prioritize tweets that contain critical informa-

tion. The implementation of LMR techniques also ensures that location data can be accurately extracted, facilitating targeted assistance and resource deployment.

3 Using transformers for automatic labelling and impact assessment

Generating accurate labels for automated natural disaster impact assessment remains a primary challenge for model development. Manual annotation by experts is both time-consuming and costly, constraining the amount of disaster data that can be exploited for analysis. Consequently, automated approaches capable of inferring labels from available unstructured data, such as social media posts, represent a crucial step toward scalable and timely impact assessment.

Recent advances in NLP, driven by the introduction of the Transformer architecture (Vaswani et al. 2017), have transformed the capacity to extract meaningful information from textual data. In particular, large pre-trained language models such as the *Bidirectional Encoder Representations from Transformers* (BERT) (Devlin et al. 2018) have achieved remarkable success across a wide range of NLP tasks, including text classification, semantic analysis, and sentiment detection. The core innovation of these models lies in the self-attention mechanism, which allows the model to capture contextual dependencies between words across long textual spans. This enables a richer and more nuanced representation of meaning compared to traditional recurrent architectures.

In this work, we employ the DistilBERT variant (Sanh et al. 2019), a lightweight and computationally efficient version of BERT that retains most of its accuracy while reducing model size and inference time. DistilBERT leverages transfer learning through two stages: a pre-training phase, during which it learns general linguistic patterns from large unlabeled corpora, and a fine-tuning phase, where it adapts to our specific task of tweet classification for hurricane-related damage assessment. This approach allows the model to generalize from limited labelled data while capturing subtle contextual cues indicative of different types and degrees of damage.

A detailed description of the BERT and DistilBERT architectures, including the self-attention and feed-forward mechanisms, is provided in Appendix A.

To fine-tune DistilBERT for tweet classification, we used a domain-specific labeled dataset and added a classifier layer. This method benefits from faster convergence due to pre-trained weights, requires less data for fine-tuning than training from scratch, and allows for versatile application in various downstream tasks. For this research, DistilBERT was fine-tuned for multiclass classification using the IDRISI dataset (see next subsection 3.1), chosen for its labelled data across multiple disaster events and its suitability for real-time social network monitoring. This approach also evaluates the model's ability to adapt to new datasets based on historical event data, with the expectation that disaster-related challenges like fires, disruptions, and injuries exhibit similar semantic patterns across different events.

3.1 IDRISI dataset

IDRISI-RE[1] (Suwaileh et al. 2023) is a dataset that compiles tweets about significant disasters in recent years, including hurricanes, earthquakes, and floods. Some tweets have been manually annotated to indicate their location and type of aid communication, such as infrastructure damage or urgent needs. The dataset comprises 16,448 tweets, each manually classified by disaster event. Table 1 provides an overview of the tweet distribution across different events while the types of labels and their percentage within the dataset, and examples of tweet content are detailed in Table 2. The variety of labels and the range of different disasters represented in the dataset align with the research objective of developing a process to analyze disasters in real-time through social media.

4 Model performance

We evaluated the fine-tuned DistilBERT model's performance using standard classification metrics, including accuracy, precision, recall, and the F1 score, to comprehensively assess its predictive capabilities across disaster-related tweet categories. Accuracy alone may overlook important nuances – particularly in imbalanced datasets - whereas precision, recall, and F1 provide complementary perspectives on the model's ability to correctly identify relevant classes and avoid misclassification. Detailed definitions of these metrics are provided in Appendix B.

Table 1 IDRISI dataset: the variety of different disasters and the number of corresponding labelled tweets

Event	Tweets count
California_wildfires_2018	1040
Canada_wildfires_2016	1040
Cyclone_idai_2019	1040
Ecuador_earthquake_2016	928
Greece_wildfires_2018	776
Hurricane_dorian_2019	1040
Hurricane_florence_2018	1040
Hurricane_harvey_2017	1040
Hurricane_irma_2017	1040
Hurricane_maria_2017	1040
Hurricane_matthew_2016	688
Italy_earthquake_aug_2016	472
Kaikoura_earthquake_2016	992
Kerala_floods_2018	1040
Maryland_floods_2018	344
Midwestern_us_floods_2019	864
Pakistan_earthquake_2019	616
Puebla_mexico_earthquake_2017	1040
Srilanka_floods_2017	368

[1] https://github.com/rsuwaileh/IDRISI/

Table 2 IDRISI dataset. Annotated label categories (left), examples of tweets (center), and percentage of each category

Label	Sample tweet	Percentage
Injured or dead people	Paradise, a town of 27,000 in the Sierra Nevada foothills, was destroyed by a wildfire in California that has killed nine people.	18.83
Infrastructure and utility damage	A major Air Force base was damaged in the Nebraska flood via @ voxdotcom @edspain tried to warn FEMA. #GIS doesn't lie.	14.87
Rescue volunteering or donation effort	Please donate to California Fire Foundation & California Community Fund, two efforts to care for the victims of the #WoolseyFire & #HillFire, and those who are putting their lives on the line to fight it. @CAFireFound: @calfund:	39.94
Displaced people and evacuations	Mandatory evacuation implemented in Victoria County before Hurricane Harvey hits #KSATnews	6.94
Requests or urgent needs	Items needed: toiletries, diapers, baby formula, non-perishbl food, clothing, blankets #HelpforHouston @CityOfBoston	5.69
Missing or found people	One missing in raging flood waters that washed out Ellicott City, Maryland	1.67
Caution and advice	Civil Defence recommends if people feel another long and strong earthquake, they should head for higher ground or as far inland as possible.	12.06

Table 3 Summary of the 10-fold cross-validation results on the IDRISI dataset for the fine-tuned DistilBERT model, reporting average performance metrics across all folds (standard deviations in parentheses).

Accuracy	F1 weighted score	Recall (mean across classes)
$0.907(\pm 0.009)$	$0.907(\pm 0.010)$	$0.856(\pm 0.010)$

Table 4 Category-wise metrics obtained from the 10-fold cross-validation on the IDRISI dataset for the fine-tuned DistilBERT model. Reported values include Accuracy, Recall, and F1 score, with standard deviations in parentheses. Values approaching 1 denote better validation performance.

Category	Accuracy	Recall	F1 score
Injured or dead people	0.970(± 0.014)	0.960(± 0.008)	0.972(± 0.014)
Infrastructure and utility damage	0.911(± 0.017)	0.910(± 0.006)	0.911(± 0.017)
Rescue volunteering or donation effort	0.929(± 0.022)	0.932(± 0.010)	0.929(± 0.022)
Displaced people and evacuations	0.925(± 0.020)	0.889(± 0.017)	0.925(± 0.020)
Requests or urgent needs	0.689(± 0.097)	0.687(± 0.053)	0.689(± 0.097)
Missing or found people	0.716(± 0.101)	0.761(± 0.060)	0.716(± 0.101)
Caution and advice	0.8490(± 0.034)	0.871(± 0.019)	0.849(± 0.034)

The evaluation was conducted using a 10-fold cross-validation scheme to ensure robust and reliable results; a detailed description of this procedure is provided in Appendix C.

Table 3 reports the overall and average performance metrics based on a 10-fold cross-validation procedure. The results indicate robust model generalization, with average accuracy and F1 scores exceeding 0.90 across folds. Performance consistency across categories is further confirmed in Table 4, which shows strong results for most disaster-related classes.

To further evaluate the model's transferability, validation was performed on the subset of IDRISI tweets specifically related to hurricane events. Table 5 presents a

Table 5 Comparison of F1 scores obtained from model validation on the subset of hurricane-related tweets within the IDRISI dataset. The column "F1-H" reports results for the model trained solely on hurricane-related data, while "F1-all" reports those for the model trained on the entire IDRISI dataset.

Category	F1-H	F1-all
Injured or dead people	0.975	0.967
Infrastructure and utility damage	0.947	0.915
Rescue volunteering or donation effort	0.932	0.922
Displaced people and evacuations	0.944	0.901
Requests or urgent needs	0.711	0.722
Missing of found people	0.400	0.500
Caution and advice	0.911	0.832

Table 6 F1 scores for model validation on a 1,000-label sample from the Hurricane Ida dataset.

Category	F1 score
Injured or dead people	0.794
Infrastructure and utility damage	0.765
Rescue volunteering or donation effort	0.485
Displaced people and evacuations	0.860
Requests or urgent needs	0.682
Missing or found people	0.500
Caution and advice	0.763

comparison of the F1 scores obtained when the model was trained exclusively on hurricane-related data versus the complete multi-disaster dataset. The similarity in performance between the two training strategies demonstrates the model's adaptability, with minor differences observed only in underrepresented categories such as *Missing or Found People* and *Injured or Dead People*, likely due to smaller sample sizes.

4.1 Ida classification challenges

For real-time analyses and to extend the classification results obtained by training BERT on the IDRISI dataset, we sought to evaluate the model performance on tweets specifically related to Hurricane Ida. To achieve this, we manually labelled a set of 1,000 tweets pertaining randomly sampled from the Hurrican Ida dataset, which served as our validation dataset.

The results in Table 6 demonstrate the variability in the classifier ability to predict tweet categories from the Hurricane Ida dataset. Categories such as *Displaced People and Evacuations* (F1 score: 0.860) and *Injured or Dead People* (F1 score: 0.794) exhibit strong performance, whereas *Rescue Volunteering or Donation Effort* (F1 score: 0.485) and *Missing or Found People* (F1 score: 0.500) perform notably worse. This disparity highlights a fundamental limitation in the current classification framework: the mismatch between the IDRISI labels used for training and the specific characteristics of Hurricane Ida tweets. Notably, Hurricane Ida introduced new damage subcategories (hereafter referred to as *Other* categories) not encompassed by the IDRISI taxonomy. For instance, tweets addressing urban flooding in densely populated areas such as New York City or infrastructure collapse in suburban regions were underrepresented in the original dataset. Additionally, many tweets

in the Hurricane Ida dataset were either irrelevant or outside the scope of disaster-related classifications.

Table 7 provides further insight by comparing the distribution of true labels in the Hurricane Ida dataset with those misclassified under the *Other* category. Specifically, tweets from the *Other* category were disproportionately misclassified as *Rescue Volunteering or Donation Effort*, with 38.6% of misclassified tweets falling into this group. Similarly, the *Caution and Advice* category, which represents a significant proportion of true labels (57.0%), frequently appears in the *Other* category (48.2%). This suggests that while the classifier effectively captures broad patterns, it struggles to differentiate finer distinctions within high-frequency categories. In contrast, the low misclassification rates for *Injured or Dead People* and *Displaced People and Evacuations* (2.6% and 1.3%, respectively) reflect the more distinct linguistic features of these categories, which align closely with the IDRISI labels. The high rate of misclassification in certain classes presents significant challenges, primarily due to the highly heterogeneous language used in tweets, which varies considerably over time, location, and other contextual factors. This variability complicates real-time analysis when relying on static, pre-existing training data. One potential solution to mitigate this issue is to expand the training dataset to include more examples from the underrepresented categories. However, the fact that many of these misclassified tweets are assigned to two less relevant categories somewhat alleviates the impact of this limitation, as it reduces the overall disruption to the classification process.

Figure 1 illustrates the final distribution of categories within the complete Hurricane Ida dataset, as classified by the fine-tuned DistilBERT model.

5 Tracking Hurricane Ida

This section examines the temporal evolution and spatial diffusion of Hurricane Ida through the lens of Twitter data. By analyzing geo-located tweets, we can trace how information about the hurricane spread over time and across different locations. In particular, we explore the relationship between the hurricane's physical trajectory and the distribution of tweets, highlighting key moments when public attention surged in response to significant events and emergency warnings. Additionally, we consider how the tweet content varied spatially, reflecting the distinct challenges faced by

Table 7 Comparison of label proportions in the Hurricane Ida validation dataset (column "true labels prop") versus the proportions predicted by the model for tweets classified as *Other* (column "Miscl-Other"). The 'Other' category includes 386 tweets, while the remaining categories collectively include 614 tweets.

Category	True labels prop	Miscl-Other
Injured or dead people	0.052	0.026
Infrastructure and Utility Damage	0.160	0.065
Rescue Volunteering or Donation Effort	0.124	0.386
Displaced People and Evacuations	0.063	0.013
Requests or Urgent Needs	0.029	0.023
Missing of Found People	0.002	0.005
Caution and Advice	0.570	0.482

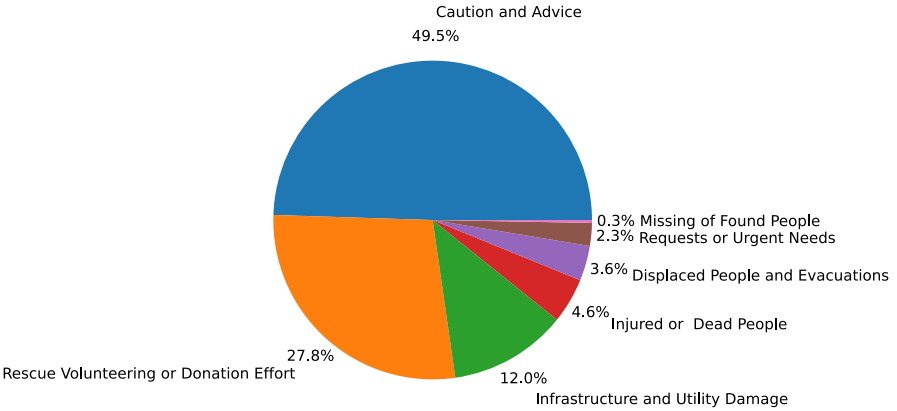

Fig. 1 Composition of the Hurricane Ida dataset based on model classification.

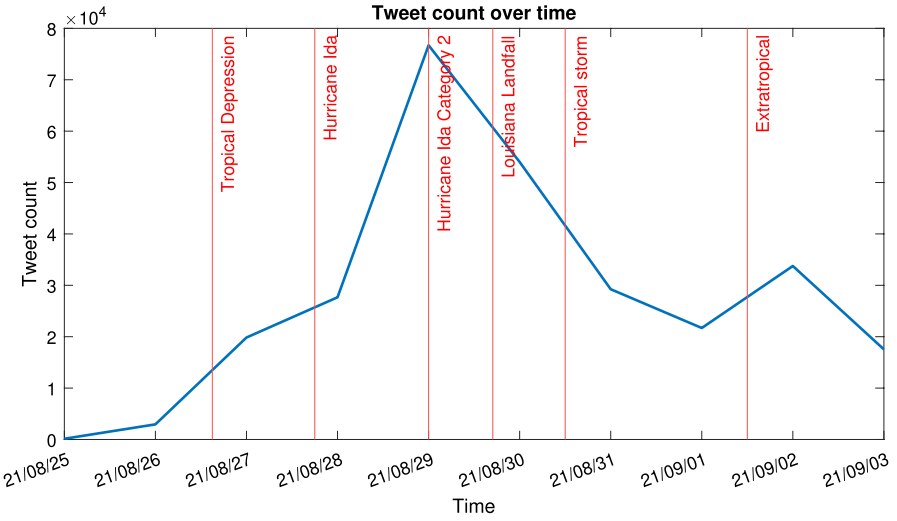

Fig. 2 Daily tweet count from the complete Hurricane Ida dataset, aligned with key events throughout the hurricane's progression.

regions in the hurricane's path. Through this analysis, we gain insights not only into the chronological progression of Hurricane Ida but also into how social media data can serve as a proxy for the public's awareness and response during extreme weather events. The findings offer a deeper understanding of the dynamics of information flow during crises, as well as the role of digital platforms in disaster management.

5.1 Temporal dynamics of Hurricane Ida

Figure 2 highlights key moments of the disaster, showing how the volume of tweets responded in syncrony with the hurricane's progression. Tweet activity increased as Hurricane Ida intensified, reaching a peak when the storm was upgraded to Category

2 and subsequently Category 4, and again during landfall in Louisiana. As the hurricane's severity diminished, tweet volume also gradually declined.

Figure 3 provides additional insights by incorporating tweet classification data, displaying the tweet flow by damage category. For clarity, only the categories with the highest tweet volumes are reported. The flow evolution aligns with reasonable expectations: categories like "Caution and Advice" and "Displaced People and Evacuations" peaked prior to the hurricane's landfall, as warnings and evacuation efforts are critical during this stage. In contrast, categories such as "Injured or Dead People" and "Infrastructure and Utility Damage" peaked shortly after landfall, reflecting the actual devastation caused by the hurricane.

5.2 Spatial distribution of Hurricane Ida

Mapping the spatial distribution of tweets provides valuable insight into both the physical trajectory of Hurricane Ida and the concurrent diffusion of information across affected regions. However, a major challenge in spatial analysis of social media data are that only a small fraction (approximately 2–3%) of tweets contain explicit geolocation metadata. To overcome this limitation, location information can be inferred from the textual content of tweets through a process known as *geoparsing*, which involves recognizing place names (toponyms) and resolving them to geographic coordinates.

In this study, we implemented a two-step geoparsing approach. First, toponym recognition was performed using the *NeuroTPR* model (Wang et al. 2020), which is designed for handling the noisy and informal nature of social media text. Second, toponyms were geocoded by linking them to their corresponding coordinates using a hierarchical dictionary derived from the publicly available U.S. National Address Database (NAD). This process enabled the reconstruction of the spatial footprint of tweet activity associated with Hurricane Ida, offering a clearer view of both the event's progression and the spatial dynamics of online communication during the disaster.

A detailed technical description of the geoparsing workflow, including the NeuroTPR architecture and geocoding procedures, is provided in Appendix D.

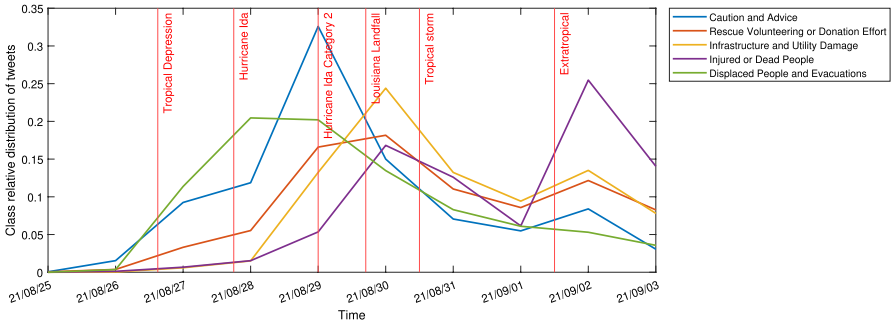

Fig. 3 Daily relative distribution of model-assigned tweet categories from the full Hurricane Ida dataset in relation to key events during the hurricane's progression.

In the hurricane dataset, out of 356,371 tweets, only 6,611 (1.85 %) were originally geolocated. Through the implemented geoparsing method, we were able to classify 84,085 tweets. The combined set of geolocated tweets amounted to 89,025 (23.6 % of all tweets).

The direct representation of geolocated tweets as points often leads to significant overlap, limiting its effectiveness in conveying geographical insights. To address this, we tessellated the U.S. territory into *n* hexagonal cells of equal area, mitigating distortions from the map projection's coordinate reference system. Each cell counted the tweets within its boundaries. However, using social media as a proxy for disaster analysis introduces inherent biases, as tweet volumes tend to be disproportionately higher in densely populated areas, such as cities, compared to rural regions. This disparity is further amplified by geoparsing techniques, which often identify locations at broader administrative levels, such as counties or states, rather than precise geographic coordinates. To address these imbalances and provide a more comprehensive representation of the disaster's impact, we apply a spatial smoothing procedure, as detailed in the following sections.

5.2.1 Spatial smoothing

Various smoothing methods, such as kernel density estimation or moving averages, are applied depending on the data type and the scale of analysis, helping researchers to better interpret spatial phenomena, such as the spread of events across a region (Ruppert et al. 2003). Given the structure of our data, this paper applies a filter designed to smooth data on the vertices of arbitrary undirected graphs with flexible, non-negative weights. Specifically, we use a graph Laplacian filter to capture similarity between signals on adjacent vertices.

Consider a simple and connected undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $|\mathcal{V}| = n$ nodes and $|\mathcal{E}| = m$ edges. We define the *self-looped graph* $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ as the graph obtained by adding a self-loop to each node in $\mathcal{G}$. The node set of $\mathcal{G}$ is denoted by $\{v_1, \ldots, v_n\}$, with each node v_i having degree (*i.e.*, number of neighbours) d_i in $\mathcal{G}$ and $d_i + 1$ in $\tilde{\mathcal{G}}$. Let $\mathrm{A}_s = (a_{ij})$ denote the $(n \times n)$ symmetric spatial adjacency matrix and $\mathrm{D} = diag(d_1, \ldots, d_n)$ the diagonal degree matrix. Since $\mathcal{G}$ is unweighted, we set $a_{ij} = 1$ if nodes v_i and v_j are connected, and $a_{ij} = 0$ otherwise. Consequently, the adjacency and diagonal degree matrices of $\tilde{\mathcal{G}}$ are defined as $\tilde{\mathrm{A}}_s = \mathrm{A}_s + \mathrm{I}$ and $\tilde{\mathrm{D}} = \mathrm{D} + \mathrm{I}$, respectively. The Laplacian matrix of $\tilde{\mathcal{G}}$ is given by $\tilde{\mathrm{L}} = \tilde{\mathrm{D}} - \tilde{\mathrm{A}}_s$, and Z denotes the *n*-dimensional vector of counts within each hexagonal cell.

To achieve smoothing, we assume that data from nodes in close proximity are highly correlated, implying that neighbouring nodes should inform each other's values. Smoothing is then performed via the graph convolution:

$$\begin{aligned} \hat{\mathrm{Z}} &= \tilde{\mathrm{D}}^{-1/2} \tilde{\mathrm{A}}_s \tilde{\mathrm{D}}^{-1/2} \mathrm{Z} \\ &= \big(\mathrm{I} - \widehat{\mathrm{L}}\big) \mathrm{Z} \\ &= \mathrm{SZ} \end{aligned} \quad , \tag{1}$$

where $\tilde{\mathrm{D}}^{-1/2}\tilde{\mathrm{A}}_s\tilde{\mathrm{D}}^{-1/2}$ is the normalized adjacency matrix with added self-loops, $\mathrm{S} = \left(\mathrm{I} - \widehat{\mathrm{L}}\right)$ and $\widehat{\mathrm{L}} = \tilde{\mathrm{D}}^{-1/2}\tilde{\mathrm{L}}\tilde{\mathrm{D}}^{-1/2}$ is the symmetric normalized graph Laplacian matrix of $\tilde{\mathcal{G}}$. Equation (1) implies that the data of neighbour nodes are aggregated and combined with the data Z_i of the current node v_i to form smoothed values:

$$\hat{Z}_i = \frac{Z_i}{d_i + 1} + \sum_{j=1}^{n} \frac{a_{ij}\ Z_j}{\sqrt{(d_i + 1)\ (d_j + 1)}}.$$

In practice, smoothing results can be improved by iterating this process multiple times, as each successive iteration allows a node to incorporate information from nodes that are *k*-hop away, where *k* represents the number of iterations. This iterative aggregation expands the influence range, gradually integrating data from increasingly distant neighbours and refining the smoothing. Consequently, by applying the *k*-th power of the graph convolution matrix S, we capture higher-order information within the graph as follows:

$$\hat{\mathrm{Z}}^{(k+1)} = \left(\mathrm{I} - \widehat{\mathrm{L}}\right)^k\ \mathrm{Z}, \quad k = 1, \ldots, K, \tag{2}$$

where each successive power, *k*, extends the smoothing process.

One crucial and interesting parameter of neighbourhood feature aggregation is the number of smoothing iterations *K*, which controls how much information is being gathered. The choice of *K* is closely related to the structural properties of graphs and has a significant impact on the model performance since, intuitively, too small or too large smoothing iterations may cause under-smoothing or over-smoothing. To provide evidence of over-smoothing, it can be noticed that by continually smoothing the node values with infinite number of propagation in (2), the final smoothed vector $\hat{\mathrm{Z}}^{(\infty)}$ is:

$$\hat{\mathrm{Z}}^{(\infty)} = \left(\mathrm{I} - \widehat{\mathrm{L}}\right)^{\infty}\ \mathrm{Z},$$

and the weight between nodes v_i and v_j is given by:

$$\left(\mathrm{S}^{\infty}\right)_{ij} = \frac{(d_i + 1)^{1/2}\ (d_j + 1)^{1/2}}{2m + n}. \tag{3}$$

Equation (3) shows that S converges to a unique stationary matrix independent of the distance between nodes, and that the final values of $\hat{\mathrm{Z}}$ are over-smoothed and unable to capture the full graph structure information since they only relate with the node degrees of target nodes and source nodes. This can be proved by the fact that, if $\mathrm{V} = (\mathrm{v}_1, \ldots, \mathrm{v}_n)$ and $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_n)$ are the matrices of eigenvectors and eigenvalues of S, respectively, by iterating the smoothing procedure *K* times we have that, by spectral decomposition:

$$\mathrm{S}^K = \mathrm{V}\boldsymbol{\Lambda}\mathrm{V}' \cdots \mathrm{V}\boldsymbol{\Lambda}\mathrm{V}' = \sum_{i=1}^{K} \lambda_i^n \mathrm{v}_i \mathrm{v}_i'.$$

Since S has maximum eigenvalue equal to 1 with unique associated eigenvector $\boldsymbol{\phi} = \mathrm{D}^{1/2}1$, and all other eigenvalues satisfy $|\lambda_i| < 1$, it follows that $\lim_{k\to\infty} \mathrm{S}^k = \tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}'$, where $\tilde{\boldsymbol{\phi}} = \boldsymbol{\phi}/\|\boldsymbol{\phi}\|$ is the normalised eigenvector.

The result above reveals that the convergence speed depends on the other eigenvalues except one of the propagation matrix S, especially the second largest eigenvalue. Intuitively, the propagation matrix is determined by the topology information of the corresponding graph and this suggests that nodes with higher degree d_i are more likely to suffer from over-smoothing.

In addressing the over-smoothing effect, we measure the spatial smoothness by evaluating the following objective function:

$$\begin{aligned} O(k) =& \hat{\mathrm{Z}}' \widehat{\mathrm{L}} \hat{\mathrm{Z}} \\ =& \mathrm{Z}' \left(\mathrm{I} - \widehat{\mathrm{L}}\right)^k \widehat{\mathrm{L}} \left(\mathrm{I} - \widehat{\mathrm{L}}\right)^k \mathrm{Z}, \quad k = 1, \ldots, K. \end{aligned} \tag{4}$$

It is worth-noting that if $\mathrm{U} = (\mathrm{u}_1, \ldots, \mathrm{u}_n)$ and $\boldsymbol{\Omega} = diag(\omega_1, \ldots, \omega_n)$ are the matrices of eigenvectors and eigenvalues of $\widehat{\mathrm{L}}$, then it follows that $\boldsymbol{\Lambda} = \mathrm{I} - \boldsymbol{\Omega}$ and $\mathrm{v}_i = \mathrm{u}_{n+1-i}$. Hence, after appropriate ordering of the eigenvalues and eigenvectors, it follows that applying the following linear *graph Fourier* transform, $\mathrm{Y} = \mathrm{U}'\mathrm{Z}$, the objective function can be rewritten as:

$$O(k) = \sum_{i}^{n} y_i^2 \left(1 - \omega_i\right)^{2k} \omega_i, \quad k = 1, \ldots, K.$$

Using $n = 6019$ hexagons to tessellate the United States and an adjacency matrix where a_{ij}=1 if hexagons *i* and *j* share a common border and $a_{ij} = 0$ otherwise, Figure 4 (*left*) illustrates the behaviour of the objective function *O*(*k*) as *k* increases. Equation (4) is a strictly decreasing and convex function. Additionally, as *k* increases from 0, Z becomes progressively smoother spatially, eventually approaching a stationary solution for $k > 35$. This observation is supported by Figure 4 (*right*), which shows how S converges towards S^∞.

To optimize the choice of the smoothing hyperparameter k, we consider three complementary approaches based on the generalized cross-validation (GCV), the modified corrected Akaike Information Criterion (AIC_c), and the identification of the elbow point in the objective function $O(k)$. The GCV method selects k by minimizing a score that balances model complexity and goodness of fit, defined as (Ruppert et al. 2003, pag.117):

$$\mathrm{GCV}(k) = \frac{\mathrm{RSS}(k)}{\left(1 - \frac{\mathrm{df}(k)}{n}\right)^2},$$

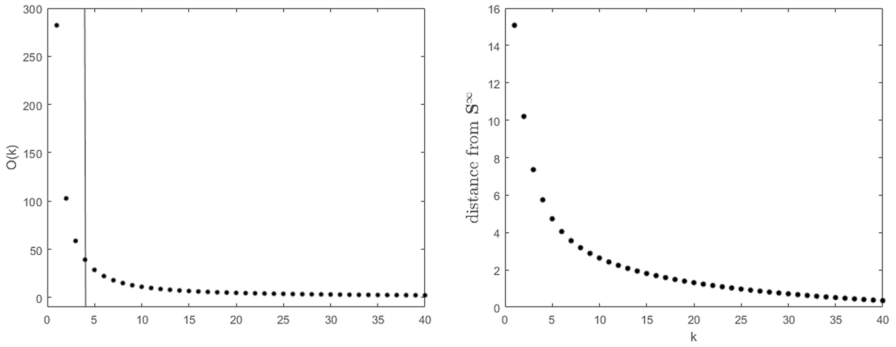

Fig. 4 *Left*: Behaviour of the objective function $O(k)$, illustrating the progression of smoothing as k increases. The vertical line corresponds to the elbow point. *Right*: The dots represents the distance between S^k and S^∞, demonstrating the convergence of S toward a stationary solution.

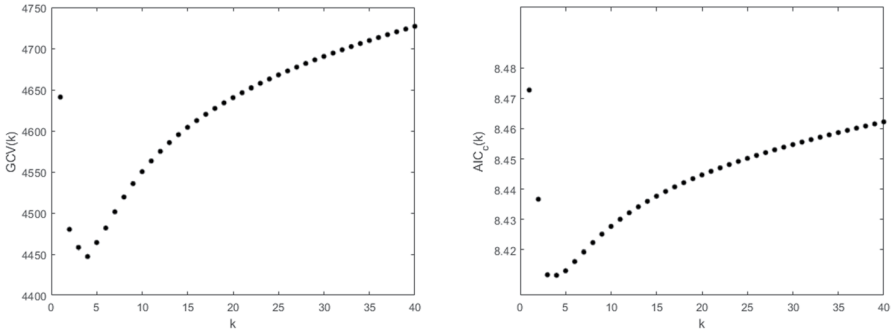

Fig. 5 Comparison of the Generalized Cross-Validation (*left*) and corrected AIC (*right*) curves, highlighting the optimal smoothing parameter $k = 4$ corresponding to their respective minimum values.

where $\mathrm{RSS}(k) = \left(\mathrm{Z} - \hat{\mathrm{Z}}^{(k)}\right)'\left(\mathrm{Z} - \hat{\mathrm{Z}}^{(k)}\right)$ is the residual sum of squares, $\mathrm{df}(k) = trace\left(\left(\mathrm{I} - \widehat{\mathrm{L}}\right)^k\right)$ represents the effective degrees of freedom, and n is the sample size. The modified corrected Akaike Information Criterion (AIC_c) accounts for the structure of the smoothing method, prioritizing both parsimony and predictive accuracy. It is given by (Ruppert et al. 2003, pag.120):

$$\mathrm{AIC}_c(k) = \log\big(\mathrm{RSS}(k)\big) + \frac{2\big(\mathrm{df}(k)+1\big)}{n - \mathrm{df}(k) - 2}.$$

The elbow point of the objective function $O(k)$ is computed to detect the optimal trade-off between complexity and fit. The elbow represents the value of k, where the rate of decrease in $O(k)$ slows significantly, forming an inflection point in the curve.

Interestingly, as shown in Figures 4 and 5, all three criteria consistently suggest $k = 4$ as the optimal choice and Figure 6 illustrates the effect of smoothing, clearly highlighting the areas most severely impacted by Hurricane Ida. The figure reveals a distinct spatial pattern of the storm's impact, with the hardest-hit zones concen-

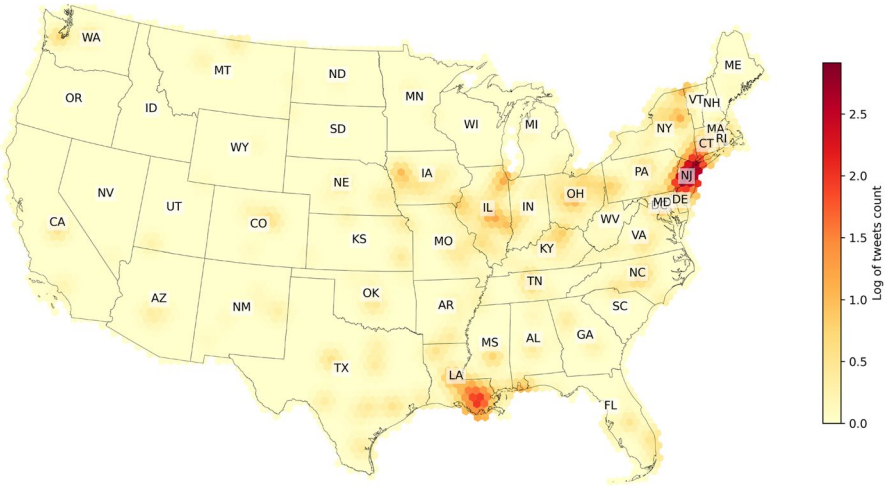

Fig. 6 The map, displayed on a logarithmic scale, illustrates the effect of smoothing on tweet counts, clearly highlighting the areas most severely impacted by Hurricane Ida.

trated along the Gulf Coast, particularly in Louisiana and Mississippi, where it made landfall, causing extensive wind damage, flooding, and infrastructure disruption. The smoothed map also emphasizes transition zones surrounding the core impacted areas, capturing the broader spatial distribution of the hurricane's effects, including regions with moderate to low impact further inland. Moreover, the analysis reflects Ida's significant impact in the Northeastern United States, where record-breaking rainfall led to widespread flash flooding and severe infrastructure disruptions, notably in New York City, New Jersey, and Pennsylvania.

The analysis can be further enhanced by illustrating the areas impacted by Hurricane Ida, categorized according to the severity of the damages. This approach provides a clearer understanding of the spatial distribution of different levels of impact, enabling more targeted insights into the extent of the destruction and facilitating the identification of regions that require priority attention for recovery and response efforts. Figure 7 displays the results of applying our smoothing methodology, separating the tweets by their assigned categories. The figure highlights how certain categories are more prevalent in specific areas, particularly around the hurricane's landfall site, and in the New York area, where the high population density led to a peak in reported damage and activity.

5.3 Tracking Hurricane Ida in space and time

By analysing the spatial distribution and density of tweets over time, it is also possible to reconstruct the storm's path and identify the regions most severely impacted. This approach uses real-time information shared by individuals on social media, offering a unique perspective that complements traditional meteorological and disaster response data.

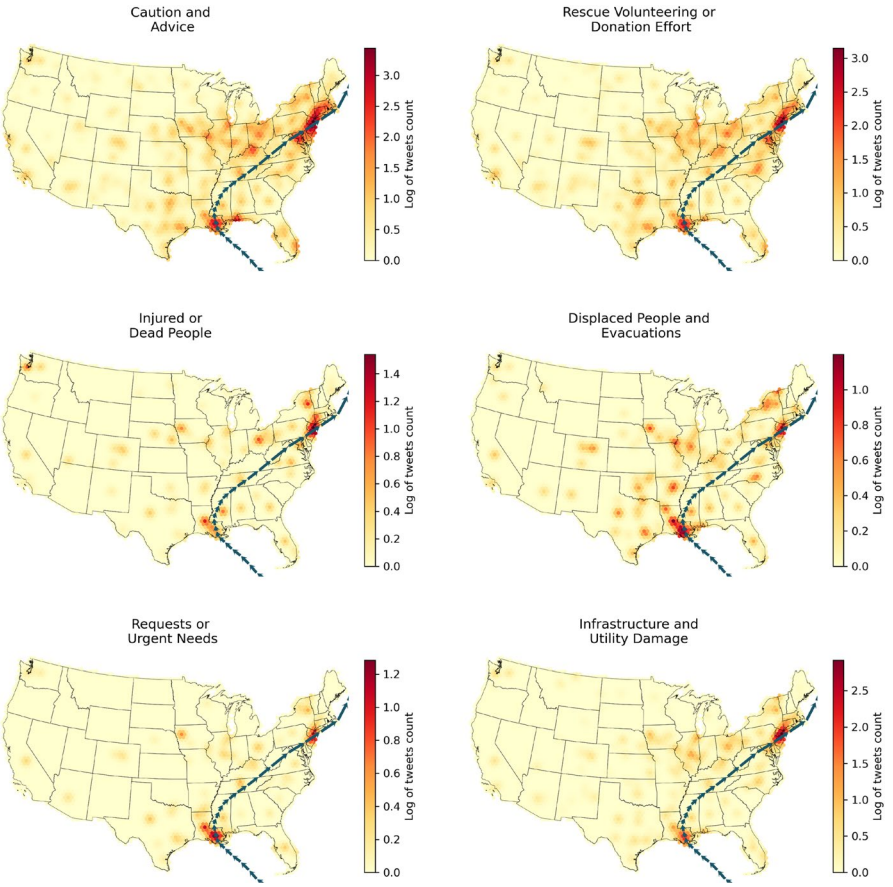

Fig. 7 Smoothed maps of tweet counts by category, displayed on a logarithmic scale. The sequence of arrows indicates the path of Hurricane Ida.

The progression of Hurricane Ida's impact can be more effectively visualized by extending the previous spatial smoothing procedure to a space-time context. This extension is achieved by expanding the neighbourhood set of each node in the network to include connections between consecutive time steps, specifically between time $t-1$ and time t. Consequently, for $t = 1, \ldots, T$, the $(nT \times nT)$ space-time adjacency matrix, $\tilde{\mathrm{A}}_{st}$, is defined as:

$$\tilde{\mathrm{A}}_{st} = \mathrm{I}_T \otimes \tilde{\mathrm{A}}_s + \mathrm{H} \otimes \mathrm{I}_n,$$

where I_r is the $(r \times r)$ identity matrix, $\mathrm{H} = \left(\boldsymbol{e}_2 \cdots \boldsymbol{e}_{T-1} 0\right)$ is a $(T \times T)$ matrix with $\boldsymbol{e}_i$ is the $(T \times 1)$ i-th unit vector, and $\tilde{\mathrm{A}}_s$ is the spatial adjacency matrix with self-loop. This formulation incorporates both spatial and temporal (one lag) dependencies, enabling a dynamic representation of the hurricane's progression. Using geolocated tweets, Figure 8 illustrates the daily progression of Hurricane Ida, with its movement over land represented by arrows. The storm's trajectory begins with

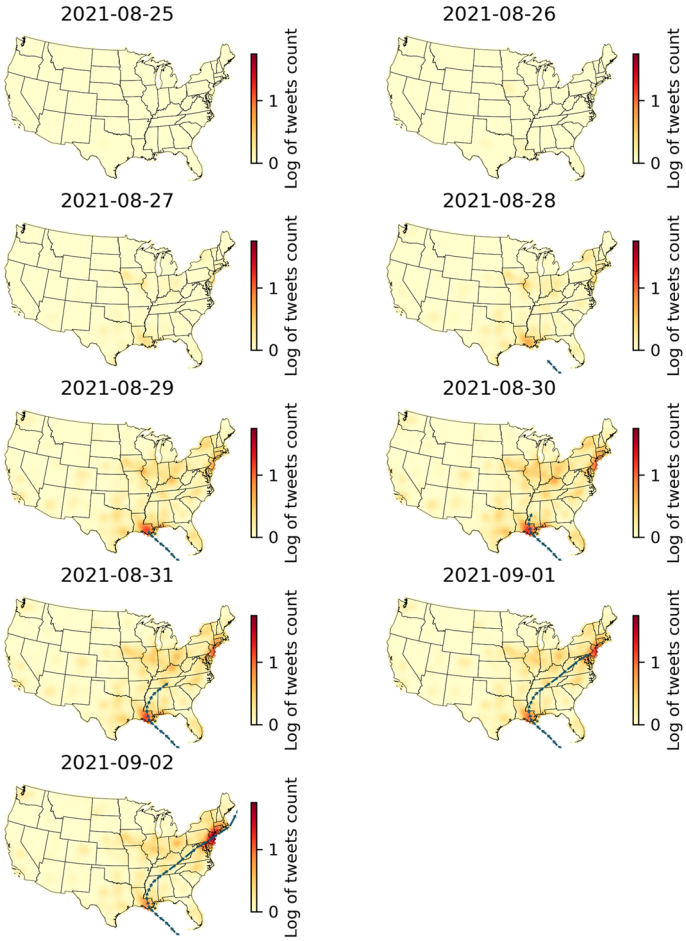

Fig. 8 Temporal evolution of tweet activity using space-time smoothing highlighting the flow of tweets over time in relation to Hurricane Ida's path.

its landfall along the Gulf Coast, followed by its inland path across the southeastern and central United States, and concludes in the Northeastern states. The space-time smoothing approach reveals geospatial patterns of tweet activity that align closely with the storm's observed path. Higher tweet densities are concentrated in areas that experienced severe wind damage, flooding, and infrastructure disruption, particularly in Louisiana and Mississippi during the early stages of the storm and in the Northeastern states as Ida's remnants caused catastrophic flooding. These high-density zones correspond well with reported damage assessments, further validating the effectiveness of this method in disaster analysis. By incorporating temporal connections into the adjacency matrix, the space-time smoothing procedure enhances the resolution of the analysis, offering a more comprehensive understanding of the spatial and temporal dynamics of Ida's impact.

6 Conclusions

Social media data serves as a unique and powerful resource for understanding societal responses to natural disasters. By addressing the informational needs of disaster management entities and crisis responders, such data captures the social dimensions and impacts of catastrophic events. Its timely and valuable insights complement traditional disaster monitoring methods, offering a richer perspective on the interplay between human behavior and disaster dynamics.

This study achieves its core objectives by establishing an effective framework for monitoring disaster events through social networks and demonstrating its applicability in tracking Hurricane Ida. The proposed framework, underpinned by a model trained on historical data, performs effectively when applied to the Hurricane Ida dataset, showcasing its potential for broader applications in disaster contexts. The toponym identification method showed considerable efficacy in markedly augmenting the quantity of geolocated tweets. Furthermore, the proposed smoothing methodology for irregular geometries enhances the creation of more significant maps for disaster tracking and possesses the potential for application to any phenomenon. It is worth noting that the proposed linear smoothing operator also forms the basis for efficient variants of Graph Convolutional neural Networks (GCN) - see, for example, Kipf and Welling (2017). In this context, iterative applications of the Laplacian filter (k iterations) correspond to propagating information across k-hop neighborhoods of the graph. While additional iterations enhance the degree of smoothing in our framework, in nonlinear architectures such as GCNs they are directly related to the network depth (i.e., the number of convolutional layers). Consequently, the proposed approach establishes a direct methodological link between classical spatial smoothing and modern graph-based deep learning, combining interpretability and scalability. From a broader perspective, this connection aligns naturally with the increasing use of neural architectures and transformer-based models for text classification and spatial smoothing tasks.

The insights achieved through the implemented methodology hold significant implications for disaster management. Spatial patterns revealed by the model can help disaster managers pinpoint regions that require targeted information dissemination, ensuring critical messages reach vulnerable communities. This dual contribution - enhancing scientific understanding and improving disaster response strategies -underscores the framework's potential as a valuable tool for managing the social dimensions of natural disasters.

Notwithstanding its strengths, the study also shows some limitations which deserve consideration. Extended BERT classification results on the IDRISI dataset showed label discrepancies in Hurricane Ida tweets. Language is very diversified and changes with time, location, and other contextual factors, making off-topic content classification difficult. Real-time analysis using only training data are complicated by this variability. This could be addressed by adding more varied and representative instances to the training dataset, especially from under-represented groups. However, misclassified tweets in this study tend to fall into two less significant groups, reducing the influence of this limitation on the research.

Additionally, most tweets in the Hurricane Ida dataset are truncated to 140 characters, a limitation that reduces the availability of contextual information for accurate classification. This restriction likely affects both the quality of categorization and subsequent analyses. Future studies could benefit from datasets that include the full text of tweets, such as those collected after Twitter's character limit increase to 280 characters.

To further enhance geolocation accuracy, machine learning-based NLP techniques, such as named entity recognition models tailored for disaster contexts, could be employed to better identify and disambiguate location entities in tweets. Combining such approaches with advanced geocoding tools would enable the development of more accurate and comprehensive geospatial datasets, ultimately supporting deeper analyses of disaster dynamics and their spatial impacts.

By addressing these limitations, future research can unlock the full potential of social media data for real-time disaster monitoring and management.

Appendix A. Technical description of BERT and DistilBERT architectures

This appendix provides a detailed technical description of the Transformer-based models discussed in Section 3. Specifically, we describe the BERT architecture, its self-attention mechanism, and the structure of the DistilBERT model used in our analysis.

Engaging with textual data presents considerable challenges primarily due to its unstructured nature. To address these challenges and extract valuable information, the field of text mining has developed multiple strategies.

Large pre-trained models, such as BERT (Devlin et al. 2018), represent a breakthrough in the application of deep learning to NLP. These models have significantly advanced numerous NLP tasks, including semantic analysis. BERT utilizes the Transformer architecture, which employs an attention mechanism to learn contextual relations between words in a text.

Unlike directional models, which read text input sequentially (either left-to-right or right-to-left), the Transformer encoder processes the entire sequence of words simultaneously. Therefore, it is considered bidirectional, though it is more accurate to describe it as non-directional. This characteristic enables the model to understand the context of a word based on all its surrounding words (both left and right), leading to a more comprehensive understanding of the text (Devlin et al. 2018).

From a technical perspective, BERT's architecture consists of an encoder and a decoder component. The encoding component is a stack of encoders, each with sublayers represented by a Feed-Forward Neural Network and a Self-Attention layer. Between each layer in both the encoder and decoder modules, residual connections are followed by layer normalization. These mechanisms facilitate the flow of gradients through the network and help stabilize training.

The encoder module processes the input sequence, extracting features and creating a rich representation of the input. Each input word, or token, is represented by three types of embeddings: a token embedding, a segment embedding, and a position

embedding. The first encoder block takes combined embeddings from BERT's first layer as input, while subsequent blocks use the output from the previous block. During training, a small fraction of tokens in each sentence is randomly masked, and the objective of the self-attention mechanism (Vaswani et al. 2017) is to predict these masked words.

Self-attention enables the model to weigh the importance of different words in a sequence, allowing it to focus on relevant information and capture long-range dependencies within the text. The mechanism involves calculating an attention score based on surrounding tokens to better represent a token embedding. If we consider token (input) embedding to be the query vector (Q), and existing information regarding the sequence to be a pair of multiple key-value ($K - V$) vectors (derived from token embeddings of surrounding tokens), the query vector's representation based on the attention mechanism is given by

$$Attention(Q, K, V) = softmax\Big(\frac{QK^T}{\sqrt{d_k}}\Big),$$

where QK^T is the dot product between the query (Q) and key vectors (K), and d_k is the dimension of the key vector. The dot product is scaled by $\sqrt{d_k}$ to avoid the value from growing very large which would push the softmax value to regions that may result in very small gradients. The query, key, and value vectors in self-attention come from the input embeddings. We use the information regarding different tokens in a sequence to generate a contextualized representation for each token in the same sequence. In other words, we represent a token in the context of other tokens in the same sequence. In Multi-Head Self-Attention, the attention mechanism is applied multiple times, each with its own set of learnable parameters. This results in multiple sets of attention weights, allowing the model to attend to different aspects of the input simultaneously.

However, self-attention is not enough to perform a complicated text embedding task. Therefore, the encoder block completes with the definition of a position-wise fully connected feed-forward network and a layer for normalization. The second layer in an encoder block is thus represented by the position-wise fully connected feed-forward network. Position-wise implies that each token (output of multi-head self-attention sub-layer) is fed individually to feed-forward networks. This network consists of two linear layers separated by a non-linear activation function $f(\cdot)$, typically ReLU (Rectified Linear Unit) such that, if we denote the input to this sublayer with x, the output is given by

$$FFNet = max(w_1 \cdot x + b_1), 0)w_2 + b_2,$$

where w_1, b_1 and w_2, b_2 are parameters for the first and second feed-forward networks, respectively. By employing linear transformations and non-linear activation functions, feed-forward networks empower the model to navigate the complex semantic landscape of language, facilitating robust comprehension and generation of text. The final step in an encoder block is to combine the output of the position-wise feedforward network and the input embeddings, and then pass them through a Nor-

malization layer. The output obtained from layer normalization is then the output for the particular encoder block.

The output of the encoder is finally fed into a decoder block, which predicts the translated sentence. The decoder also consists of multiple blocks of multi-head attention, which are fed into feed-forward neural networks and add positional embeddings to the inputs.

Since its debut, a number of substitute versions have been introduced for BERT. Sentimental analysis, phrase prediction, abstract summarization, question-answering, natural language inference, and a host of other NLP tasks have all been revolutionised by BERT Technology. In this paper, we used the DistilBERT variant which is a small, fast, cheap and light Transformer model trained by distilling BERT base. It has 40% less parameters than the BERT-base model, runs 60% faster while preserving over 97% of BERT's performances as measured on the GLUE language understanding benchmark (Sanh et al. 2019). The model has 6 layers, 768 hidden dimensions and 12 Self-Attention Heads, totalizing 66 Millions parameters - compared to 340M parameters for BERT-Large and 110M for BERT-base.

There are two steps in training BERT: pre-training and fine-tuning. During the pre-training stage, using the same corpus as the original BERT model, DistilBERT learns the distribution of the words in the general language from the large corpus of unlabeled text including BooksCorpus (800M words) and English Wikipedia (2,500M words) using masked language modeling and next-sentence prediction methods. Hence, transfer learning strategies significantly boosts BERT's effectiveness in downstream NLP tasks by utilizing pre-existing knowledge encoded in pre-trained models, thereby reducing the dependency on large training datasets (Pan and Yang 2010).

Crucially, the knowledge acquired during pre-training is transferable to a wide range of NLP tasks. When fine-tuning BERT for specific tasks, such as text classification, named entity recognition, or question answering, transfer learning enables the model to adapt its learned representations to the nuances and requirements of the target task. Accordingly, at the fine-tuning stage, DistilBERT is initialized with the pre-trained weights from the previous stage, and all the parameters are fine-tuned by labelled data from downstream tasks.

A schematic representation of DistilBERT's architecture is provided in Figure 9.

Appendix B. Evaluation metrics

This appendix provides the formal definitions of the evaluation metrics used to assess classification performance in Section 4.

Model accuracy measures the proportion of correctly classified samples out of the total number of samples:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

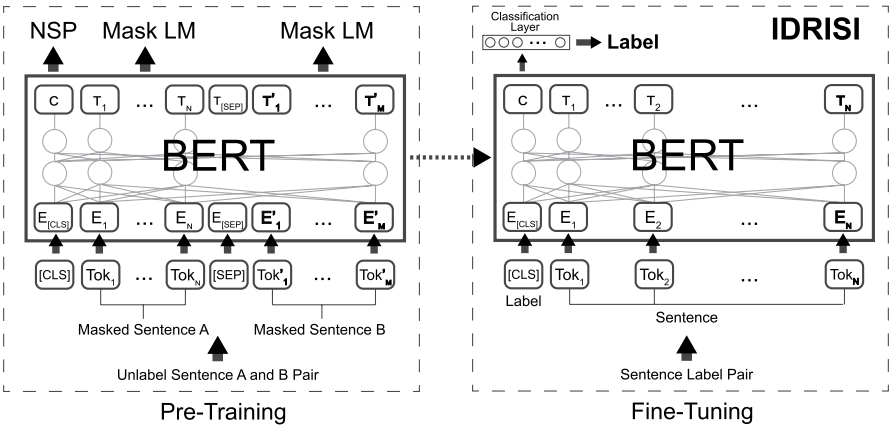

Fig. 9 Pre-training and fine-tuning procedure of DistilBERT for IDRISI classification task. During the fine-tuning a layer for classification is added and the full network weights are trained for the classification task, with starting weights coming from the pre-trained model. E_i denotes the token embeddings, C represents the final hidden state of the special *[CLS]* (Devlin et al. 2019) token used as the aggregate sequence representation for classification, and T_i indicates the final hidden state corresponding to the i-th input token.

where *TP*, *TN*, *FP*, and *FN* denote true positives, true negatives, false positives, and false negatives, respectively.

Precision (P) quantifies the proportion of correctly predicted positive samples among all samples assigned to the positive class:

$$P = \frac{TP}{TP + FP}.$$

Recall (R) measures the proportion of actual positive samples correctly identified by the model:

$$R = \frac{TP}{TP + FN}.$$

The F1 score is the harmonic mean of precision and recall, balancing both false positives and false negatives:

$$F1 = \frac{2 \times P \times R}{P + R}.$$

Together, these metrics provide a comprehensive evaluation framework: accuracy reflects overall correctness, while precision and recall characterize the model's sensitivity to positive classes and its robustness to misclassification. The F1 score synthesizes these aspects into a single measure of predictive quality.

Appendix C. N-folds cross-validation

Cross-validation is a widely employed resampling technique for assessing the predictive performance and generalization ability of machine learning models (Hastie et al. 2009). Instead of relying on a single train–test partition, the available dataset is divided into N equally sized subsets, or folds. During each iteration, one fold is reserved for validation, while the remaining N - 1 folds are used for model training. The process is repeated N times so that each subset serves once as the validation fold.

This procedure yields a more robust and less biased estimate of model performance by reducing the dependence on any particular data split. It is particularly beneficial when the available training data are limited, since all observations are utilized for both training and validation across the iterations.

After completion of the N cycles, the evaluation metrics are averaged over all folds, and the corresponding standard deviations are computed to quantify variability in model performance. In the present study, a 10-fold cross-validation scheme (N = 10) was adopted, providing a well-established balance between computational efficiency and statistical reliability.

Appendix D. Geoparsing and extraction of spatial information from tweets

This appendix provides a detailed description of the methodology used to infer geolocation information from tweet text, as summarized in Section 5.2.

Geoparsing involves two key stages: (1) *Toponym Recognition*, the identification of place names within textual data, and (2) *Toponym Resolution*, the translation of these names into geographic coordinates. Figure 10 illustrates this workflow.

For the recognition step, we adopted the *NeuroTPR* model (Wang et al. 2020), specifically developed for toponym recognition in social media content. NeuroTPR builds upon the Bidirectional Long Short-Term Memory Conditional Random Field (BiLSTM-CRF) architecture introduced by Lample et al. (2016), integrating multiple enhancements that improve accuracy under informal linguistic conditions, such as inconsistent capitalization and spelling variations.

The architecture incorporates several layers:

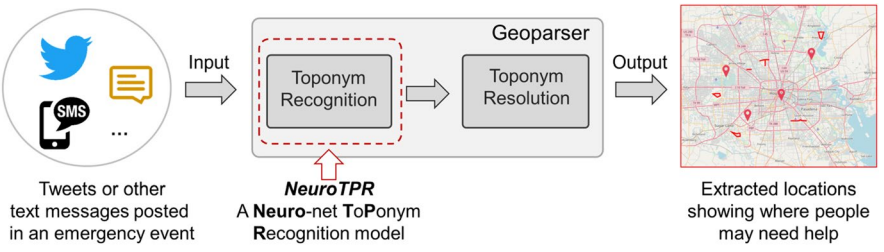

Fig. 10 Extraction of location from the textual content of social media messages, based on Wang et al. (2020).

Fig. 11 Example of selection of a toponym from a message.

- *Character embeddings.* Dual layers capture both case-sensitive and caseless morphological features, improving recognition of irregular or misspelled words.
- *Word embeddings.* Pre-trained GloVe embeddings capture semantic associations based on word co-occurrence statistics.
- *Contextualized embeddings.* ELMo embeddings (Peters et al. 2018) encode contextual semantics by modeling word usage in surrounding text.
- *Sequential labeling.* A BiLSTM layer models dependencies between tokens, while a CRF layer enforces consistency in predicted label sequences.Figure 11 illustrates the toponym recognition process. In this study, the general pretrained NeuroTPR model was employed without further task-specific training.

To convert recognized toponyms into geographic coordinates, we constructed a hierarchical dictionary using the U.S. *National Address Database (NAD)*[2], which links locality names, postcodes, counties, and states with their corresponding coordinates. Counties and states were assigned the average coordinates of their constituent postcodes. Each recognized toponym was then matched to this dictionary to obtain latitude–longitude pairs.

This approach is limited by common challenges in geographic name resolution, such as ambiguous toponyms (e.g., "Washington" as both a city and a state) and minor spelling variations. While our implementation relied on exact string matching, more advanced systems could employ fuzzy matching or contextual disambiguation to improve accuracy.

The spatial tessellation was implemented using the *H3*[3] geospatial indexing system developed by Uber. This system partitions the Earth's surface into hexagonal cells defined across several predefined resolution levels, each corresponding to a fixed and globally consistent number and size of hexagons.

For this study, resolution level 4 was selected, as it generates approximately 6,019 hexagonal cells-closely matching the number of U.S. counties-and thus provides an appropriate geographic granularity for the analysis. Using finer resolutions would result in excessively small cells that, given the toponym-based geocoding approach and the inherent limitations in positional precision, could introduce noise or misleading spatial patterns rather than additional reliable detail. This resolution level therefore represents a balanced choice, consistent with both the large spatial footprint of hurricanes and the accuracy of the available geotagging.

The use of the *H3* indexing system was further motivated by its scalability and interoperability. These characteristics facilitate the integration of the proposed analytical framework with other spatial datasets and geospatial infrastructures that rely on the same hierarchical indexing logic. Such interoperability also supports potential extensions of the approach to alternative spatial resolutions or additional data sources.

[2] https://www.transportation.gov/gis/national-address-database

[3] https://h3geo.org/docs

Acknowledgements The research of L. Ippoliti is funded by the European Union - NextGenerationEU, research project PRIN2022 PNRR SLIDE - Stochastic Modeling of Compound Events, CUP D53D23018920001.

References

Alam F, Ofli F, Imran M, et al (2020) Deep Learning Benchmarks and Datasets for Social Media Image Classification for Disaster Response. CoRR abs/2011.08916., arXiv:2011.08916

Alam F, Sajjad H, Imran M et al (2021) CrisisBench: Benchmarking Crisis-related Social Media Datasets for Humanitarian Information Processing. Proceedings of the International AAAI Conference on Web and Social Media 15(1):923–932. https://doi.org/10.1609/icwsm.v15i1.18115

Bossu R, Laurin M, Mazet-Roux G et al (2015) The Importance of Smartphones as Public Earthquake-Information Tools and Tools for the Rapid Engagement with Eyewitnesses: A Case Study of the 2015 Nepal Earthquake Sequence. Seismol Res Lett 86:1587–1592. https://doi.org/10.1785/0220150147

Burel G, Alani H (2018) Crisis Event Extraction Service (CREES) - Automatic Detection and Classification of Crisis-related Content on Social Media. In: Boersma K, Tomaszeski B (eds) ISCRAM 2018 Conference Proceedings - 15th International Conference on Information Systems for Crisis Response and Management. Rochester Institute of Technolog, Rochester, NY (USA), pp 597–608

Castillo C (2016) Big Crisis Data: Social Media in Disasters and Time-Critical Situations. Cambridge University Press. https://doi.org/10.1017/CBO9781316476840

Devlin J, Chang M, Lee K, et al (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR abs/1810.04805. arXiv:1810.04805

Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer, New York, NY, https://doi.org/10.1007/978-0-387-84858-7

Imran M, Elbassuoni S, Castillo C et al (2013) Extracting Information Nuggets from Disaster- Related Messages in Social Media. In: Comes T, Fiedrich F, Fortier S et al (eds) ISCRAM 2013 Conference Proceedings - 10th International Conference on Information Systems for Crisis Response and Management. Karlsruher Institut fur Technologie, KIT; Baden-Baden, pp 791–801

Kipf TN, Welling M (2017) Semi-Supervised Classification with Graph Convolutional Networks. In: International Conference on Learning Representations (ICLR), https://arxiv.org/abs/1609.02907

Kumar A, Singh JP, Saumya S (2019) A Comparative Analysis of Machine Learning Techniques for Disaster-Related Tweet Classification. In: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129), pp 222–227, https://doi.org/10.1109/R10-HTC47129.2019.9042443

Lample G, Ballesteros M, Subramanian S, et al (2016) Neural Architectures for Named Entity Recognition. In: Knight K, Nenkova A, Rambow O (eds) Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, San Diego, California, pp 260–270, https://doi.org/10.18653/v1/N16-1030

Landwehr PM, Carley KM (2014) Social Media in Disaster Relief, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 225–257. https://doi.org/10.1007/978-3-642-40837-3_7

Pan SJ, Yang Q (2010) A Survey on Transfer Learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191

Peters ME, Neumann M, Iyyer M, et al (2018) Deep contextualized word representations. CoRR abs/1802.05365. arXiv:1802.05365

Phillips ME (2021) Hurricane Ida Twitter Dataset. Https://digital.library.unt.edu/ark:/67531/metadc1913080/, accessed May 16, 2023), University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge

Sanh V, Debut L, Chaumond J, et al (2019) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR abs/1910.01108. arXiv:1910.01108

Starbird K, Palen L, Liu SB, et al (2012) 3 - Promoting structured data in citizen communications during disaster response: an account of strategies for diffusion of the 'Tweak the Tweet' syntax. In: Hagar C (ed) Crisis Information Management. Chandos Information Professional Series, Chandos Publishing, p 43–63, https://doi.org/10.1016/B978-1-84334-647-0.50003-5

Subba R (2016) Online Convergence Behavior, Social Media Communications and Crisis Response: An Empirical Study of the 2015 Nepal Earthquake Police Twitter Project. In: Hawaii International Conference on System Sciences, https://doi.org/10.24251/HICSS.2017.034

Suwaileh R, Elsayed T, Imran M (2023) IDRISI-RE: a generalizable dataset with benchmarks for location mention recognition on disaster tweets. Inform Processing Manag 60(3):103340. https://doi.org/10.1016/j.ipm.2023.103340

U.S. Environmental Protection Agency (2021) Hurricane ida response site. https://response.epa.gov/HurricaneIda, accessed: 2025-11-06

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention Is All You Need. CoRR abs/1706.03762. arXiv:1706.03762

Wang J, Hu Y, Joseph K (2020) NeuroTPR: a neuro-net toponym recognition model for extracting locations from social media messages. Trans GIS 24(3):719–735. https://doi.org/10.1111/tgis.12627

Wang Z, Ye X (2018) Social media analytics for natural disaster management. Int J Geogr Inf Sci 32(1):49–72. https://doi.org/10.1080/13658816.2017.1367003

Young C, Kuligowski E, Pradhan A (2020). A Review of Social Media Use During Disaster Response and Recovery Phases. https://doi.org/10.6028/NIST.TN.2086

Authors and Affiliations

Emiliano del Gobbo[1] · Luigi Ippoliti[2] · Lara Fontanella[3] · Barbara Cafarelli[4]

✉ Emiliano del Gobbo
emiliano.delgobbo@unina.it

Luigi Ippoliti
luigi.ippoliti@unich.it

Lara Fontanella
lara.fontanella@unich.it

Barbara Cafarelli
barbara.cafarelli@unifg.it

[1] Department of Electrical Engineering and Information Technology, University of Naples Federico II, Via Claudio 21, 80125 Naples, NA, Italy

[2] Department of Economics, University of Chieti-Pescara, Via dei Vestini 42, 65127 Pescara, PE, Italy

[3] Department of Socio-Economic, Managerial, and Statistical Studies, University of Chieti-Pescara, Via dei Vestini 42, 65127 Pescara, PE, Italy

[4] Department of Economics, Management and Territory, University of Foggia, Via da Zara 11, 71121 Foggia, FG, Italy