DisImpact: Quantifying the Physi-Social Impact of Natural Disasters Through Social Media

Ruichen Yao[1], Tejna Dasari[1], Xuanyu Meng[1], Elliot Cao[1], Zelin Li[1], Yifan Liu[1], Yaokun Liu[1], Lanyu Shang[2], Dong Wang[1]

[1]University of Illinois Urbana-Champaign, Illinois, USA
[2]Loyola Marymount University, California, USA
[1]{ryao8, tejnad2, xuanyum2, elliotc4, zelin3, yifan40, yaokunl2, dwang24}@illinois.edu; [2]lanyu.shang@lmu.edu

Abstract

Natural disasters not only cause large-scale physical destruction, but also cascading social consequences that are difficult to quantify with traditional surveys and reports. Social media platforms offer an alternative perspective that captures multimodal, real-time, and user-generated content that can be leveraged for disaster impacts. In this paper, we introduce DisImpact, a two-stage framework that systematically quantifies the physi-social impacts of disasters via a Multimodal Large Language Model (MLLM). The social media posts are first classified into ten disaster impact categories that cover both physical and social domains. We then construct a disaster impact index that integrates the relative prominence of each category with the intensity of public engagement on a weekly basis. This design provides a unified scale for representing disaster impacts across both individual disaster impact categories and the broader physical and social domains. The unified representation enables direct comparison across categories and allows the impacts to be flexibly aggregated to reveal higher-level patterns and overall trends. We validate the impact indices against authoritative ground-truth data, including FEMA Public Assistance data and NASA FIRMS fire detections, observing consistent lead-lag correlations that demonstrate strong validity across both social and physical impact dimensions. We further conduct temporal and spatial analyses, and the results show that physical impacts are often peak during the disasters and localized in regions that are directly affected by disasters, while social impacts often emerge later and spread more broadly across time and space. To the best of our knowledge, this is the first framework to comprehensively quantify disaster impacts across their physical and social dimensions using multimodal data from multiple social media platforms.

Introduction

The increasing frequency and severity of natural disasters (e.g., hurricanes, wildfires, earthquakes) cause not only large-scale physical destruction but also profound social disruption, posing urgent challenges to human society (Zhang et al. 2023; Yao et al. 2025; Marshall and Wang 2016). For instance, in September and October 2024, a series of hurricanes struck the United States, together causing more than 250 casualties and more than $300 billion in economic losses (Wilcox and Jacobs 2024). In addition, a dozen of wildfires affected California in January 2025, causing 440 estimated deaths and destroying more than 15,000 houses (Paglino, Raquib, and Stokes 2025; Lindsey 2025). Beyond physical destruction, disasters also trigger cascading social consequences, ranging from public discourse on resource allocation to heightened perceptions of risk and inequity (Basu et al. 2024; Li et al. 2025).

Traditional documentation of physi-social impacts relies on surveys and government reports. However, these sources often involve significant delays in data collection, capture only a subset of impact dimensions, and are therefore insufficient for capturing the dynamic, real-time evolution of disaster impacts. In contrast, social media posts provide timely, multimodal, and user-driven observations. The posts can reflect both physical consequences, such as damage to infrastructure, resource shortages, and casualties; and social responses, such as distress, blame, and solidarity. Leveraging the real-time social media data enables researchers to move beyond isolated case studies by providing continuous and large-scale observations of how disasters affect communities over time. In this paper, we introduce *DisImpact*, a framework that quantifies physi-social impacts of disasters in ten different impact categories, covering both physical (e.g., casualties, damage, resource shortage) and social (e.g., distress, assistance) dimensions, by analyzing multimodal posts from multiple social media platforms. By applying the defined impact categories to posts, our framework captures how different types of impact evolve and interact, offering structured insights that can support disaster response, risk communication, and policy development.

Existing studies have developed multiple ways to quantify the impacts of disasters, but most focus on a single dimension. For example, Miller et al. (2025) and Mitchell, Maheen, and Bowen (2024) measured how exposure to disaster would increase public perceptions of crisis, depression, and suicidal thoughts. Yang et al. (2025) focus on infrastructure disruptions, showing how damage to physical systems undermines community well-being. Jamal and Hasan (2023) and Basu et al. (2024) evaluate preparedness and adaptive capacity through community surveys. Although these approaches provide valuable insights, they typically examine only one dimension of impact at a time, such as mental-health outcomes measured through clinical symptom scales (Miller et al. 2025), infrastructure disruptions

quantified through service-interruption metrics (Yang et al. 2025), or community-resilience indicators derived from social and behavioral data (Jamal and Hasan 2023). Because these measurements rely on fundamentally different scales, they cannot be directly compared, leading to fragmented assessments that obscure how physical and social impacts relate to each other, as infrastructure damage may amplify psychological distress and existing inequities may determine recovery outcomes. In contrast, an integral framework that quantifies both physical and social indicators in a unified scale enables a comprehensive understanding of the impacts of disasters and can provide stronger empirical support for response planning and policy development.

To address these limitations, we propose **DisImpact**, a two-stage framework that systematically quantifies physi-social impacts using multimodal social media data from multiple platforms. Different from previous studies that lack quantification or examine only a single dimension such as mental health or infrastructure damage, our framework introduces ten impact categories spanning both physical and social dimensions. We first leverage a Multimodal Large Language Model (MLLM) to classify a post into defined disaster impact categories based on its multimodal content. To capture the evolving prominence of each category over time, we then calculate an impact index that combines the proportional prevalence of each category with a measure of public engagement intensity within each time window. This formulation allows us to capture both the relative prominence of each impact category and the overall intensity of public engagement within each time window, offering a comprehensive and scalable measure of disaster impacts. Moreover, as all category-level indices are defined in a unified scale, they function as modular components that can be flexibly combined to construct higher-level composite impact measures, such as physical or social impact dimensions.

To demonstrate the effectiveness of **DisImpact**, we evaluated it on two large-scale real-world disaster datasets: the *2024 Atlantic Hurricanes* and the *2025 California wildfires*. These events differ in the hazard type, geography, and social context, providing a rigorous testbed for our framework. Through validation studies, DisImpact shows consistent alignment with authoritative ground-truth data: FEMA Assistance data and NASA FIRMS fire detections, indicating that the impact indices can accurately capture both social and physical disaster impacts. Using DisImpact, we perform both temporal and spatial analysis of the impact indices, revealing that physical impacts peak during the disaster onset, whereas social impacts often rise in the aftermath and diffuse more broadly across time and space. We envision that the framework can inform a deeper understanding of physi-social impacts, support disaster recovery planning, and enhance real-time disaster management strategies.

Related Works

Physical and Social Impacts in Disaster

Existing research highlights the impacts of disasters in both the physical realm and social dimensions, both having substantial effects on communities.

Physical Impacts. The most visible effect of disasters is immediate physical destruction and its cascading effects. *Casualties and injuries* rise dramatically from structural collapse, flooding, and entrapment. Parks et al. (2021) demonstrat that tropical cyclone exposure increases respiratory illness from mold, smoke, particulates, and contaminated water. These mortality hazards remain elevated for many years after severe flooding, according to a Hurricane study by Keenan et al. (2025). Beyond health risks, disasters force families to evacuate, exposing them to secondary hazards. Jamal and Hasan (2023) highlight that the *evacuation* process itself exposes individuals to dangerous traffic conditions and overcrowded shelters. Severe disasters can cause extensive *damage to infrastructure and utility systems*. Yang et al. (2025) reporte collapsed buildings, flooded roads, and widespread power outages, cutting access to essential lifelines, including food, water, and healthcare. Disasters also cause significant *environmental damage* to the ecosystem. Basu et al. (2024) assert that disasters routinely cause unsafe drinking water, dwindling fuel reserves, and depleted medical stockpiles. Breakdowns caused by disasters often cascade into broader *environmental damage*. For example, Juarez et al. (2025) indicate that wildfires leave behind toxic smoke, devastate agricultural land, and accelerate ecosystem collapse.

Social Impacts. Disasters also have a substantial impact on society and local communities. Miller et al. (2025) show that extreme weather events often result in severe *public health consequences*, including disease outbreaks, mold-related health issues, and disruptions to hospital services. Garfin et al. (2022) demonstrate that repeated exposure to disasters heightens like PTSD, depression, and functional impairment. For example, victims exposed to fires and smoke particles experience increased *emotional and psychological distress*, leading to spikes in mental health visits (Juarez et al. 2025; Jung et al. 2025). In addition, Wertis et al. (2023) demonstrate that disasters continue to cause *socioeconomic disruptions* such as forced business closures, educational failures, and increased housing insecurity, leading to community destabilization. In addition, disasters also shape social narratives. Abid et al. (2024) highlight that disasters exacerbate *inequities* by fueling political blame, discrimination, or the disproportionate suffering of marginalized populations. Li et al. (2025) show that local community networks, governments, and NGOs collaborate to mobilize *assistance and recovery*, such as providing resources and rebuilding infrastructure, to foster long-term resilience.

Prior studies examine disaster impacts in narrow classes and rarely quantify them. In this work, we introduce ten physical and social categories and use them to label social media data, producing a holistic and measurable framework.

Existing Quantification Methods

Prior work has quantified disaster impacts using both physical and social measures through distinct, domain-specific methodologies. Physical impacts are commonly assessed through official assessment frameworks such as the Post-Disaster Needs Assessment (PDNA), which aggregates government-led evaluations of damages, losses, and recovery needs using structured surveys, field inspections, and

administrative records collected after major events (Global Facility for Disaster Reduction and Recovery (GFDRR) et al. 2017). Complementing these institutional assessments, Keenan et al. (2025) estimate excess mortality following Hurricane Sandy by linking longitudinal Medicare Fee-for-Service beneficiary records with flood exposure at the ZIP Code Tabulation Area level, enabling population-scale inference of disaster-related health effects. Wang et al. (2012) introduce a framework to rigorously assess the accuracy of the quantification using estimation theoretical methods. Social impacts have been examined using a range of survey-based, behavioral, and communication-centered approaches. Garfin et al. (2022) assess post-disaster mental health outcomes among hurricane-exposed populations using standardized post-traumatic stress symptom (PTSS) scales and measures of functional impairment collected through longitudinal surveys. Mansur, Langhorn, and Nelson (2024) analyze public discourse on X during Hurricane Ida using large-scale textual analysis to examine narratives of government responsibility and infrastructure inequality. Wertis et al. (2023) employ a quasi-experimental design leveraging crisis-text message volumes to measure changes in anxiety, suicidal ideation, and bereavement following Hurricane Ida, enabling causal inference on socioeconomic and mental-health disruption. Despite these advances, existing methods remain largely isolated, with most studies focusing on a single impact dimension (e.g., mortality, mental health, or infrastructure damage) and employing distinct measurement pipelines that limit direct comparison across disaster types. Official assessments are systematic but slow and resource-intensive, while social media-based approaches offer faster signals but often rely on a single platform. Our framework addresses these gaps by organizing disaster impacts into ten physical and social categories and introducing a unified index that quantifies both prevalence and intensity across hazards, modalities, and platforms. Leveraging data from Reddit, TikTok, and YouTube, DisImpact expands the representation of public experiences and provides a scalable complement to traditional assessment methods.

Data Collection

In this study, we focus on two recent disasters, 2024 Pacific Hurricanes and 2025 California Wildfires. For the 2024 Pacific Hurricanes, we focus on the three major hurricanes that hit Florida and North Carolina during the fall of 2024: *Hurricane Francine*, *Hurricane Helene*, and *Hurricane Milton*. According to the National Hurricane Center (NHC), Hurricane Francine was a Category 2 Storm formed on September 9, 2024; Hurricane Helene was a Category 4 Storm formed on September 24, 2024; and Hurricane Miltion was a Category 5 Storm formed on October 5, 2024. The combined damage of these consecutive hurricanes is estimated to be more than $300 billion (Wilcox and Jacobs 2024). For the 2025 California Wildfires, we focus on the wildfires that hit Los Angeles in January 2025. According to NOAA Climate (Lindsey 2025) and the study of Paglino, Raquib, and Stokes (2025), the 2025 California Wildfires caused 440 estimated deaths and destroyed more than 15,000 houses. To ensure timely and relevant coverage of the disaster-related discussions, we collect posts for hurricanes from September to November 2024, and for wildfires from December 2024 to March 2025. This collection window spans approximately one month before disaster onset and one month after the main impact period, enabling us to capture pre-event signals, peak impacts, and recovery discussions.

	Reddit	**TikTok**	**YouTube**
Collected Raw Posts	12,301	67,027	2,767
Relevant Posts (%)	9,666 (79%)	47,921 (71%)	1,707 (62%)
Avg. #Words per Rel. Post	119	21	18
#Images on Rel. Posts	1,579	–	–
#Videos on Rel. Posts	878	47,921	1,707

Table 1: Distribution of Collected Posts on Hurricanes

	Reddit	**TikTok**	**YouTube**
Collected Raw Posts	9,456	40,208	2,294
Relevant Posts (%)	6,204 (66%)	17,567 (44%)	1,339 (58%)
Avg. #Words per Rel. Post	92	35	23
#Images on Rel. Posts	923	–	–
#Videos on Rel. Posts	382	17,567	1,339

Table 2: Distribution of Collected Posts on Wildfires

We collected posts from three platforms: Reddit, TikTok, and YouTube with a total of 134,053 posts. These three platforms are widely adopted across diverse demographics and communities, enabling the dissemination of information through multiple modalities such as text, images, and videos (Sihag et al. 2023; Feng et al. 2023). The social media data collection adopted the Reddit API, TikTok Research Tools, and YouTube API. We did not include data from X because the platform terminated the free academic API access since 2023 and replaced it with expensive paid tiers, making large-scale data collection financially infeasible for this project (Murtfeldt et al. 2024; Pehlivan et al. 2025; Bisiani et al. 2025). We utilized different keywords such as *Hurricane Helene*, *Hurricane Milton*, *Los Angeles Wildfire*, and *Palisades Wildfire* to retrieve relevant content for different disasters. To ensure the ethics of the research and protect users' privacy, we only collected post information (i.e., post ID, post content, post location, and post time) without users' identity information. Tables 1 and 2 present the number of posts collected from various social media platforms. For text-centric platforms like Reddit, we not only collected the title and description of the post but also collected the attached images and videos. For video-based platforms such as TikTok and YouTube, we collected each post's video content along with its title and description, while restricting videos to a maximum length of five minutes to reduce the cost and workload for further cleaning and annotation tasks.

Data Cleaning

We observe that the collected raw social media posts contain irrelevant and off-topic posts. For example, the keyword "hurricane" can appear on posts related to the Miami Hurricanes football team or the Carolina Hurricanes hockey

Domain	Category	Definition
Physical Impact	**Casualties & Injuries (CINJ)**	Posts describing people or animals who are killed, seriously injured, missing, or experiencing immediate medical emergencies.
	Evacuations & Displacement (EVAC)	Posts describing people being forced to evacuate, or relocations to shelters caused by unsafe conditions during a disaster.
	Infrastructure & Utility Damage (INFR)	Posts reporting physical damage to infrastructure or utility systems, such as roads, buildings, bridges, or disruptions to essential services such as electricity, water, or communication networks.
	Environmental Damage (ENVD)	Posts describing harm to the natural environment or resources caused by the disaster, such as damage to ecosystems, agriculture, or coastlines, as well as contamination of water, soil, or air.
	Resource Shortages (RSRC)	Posts requesting essential survival resources during a disaster, such as clean water, food, shelter, clothing, or other urgent supplies.
Social Impact	**Public Health (PUBH)**	Posts describing public health consequences following a disaster, such as outbreaks of infectious diseases, disruption of chronic illness care, and shortages of medical supplies.
	Emotional & Psychological Distress (EMOT)	Posts describing psychological distress or emotional suffering experienced by oneself or others as a result of a disaster, such as trauma, anxiety, depression, grief, or cumulative emotional strain.
	Bias Narratives (BIAS)	Posts revealing social bias, discrimination, or unequal impacts of the disaster across different groups, such as political blame, hate speech, or highlighting the disproportionate suffering of vulnerable populations.
	Assistance & Recovery (ASST)	Posts describing efforts to support recovery and rebuilding after a disaster, such as formal aid from governments or NGOs, informal community-based assistance, long-term infrastructure repair, and expressions of resilience or adaptation by affected communities.
	Socioeconomic Disruption (SECO)	Posts describing economic or social disruptions caused by the disaster, such as financial losses, business interruptions, housing insecurity, educational setbacks, job loss, or the decline of local industries.

Table 3: Definition of Each Disaster Impact Category

	Reddit	TikTok	YouTube
CINJ	332 (3%)	1,207 (3%)	98 (6%)
EVAC	368 (3%)	3,212 (6%)	64 (4%)
INFR	**1,720 (18%)**	**12,118 (25%)**	**681 (40%)**
ENVD	738 (8%)	2,662 (6%)	212 (12%)
RSRC	174 (2%)	2,237 (4%)	11 (1%)
PUBH	155 (2%)	210 (1%)	11 (1%)
EMOT	623 (6%)	3,368 (7%)	25 (1%)
BIAS	504 (5%)	1,964 (4%)	20 (1%)
ASST	866 (9%)	7,048 (15%)	336 (20%)
SECO	1603 (17%)	2,414 (5%)	90 (5%)
OTHER	2,583 (27%)	11,481 (24%)	159 (9%)

Table 4: Physi-Social Impact Distribution (Hurricanes)

	Reddit	TikTok	YouTube
CINJ	144 (3%)	458 (3%)	92 (7%)
EVAC	283 (5%)	1,345 (8%)	160 (12%)
INFR	541 (9%)	2,745 (16%)	284 (20%)
ENVD	**1312 (21%)**	**5,467 (30%)**	**508 (37%)**
RSRC	28 (1%)	80 (1%)	1 (1%)
PUBH	89 (1%)	336 (2%)	21 (2%)
EMOT	204 (3%)	724 (4%)	23 (2%)
BIAS	462 (7%)	790 (5%)	15 (1%)
ASST	1,137 (18%)	3,621 (20%)	162 (12%)
SECO	986 (16%)	1,282 (7%)	51 (4%)
OTHER	1,018 (16%)	719 (4%)	22 (2%)

Table 5: Physi-Social Impact Distribution (Wildfires)

team, while "wildfire" can be used metaphorically to describe rapidly spreading content or emotionally charged discussions. In addition, some posts use disaster-related hashtags solely for promotional purposes, such as adding #hurricane or #wildfire to increase exposure and attract more attention to their advertisement. These practices often result in posts with irrelevant content being included in the dataset.

Inspired by Yao et al. (2025) and Matoshi et al. (2025), we prompt a Multimodal Large Language Model (MLLM) to label the relevance of each post based on its multimodal content (textual descriptions, images, and videos). We applied separate filtering prompts to each of the hurricane and wildfire datasets, by classifying each post as relevant or irrelevant to pertaining disaster type. Posts that mentioned unrelated events, were removed to ensure that only content related to the disasters was retained. This cleaning process al-

lowed us to reduce the noise from off-topic or promotional posts and focus on the target events, while maintaining scalable volumes of content. We used Gemini-2.0-flash as the MLLM because of its cost efficiency and advanced capabilities over multimodal inputs (Jegham, Abdelatti, and Hendawi 2025; Hirosawa et al. 2024; Team et al. 2024). The detailed prompt for each disaster can be found in the Appendix Figures 9 and 10. In addition, we validated the MLLM-generated annotations through human review on a randomly sampled subset. The results show high agreement between human annotators and between human and MLLM annotations, with detailed results reported in Section *Data Verification*. From Tables 1 and 2, a total of 59,294 posts were identified as relevant to hurricanes, and 25,110 posts were identified as relevant to wildfires.

Quantify Physi-Social Impact Index

In this section, we propose a two-stage framework, DisImpact, to quantify the physi-social impacts of disasters. In the first stage, we prompt an MLLM to annotate each social media post with one of ten predefined impact categories, with five categories representing physical impacts and the other five representing social impacts. The model analyzes the post's textual and visual content together to assign the most dominant impact category from the predefined set, following the single-label annotation scheme of prior disaster-related social media datasets such as CrisisMMD (Alam, Ofli, and Imran 2018) and HumAID (Alam et al. 2021). An "Other" label is included for posts that do not fit any of the ten predefined categories. In the second stage, we compute category-level impact index for each temporal window by integrating 1) the relative representation of each category within the window and 2) a window-specific weight that reflects the intensity of activity. The resulting indices reflect both prominence of categories and level of public engagement.

Data Annotation

In the first stage of the DisImpact framework, our goal is to map each social media post to a disaster impact category. We first define ten impact categories informed by prior research, including five categories representing physical impacts and five representing social impacts. The ten disaster impact categories were derived through a multi-stage synthesis process combining prior literature and expert consultation. First, we conducted a structured review of disaster-related studies to extract comprehensive impact concepts in physical and social domains, such as infrastructure damage, environmental degradation, casualties, health outcomes, socioeconomic disruption, and inequality. The detailed summation appears at *Section Physical and Social Impacts in Disaster*. Second, we grouped conceptually similar impacts such as housing damage and infrastructure failure, and political blame and hate speech toward vulnerable groups. Third, experts from interdisciplinary areas were discussed to refine category boundaries and enforce mutual exclusivity at the conceptual level. Impacts were assigned to either the physical or social domain based on whether they primarily reflected material consequences (e.g., casualties, environmental damage, infrastructure loss) or societal and human dimensions (e.g., socioeconomic disruption, inequality, psychological stress). Table 3 presents the list of categories and their definitions.

To map the content of social media posts to the defined impact category, the most straightforward method is to manually label each post. However, this approach is not only time-consuming and labor-intensive but also financially costly, as our dataset includes over 80,000 relevant posts. To address this challenge, we propose an MLLM-driven approach, supplemented with human verification to assess the reliability of the label. Specifically, the model is prompted to assign the most dominant disaster impact category based on the comprehensive understanding of each post's multimodal content. While disaster-related posts may in reality reflect multiple co-occurring impacts, the single-label annotation scheme follows prior disaster-related social media datasets such as CrisisMMD (Alam, Ofli, and Imran 2018) and HumAID (Alam et al. 2021) to enable scalable and reproducible annotation at a large-scale dataset. This design provides clear and consistent signals for large-scale annotation, avoiding ambiguity that can arise when multiple overlapping labels are assigned to a single post. To protect user privacy, all tagged usernames, starting with "@", are replaced with "@user". The detailed prompt can be found in the Appendix Figure 11. If the content of the post does not correspond to any of the predefined disaster impact categories, the model is instructed to assign the label as "Other". Similar to the data cleaning task, we employ Gemini-2.0-Flash as the MLLM due to its strong multimodal reasoning capabilities, rapid inference speed, and low computational cost, which allow efficient processing of large-scale dataset (Jegham, Abdelatti, and Hendawi 2025; Hirosawa et al. 2024; Team et al. 2024). To ensure the reliability of the annotations, we randomly sample a subset of posts for manual verification by human annotators from diverse disciplines. As shown in Section *Data Verification*, we observe strong agreement both among human annotators and between human annotators and the MLLM, demonstrating the reliability of the annotations.

Tables 4 and 5 present the distribution of disaster impact categories for hurricanes and wildfires, respectively. The largest value in each platform is highlighted in bold and the second-largest value is indicated with underline. We observe both commonalities and divergences across different disasters. For instance, from Tables 4 and 5, both disasters consistently show high public concern regarding assistance and recovery. This indicates a shared public focus on disaster support and community rebuilding, reflecting a fundamental dimension of societal response that transcends specific disasters. In addition to the commonalities, we also identify differences between disasters. As shown in Table 4, discussions during hurricanes focus more on infrastructure damage, whereas Table 5 reveals that environmental damage is more frequently mentioned during wildfires. This divergence is likely attributed to the pervasive dispersion of wildfire smoke, which can impact regions far beyond the immediate burn zones, thereby elevating broader environmental concerns compared to the typically localized effects of hurricanes. Moreover, platform-level differences also emerge. For example, discussions of socioeconomic disruption ap-

	Human-Human Consistency	Fleiss' κ	Human-MLLM Consistency	Cohen's κ
H-Clean	0.9067	0.6794	0.9533	0.7828
W-Clean	0.7533	0.6450	0.9100	0.8112
H-Anno	0.7133	0.7796	0.9056	0.8895
W-Anno	0.7267	0.7816	0.8800	0.8621

Table 6: Annotation Verification Results. H stands for Hurricane, and W stands for Wildfire.

pear more prominent on Reddit compared to other platforms.

Data Verification

To verify the reliability of the annotation, we invite three human annotators across different disciplines to manually verify the results, similar to the work by Manakul, Liusie, and Gales (2023); Liu et al. (2025). Specifically, we conducted human verification for two tasks: data cleaning and disaster impact classification. For the data cleaning task, we randomly sampled a subset of posts from the raw collected dataset for each annotator. For the disaster impact classification task, we separately sampled posts from the cleaned relevant dataset. In both tasks, we sampled 150 posts for each disaster–platform pair, including 50 overlapping posts for measuring inter-annotator agreement. This sampling scheme results in 900 posts per annotator for each task, corresponding to all combinations of two disasters and three platforms. To verify the agreement of annotations between human annotators, we adopted two metrics, Consistency and Fleiss' Kappa κ Score, for evaluation. Consistency calculates the proportion of samples that are completely consistent among all annotators, while Fleiss' Kappa Score calculates the overall consistency after random consistency correction between annotators (Fleiss 1971). Let A, B, and C be the labels assigned by each annotator, i represents the index of the data sample, and N is the total number of data samples. The Consistency is defined as: Consistency $= \frac{\sum_{i=1}^{N}(A_i=B_i=C_i)}{N}$. Let $\bar{P}$ denote actual agreement (i.e., the average proportion of actual agreement between annotators) and $\bar{P}_e$ denote the expected agreement (i.e., the expected agreement if annotators randomly choose categories). The Fleiss' Kappa Score is defined as: Fleiss' $\kappa = \frac{\bar{P}-\bar{P}_e}{1-\bar{P}_e}$. Subsequently, we compute Consistency and Cohen's Kappa κ Score to evaluate the agreement between human consensus and MLLM's annotations. Different from Fleiss' Kappa, Cohen's Kappa accounts for chance agreement between two raters and is therefore more appropriate for assessing pairwise annotator reliability (Cohen 1960). Let P_o denote the observed agreement between two raters, and P_e denote the expected agreement by chance. The Cohen's Kappa score is defined as: Cohen's $\kappa = \frac{P_o-P_e}{1-P_e}$.

Table 6 presents the inter-annotator agreement, as well as the agreement between human consensus and the MLLM generated outputs. For both the data-cleaning task and the disaster-impact classification task, the results demonstrate high agreement among human annotators as well as between humans and the MLLM, reflecting the reliability of MLLM.

Impact Index Calculation

In the second stage of the DisImpact framework, we design a method to calculate the intensity of each disaster impact category within fixed-length time windows, based on the category labels assigned to individual posts during the first stage. Following the design in Shang et al. (2025), we segment the timeline into weekly intervals (7 days), allowing for consistent temporal aggregation and comparison across disaster phases. We first calculate the smoothed proportion P (in the interval (0, 1)) of posts assigned to each category c within a given time window t to capture the relative prominence of different impact categories:

$$P_t(c) = \frac{n_t(c) + \alpha}{N_t + \alpha \cdot C} \tag{1}$$

where $n_t(c)$ is the number of posts classified into the category c during the time window t, C is the total number of categories, and $N_t = \sum_{c=1}^{C} n_t(c)$ is the total number of posts from all categories during time window t. The number of posts within a given window can fluctuate greatly depending on the disaster's severity. For example, categories with relatively small absolute counts may be overshadowed during periods of intense public discourse, whereas rare categories may exert a large influence during periods of limited activity due to inflated relative proportions, which distorts the overall distribution. To address this challenge, we adopt additive smoothing instead of raw proportions. The additive smoothing ensures that every category maintains a nonzero contribution across time windows, thereby stabilizing the distributions and mitigating the impact of fluctuations. Following the Jeffreys prior convention (Jeffreys 1998) and the study of Chen and Goodman (1999), we set the smoothing parameter $\alpha = 0.5$ to correct the instability due to small counts while avoiding over-smoothing.

Subsequently, we compute the intensity weight w_t for each time window t to modulate the impact index based on the level of discussion activity. The goal of introducing intensity weight is to amplify the influence of high-volume periods while attenuating the impact of windows with sparse discussions, which are more prone to unstable variations. The design of weight is grounded in the intuition that periods of intense public discourse tend to reflect stronger or more widespread societal impacts, whereas low-activity periods, even if dominated by a specific category, may not signify a meaningful or large-scale effect. During the active phase of a disaster, public attention typically intensifies and social media platforms exhibit substantial discussions across multiple impact-related categories. In such high-activity windows, even categories with small relative proportions can correspond to large absolute volume of posts, and should therefore contribute more strongly to the impact index. Conversely, in later stages when overall discussion volumes decline, a category may dominate in a time window with a small number of remaining posts. Such dominance only reflects isolated or residual conversations rather than sustained or widespread impact. By incorporating an intensity weight that reflects the overall discussion volume in each time window, the index avoids overemphasizing noisy low-volume

Indicator–Disaster	-3 week	-2 week	-1 week	0 week	+1 week	+2 week	+3 week
FEMA–Hurricane (Impacts in Social Domain)	0.440	0.214	-0.008	-0.309	-0.528	-0.552	-0.466
FIRMS–Wildfire (Impacts in Physical Domain)	0.105	-0.228	-0.289	0.424	-0.546	0.533	-0.333

Table 7: Lead-Lag Spearman Correlations with Authorized Ground-truth Data

windows or underestimating categories solely due to small relative proportions, thereby reducing the risk of obscuring key impact patterns and ensuring consistency in the resulting index values. The intensity weight is defined as:

$$w_t = \arctan\left(\frac{N_t - N_{mean}}{\text{IQR}}\right) + \frac{\pi}{2} \tag{2}$$

Specifically, we first measure the deviation between the total number of posts N_t in window t and the overall mean posting volume $N_{\text{mean}} = \frac{1}{T}\sum_{t=1}^{T} N_t$, and normalize the difference using the interquartile range (IQR) of post counts across all time windows. We then apply the arctangent function as a soft activation, which smoothly distinguishes between low- and high-activity periods while preventing sharp spikes caused by extreme outliers. Finally, we add an offset of $\frac{\pi}{2}$ to ensure that all weights remain positive, thereby keeping the resulting impact index values within the interval (0, π). In practice, w_t approaches 0 when the posting volume is far below the mean ($N_t \ll N_{\text{mean}}$), approaches π when posting volumne is far above the mean ($N_t \gg N_{\text{mean}}$), and takes the mid-range value $\frac{\pi}{2}$ when posting volumne approximately equals the mean ($N_t \approx N_{\text{mean}}$). The impact index for category c in time window t is defined as the product of the smoothed proportion $P_t(c)$ and the intensity weight w_t:

$$I_t(c) = P_t(c) \cdot w_t \tag{3}$$

This formulation ensures that the impact index captures both the relative prominence of category c within the window and the overall level of public attention during that period, allowing us to jointly account for content-specific importance and temporal discussion intensity. As a result, the final index reflects both the absolute level of activity within a window and the relative prominence of each impact category. Moreover, because all category-level indices are expressed on a unified scale, they can be flexibly combined as modular components to construct higher-level physical and social impact measures in a modular manner.

Impact Index Analyses

Impact Index Validation

To validate DisImpact, we compare our indices against authoritative ground-truth sources selected for construct validity. Specifically, to reflect community-level social needs, we select FEMA Public Assistance Allocations data, which capture federal recovery funding distributions following disaster events. To represent objective indicators of wildfire intensity, we select NASA FIRMS fire detection data, which provide satellite-based measurements of active fire radiative power. Both external signals are aggregated to weekly intervals to align with the temporal resolution of our impact indices. In particular, we compare social impact indices derived from hurricane-related social media data with FEMA Public Assistance allocations, and physical impact indices derived from wildfire-related data with FIRMS fire-radiative-power measurements.

We use the Spearman rank correlation coefficient (ρ) to quantify monotonic association between the calculated impact indices and ground-truth measurements. Spearman ρ is suitable because it captures monotonic trends without assuming linearity, tolerates scale differences, and is robust to outliers and non-Gaussian distributions (Schober, Boer, and Schwarte 2018). To account for timing differences between real-time social media activity and delayed official reporting, we compute correlations at temporal offsets of ± 3 weeks: negative lags ($\ell < 0$) indicate social media leads, zero lag ($\ell = 0$) indicates concurrent movement, and positive lags ($\ell > 0$) indicate ground truth leads. For each temporal lag ℓ, we correlate the impact index in week t with the ground-truth measurement in week $t+\ell$, allowing us to evaluate how social-media signals align with, lead, or lag behind official indicators. The lead-lag correlation results are available at Table 7. Following conventions in disaster informatics, correlations in the 0.3-0.5 range are typically viewed as meaningful due to the inherent cross-domain noise between social and physical indicators (Kryvasheyeu et al. 2016).

From Table 7, the social impact indices demonstrate construct validity when compared with FEMA Public Assistance data, with the strongest association observed at a 3-week lead. This lead–lag trend indicates that the social impact indices systematically precede FEMA disbursements by approximately three weeks, reflecting its ability to capture emerging community needs earlier than official financial processes. The negative contemporaneous correlation further aligns with the documented delays inherent in FEMA's recovery-phase funding processes. In addition, the physical impact indices show moderate construct validity when evaluated against FIRMS fire-radiative-power measurements. The strongest contemporaneous association occurs at concurring, indicating that physical impact indices can capture real-time wildfire intensity as measured by FIRMS. Together, the hurricane and wildfire validation results provide convergent evidence that the impact indices can produce timely, meaningful, and externally grounded measurements of both social and physical disaster impacts.

Temporal Analysis

We aggregate posts from all social media platforms to investigate the comprehensive impact of each disaster. Figure 1 reports the aggregated physical and social impact indices for hurricanes and wildfires. We subsequently decompose

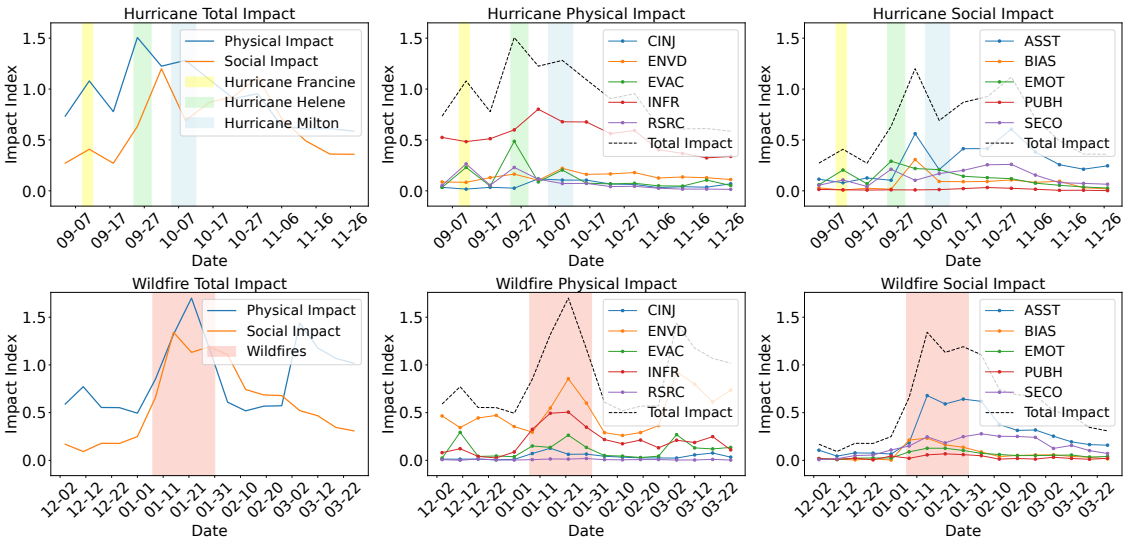

Figure 1: Physical and Social Impact Over Time

these aggregated trends by illustrating the temporal dynamics of each disaster impact category. We observe that both the physical and social impact indices show increases during the core disaster periods. This temporal alignment between peak impact values and known disaster timelines provides support for the validity of the disImpact framework. It indicates that our method is capable of capturing time-sensitive signals that reflect shifts in public attention, concern, and discourse patterns during disasters.

In addition, we observe that the physical impact index decreases rapidly after the disaster, while the social impact index exceeds the physical impact index and remains at a relatively high level after hurricanes. This transition suggests that after the immediate physical consequences of the disasters, public attention shifts more toward social dimensions such as recovery, inequality, and socioeconomic disruption. A similar pattern emerges in the wildfire case, where the social impact index temporarily exceeds the physical impact index after the disaster. However, in contrast to the hurricane case where the physical impact steadily declines after the disaster, the physical impact index in the wildfire case shows a resurgence after an initial decrease. A closer examination of category-level trends reveals that this rebound is primarily driven by a noticeable increase in the environmental damage index. This resurgence in the physical impact index is likely driven by increased public concern over the widespread dispersion of wildfire smoke, which brings renewed attention to environmental damage and broader physical consequences, even after the fire itself has subsided.

Subsequently, we compute the disaster impact index for each category separately for each platform, allowing us to analyze both the convergent and divergent patterns of category-level impact across platforms. Figures 2 and 3 present the impact index for categories in the physical and social domains, respectively. For categories within the physical domain, we observe a high degree of consistency across platforms in terms of impact index patterns. During hurricanes, infrastructure damage stands out as one of the most prominent physical impact categories, indicating a shared emphasis on disruptions such as road blockages, power outages, and building damage. Similarly, in the wildfire dataset, environmental damage emerges as the most dominant physical impact category, reflecting a common concern with ecological consequences such as smoke and habitat loss.

However, for social-domain categories, we observe more pronounced differences across platforms. For instance, the socioeconomic impact index on Reddit is substantially higher than that of other platforms, suggesting that Reddit users tend to engage more with topics related to economic hardship, job loss, and resource access. These findings highlight that different platforms may emphasize different facets of social impact, shaped by their respective user bases, content norms, and communicative affordances. This observation underscores the critical importance of incorporating data from multiple social media platforms, as relying on a single platform may result in an incomplete or biased representation of public discourse.

Spatial Analysis

To examine the spatial dimension of the disaster impact index, we employ two approaches to infer the geographic location of each post. The first way is to directly utilize location metadata if the post was explicitly provided. The second way applies to posts lacking metadata-based geolocation, following prior works (Yao et al. 2025; Vračević et al. 2025) in social media analysis. The distribution of posts labeled by each location method across is reported in Table 8. Posts for which neither method yields a valid geographic reference are excluded from the spatial analysis.

Figures 4 and 5 present the monthly average physical and social impact indices for each U.S. state during the hurricane and wildfire periods, respectively. These two approaches correspond to different sources of spatial informa-

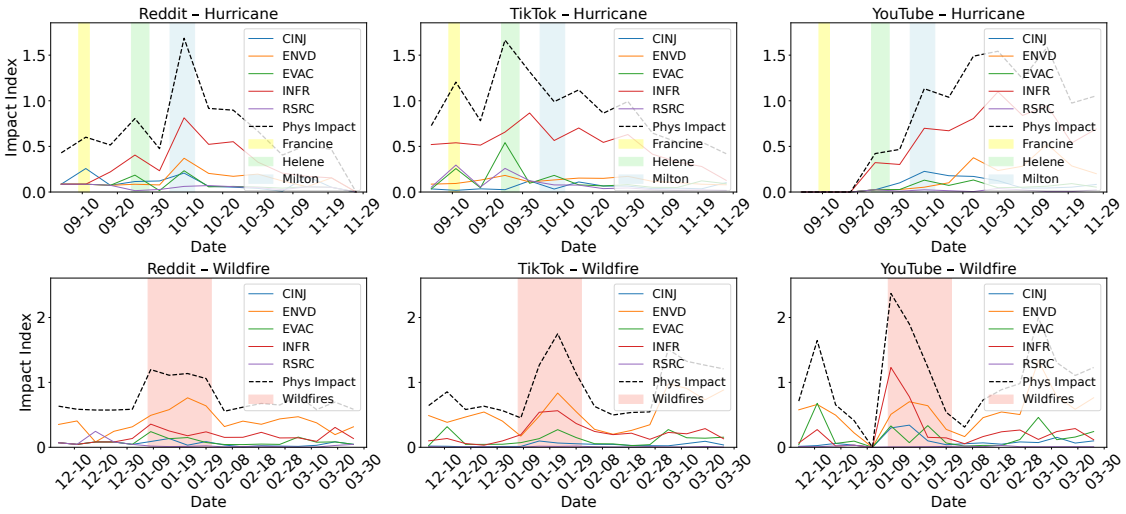

Figure 2: Physical Impact Over Time for each Platform

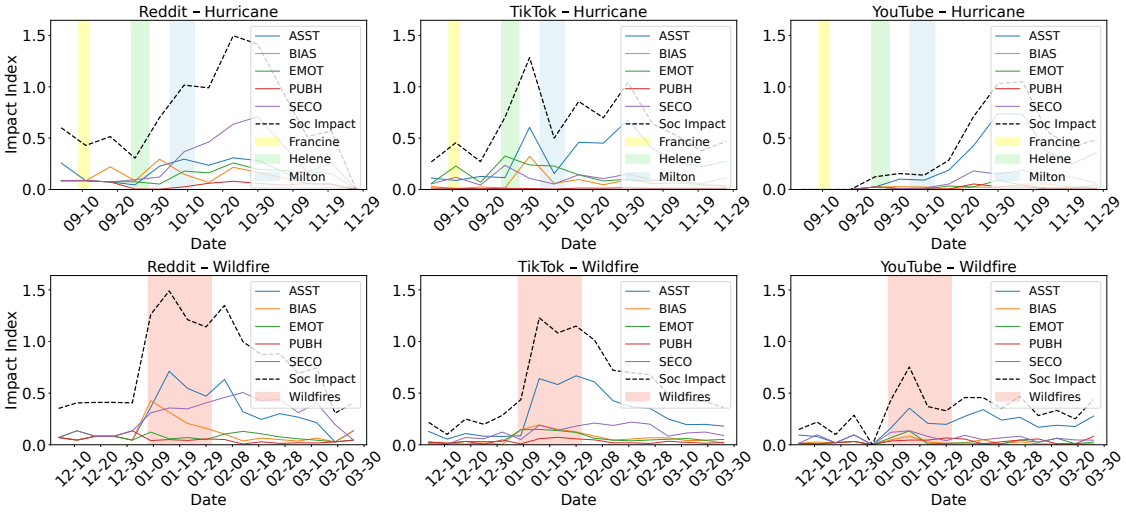

Figure 3: Social Impact Over Time for each Platform

	Reddit		**TikTok**		**YouTube**	
	Hurricane	Wildfire	Hurricane	Wildfire	Hurricane	Wildfire
Posts with Location from Metadata	0 (0%)	0 (0%)	13,903 (38%)	4,061 (24%)	243 (14%)	106 (8%)
Posts with Location from Textual Content	1,374 (19%)	2,014 (39%)	7,707 (21%)	6,945 (41%)	524 (31%)	623 (47%)
Total Number of Posts with Location	1,374 (19%)	2,014 (39%)	21,610 (59%)	11,006 (65%)	767 (45%)	729 (55%)

Table 8: Distribution of Location Information

tion. One captures the location where a post was published, while the other identifies the location referenced in the post content. We therefore visually distinguish them as separate layers in the spatial analysis. Specifically, locations inferred from metadata are shown in blue, while locations inferred from textual content are shown in red and overlaid on top of the metadata-based layer for visualization. In addition, we also integrate the posts with geolocation from both methods and the results are available at Appendix Figures 7 and 8. Through the analysis, we observe that the physical impact index tends to be significantly higher than the social impact index during the corresponding month in disaster-affected states. For example, in the hurricane case, states along the storm trajectory, such as Florida, North Carolina, and Georgia, exhibited markedly elevated physical impact indices in September. Similarly, in the wildfire case, Cali-

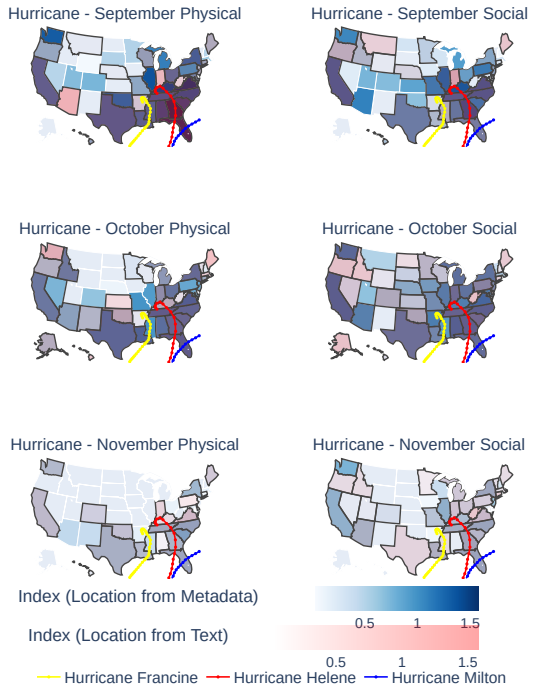

Figure 4: Physi-Social Impact during Hurricanes

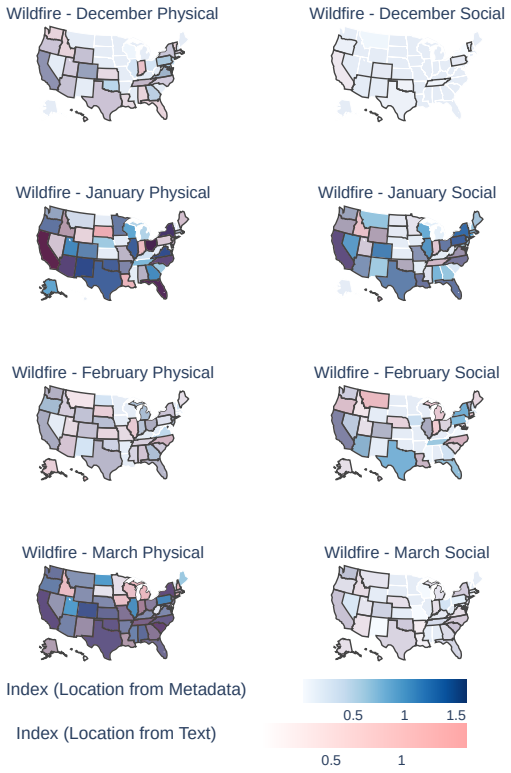

Figure 5: Physi-Social Impact during Wildfires

fornia showed a substantially higher physical impact index in January, reflecting its central role in the wildfire events. This finding suggests that the spatial distribution of physical impact aligns closely with the actual geographic footprint of the disaster, validating the effectiveness of our framework in capturing region-specific disaster effects. Moreover, we find that after the disaster, states outside the directly affected regions often exhibit higher social impact indices than physical ones. This is likely because, although these states are not physically impacted, users in those states still actively engage in discussions around social consequences, such as assistance and recovery efforts.

Conclusion

This paper introduces **DisImpact**, a two-stage framework for quantifying the physicosocial impacts of disasters using multimodal content from multiple social media platforms. In stage one, an MLLM classifies posts into ten predefined disaster impact categories. In stage two, we compute category-specific indices within each time window using a smoothed proportion of posts and a window-level weight reflecting overall disaster-related activity. Since all indices share a unified scale, they can be compared directly or aggregated into broader physical and social impact measures, enabling both fine-grained category analysis and high-level assessments of overall disaster effects. We apply DisImpact to the *2024 Atlantic Hurricanes* and *2025 California wildfires*. Validation against FEMA and NASA FIRMS ground-truth data further demonstrates that the resulting impact indices are timely and reliable. Temporal analysis shows that physical impact indices peak during disaster onset, while social impact indices rise in the aftermath, reflecting shifts in public attention from immediate damage to longer-term social concerns. We also find consistent physical impact patterns across platforms and substantial variation in social impacts, highlighting the value of multi-platform integration.

Limitations and Future Work

Although DisImpact provides a novel and effective framework for quantifying physi-social impacts of disasters, several limitations remain.

1. **Single-Label Annotation**: Our framework adopts a single-label annotation scheme that assigns each post to its most dominant impact category following the setting of prior works (Alam et al. 2021; Alam, Ofli, and Imran 2018). However, it loses information about posts that might contain multiple impact types. Future work could extend DisImpact to multi-label annotation and more compositional impact representations, enabling finer-grained modeling for disaster impacts.
2. **Data Source**: We also acknowledge that using social media data alone cannot fully capture population-level public sentiment or objective disaster impact. For example, our study finds weak signals of psychological distress. This underrepresentation may reflect both underreporting on public platforms and the long-term nature of mental health impacts. Future work should incorporate complementary data sources (e.g., surveys, NGO reports, or community records) to capture more information.

3. **Temporal Scope**: By focusing on posts within a few months before and after disaster events, our framework captures only short-term impacts. Extending the temporal horizon would allow analysis of longer-term recovery processes, including economic resilience, community rebuilding, and mental health outcomes.
4. **Scope of the Unified Index:** Although the unified index enables direct comparison and aggregation across impact categories, it is not intended to replace domain-specific metrics used by practitioners, which often provide higher precision and interpretability within specialized contexts. In practice, the index should be interpreted in conjunction with its underlying category-level composition. While experts from different domains may observe the same index value, examining the category breakdown reveals whether observed changes are driven primarily by physical impacts, social impacts, or their interaction. Accordingly, the unified index supports cross-dimensional and cross-platform comparison without collapsing domain-specific meaning.

Acknowledgments

This research is supported in part by the National Science Foundation under Grant No. CNS-2427070, IIS-2331069, IIS-2202481, IIS-2130263, CNS-2131622. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Abid, S.; Mishra, B. K.; Thakker, D.; and Mishra, N. 2024. Enhancing Trustworthiness and Minimising Bias Issues in Leveraging Social Media Data for Disaster Management Response. *arXiv preprint arXiv:2409.00004*.

Alam, F.; Ofli, F.; and Imran, M. 2018. CrisisMMD: Multimodal Twitter Datasets from Natural Disasters. In *Proceedings of the 12th International AAAI Conference on Web and Social Media (ICWSM)*.

Alam, F.; Qazi, U.; Imran, M.; and Ofli, F. 2021. Humaid: Human-annotated disaster incidents data from twitter with deep learning benchmarks. In *Proceedings of the International AAAI Conference on Web and social media*, volume 15, 933–942.

Basu, R.; Chaudhary, S.; Deval, C.; Sayeed, A.; Herndon, K.; and Griffin, R. 2024. Estimating Disaster Resilience of Hurricane Helene on Florida Counties. *arXiv preprint arXiv:2410.02071*.

Bisiani, S.; Gulyas, A.; Wihbey, J.; and Heravi, B. 2025. UKTwitNewsCor: A Dataset of Online Local News Articles for the Study of Local News Provision. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2371–2384.

Chen, S. F.; and Goodman, J. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4): 359–394.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1): 37–46.

Feng, Y.; Poralla, P.; Dash, S.; Li, K.; Desai, V.; and Qiu, M. 2023. The impact of chatgpt on streaming media: a crowdsourced and data-driven analysis using twitter and reddit. In *2023 IEEE 9th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)*, 222–227. IEEE.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Garfin, D. R.; Thompson, R. R.; Holman, E. A.; Wong-Parodi, G.; and Silver, R. C. 2022. Association Between Repeated Exposure to Hurricanes and Mental Health in a Representative Sample of Florida Residents. *JAMA Network Open*, 5(6): e2217251.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Global Facility for Disaster Reduction and Recovery (GFDRR); European Union; United Nations Development Group; and World Bank. 2017. Post-Disaster Needs Assessment Guidelines, Volume A. Comprehensive guidelines for post-disaster damage, loss, and recovery needs assessment.

Hirosawa, T.; Harada, Y.; Tokumasu, K.; Ito, T.; Suzuki, T.; and Shimizu, T. 2024. Comparative study to evaluate the accuracy of differential diagnosis lists generated by gemini advanced, gemini, and bard for a case report series analysis: cross-sectional study. *JMIR Medical Informatics*, 12: e63010.

Jamal, T. B.; and Hasan, S. 2023. Understanding the loss in community resilience due to hurricanes using Facebook Data. *International journal of disaster risk reduction*, 97: 104036.

Jeffreys, H. 1998. *The theory of probability*. OuP Oxford.

Jegham, N.; Abdelatti, M.; and Hendawi, A. 2025. Visual Reasoning Evaluation of Grok, Deepseek Janus, Gemini, Qwen, Mistral, and ChatGPT. *arXiv preprint arXiv:2502.16428*.

Juarez, R.; Maunakea, A.; Bonham, C.; Bond Smith, D.; et al. 2025. Health and Social Support in the Aftermath of the Maui Wildfires: The Maui Wildfire Exposure Study (MauiWES). *JAMA Network Open*. Community-based cohort 6–14 months post-2023 wildfires.

Jung, Y. S.; Chinthrajah, S.; Johnson, M.; Dresser, C.; et al. 2025. Fine Particulate Matter From 2020 California Wildfires and Mental Health–Related Emergency Department Visits. *JAMA Network Open*.

Keenan, J. M.; Elkins, C.; Hino, M.; Jacobson, M.; and Wing, I. S. 2025. Long-term impacts of hurricanes on mortality among Medicare beneficiaries: Evidence from Hurricane Sandy. *Frontiers in Public Health*, 13: 1239419.

Kryvasheyeu, Y.; Chen, H.; Obradovich, N.; Moro, E.; Van Hentenryck, P.; Fowler, J.; and Cebrian, M. 2016. Rapid assessment of disaster damage using social media activity. *Science advances*, 2(3): e1500779.

Li, X.; Ma, J.; Li, B.; and Mostafavi, A. 2025. Quantifying the Social Costs of Power Outages and Restoration Disparities Across Four US Hurricanes. *arXiv preprint arXiv:2509.02653*.

Lindsey, R. 2025. The Weather and Climate Influences on the January 2025 Fires around Los Angeles. NOAA Climate.gov, accessed November 24, 2025.

Liu, Y.; Maussymbayeva, A.; Murzakhmetov, A.; Nemerenco, A.; Li, Y.; and Wang, D. 2025. Reasoning-based uncertainty estimation for scalable multidimensional media bias annotation: A benchmark across diverse media spaces. *Knowledge-Based Systems*, 115058.

Manakul, P.; Liusie, A.; and Gales, M. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Mansur, A. V.; Langhorn, G.; and Nelson, D. R. 2024. Exploring Social Contracts of Disaster Risk Through Twitter Narratives During a Major Storm. SSRN Electronic Journal. Preprint. Analysis of inequity and blame narratives around government infrastructure/response using Twitter data during Hurricane Ida.

Marshall, J.; and Wang, D. 2016. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *Proceedings of the 10th ACM conference on recommender systems*, 167–174.

Matoshi, V.; De Vuono, M. C.; Gaspari, R.; Kröll, M.; Jantscher, M.; Nicolardi, S. L.; Mazzola, G.; Rauch, M.; Sabol, V.; Salhofer, E.; et al. 2025. One size fits all: Enhanced zero-shot text classification for patient listening on social media. *Frontiers in Artificial Intelligence*, 7: 1397470.

Miller, V. E.; Fitch, K. V.; Swilley-Martinez, M. E.; Agha, E.; Alam, I. Z.; Kavee, A. L.; Cooper, T.; Gaynes, B. N.; Carey, T. S.; Goldston, D. B.; Ranapurwala, S. I.; and Pence, B. W. 2025. Impact of Hurricanes and Floodings on Mental Health Outcomes Within the United States: A Systematic Review and Meta-Analysis. *Disaster Medicine and Public Health Preparedness*, 18: e335.

Mitchell, A.; Maheen, H.; and Bowen, K. 2024. Mental health impacts from repeated climate disasters: an Australian longitudinal analysis. *The Lancet Regional Health–Western Pacific*, 47.

Murtfeldt, R.; Paik, S.; Alterman, N.; Kahveci, I.; and West, J. D. 2024. RIP Twitter API: A eulogy to its vast research contributions. *arXiv preprint arXiv:2404.07340*.

Paglino, E.; Raquib, R. V.; and Stokes, A. C. 2025. Excess deaths attributable to the Los Angeles wildfires from January 5 to February 1, 2025. *JAMA*, 334(11): 1018–1019.

Parks, R. M.; Benmarhnia, T.; Sheridan, P.; Hsu, A.; Burkart, K.; Anderson, G. B.; and Bell, M. L. 2021. Tropical cyclone exposure is associated with increased hospitalization. *Nature Communications*, 12: 6561.

Pehlivan, Z.; Park, S.; Abrahams, A. S.; Desblancs, M. J. P.; Steel, B. D.; and Bridgman, A. 2025. Can-PolNews: A Multi-Platform Dataset of Political Discourse in Canada. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, 2550–2559.

Schober, P.; Boer, C.; and Schwarte, L. A. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5): 1763–1768.

Shang, L.; Chen, B.; Liu, S.; Zhang, Y.; Zong, R.; Vora, A.; Cai, X.; Wei, N.; and Wang, D. 2025. SIDE: Socially Informed Drought Estimation Toward Understanding Societal Impact Dynamics of Environmental Crisis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28359–28367.

Sihag, M.; Li, Z. S.; Dash, A.; Arony, N. N.; Devathasan, K.; Ernst, N.; Albu, A. B.; and Damian, D. 2023. A data-driven approach for finding requirements relevant feedback from tiktok and youtube. In *2023 IEEE 31st International Requirements Engineering Conference (RE)*, 111–122. IEEE.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Vračević, M.; Škunca, S.; Benatallah, B.; et al. 2025. The purpose of alternative geo-social media data in disaster research: A case study of Hurricane Ian. *Social Network Analysis and Mining*, 15(1): 68.

Wang, D.; Kaplan, L.; Abdelzaher, T.; and Aggarwal, C. C. 2012. On scalability and robustness limitations of real and asymptotic confidence bounds in social sensing. In *2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, 506–514. IEEE.

Wertis, L.; Runkle, J. D.; Sugg, M. M.; and Singh, D. 2023. Examining Hurricane Ida's impact on mental health: results from a quasi-experimental analysis. *Geohealth*, 7(2): e2022GH000707.

Wilcox, C.; and Jacobs, P. 2024. Hurricane-battered researchers assess damage. *Science (New York, N.Y.)*, 386: 367.

Yang, Y.; Liu, H.; Mostafavi, A.; and Tatano, H. 2025. Review on modeling the societal impact of infrastructure disruptions due to disasters. *Reliability Engineering & System Safety*, 110879.

Yao, R.; Murzakhmetov, A.; Pillai, R.; Maussymbayeva, A.; Li, Z.; Liu, Y.; Liu, Y.; Shang, L.; Zhang, Y.; Wei, N.; et al. 2025. MASH: A Multiplatform and Multimodal Annotated Dataset for Societal Impact of Hurricane. *arXiv preprint arXiv:2509.23627*.

Zhang, Y.; Zong, R.; Shang, L.; Zeng, H.; Yue, Z.; Wei, N.; and Wang, D. 2023. On Optimizing Model Generality in AI-based Disaster Damage Assessment: A Subjective Logic-driven Crowd-AI Hybrid Learning Approach. In *IJCAI*, 6317–6325.

Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes
   (e) Did you describe the limitations of your work? Yes
   (f) Did you discuss any potential negative societal impacts of your work? Yes
   (g) Did you discuss any potential misuse of your work? Yes
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA
   (e) Did you address potential biases or limitations in your theoretical framework? NA
   (f) Have you related your theoretical results to the existing literature in social science? NA
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? NA
   (b) Did you include complete proofs of all theoretical results? NA
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes
   (b) Did you mention the license of the assets? Yes
   (c) Did you include any new assets in the supplemental material or as a URL? Yes
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? NA
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
   (d) Did you discuss how data is stored, shared, and deidentified? NA

	Hurricane		Wildfire	
	Physical Domain	Social Domain	Physical Domain	Social Domain
All Relevant Posts	25,945 (57%)	19,281 (43%)	13,448 (58%)	9,903 (42%)
Posts with Location from Metadata	8,792 (62%)	5,339 (38%)	2,671 (64%)	1,493 (36%)
Posts with Location from Textual Content	5,852 (60%)	3,836 (40%)	5,507 (56%)	4,254 (44%)

Table 9: Distribution of Posts in Physical and Social Domain for Different Settings.

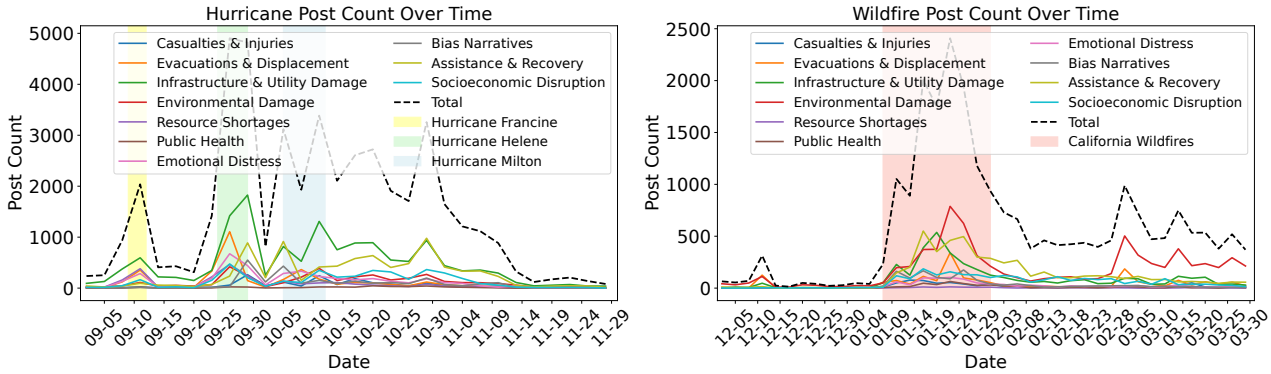

Figure 6: Post Count of each Category Over Time

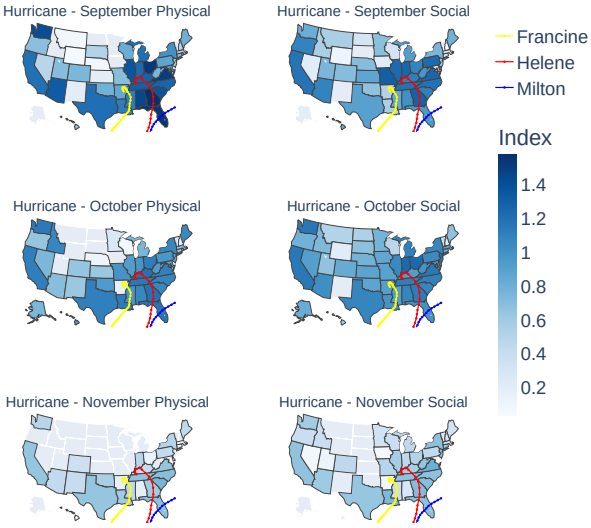

Figure 7: Physi-Social Impact Across U.S. during Hurricanes

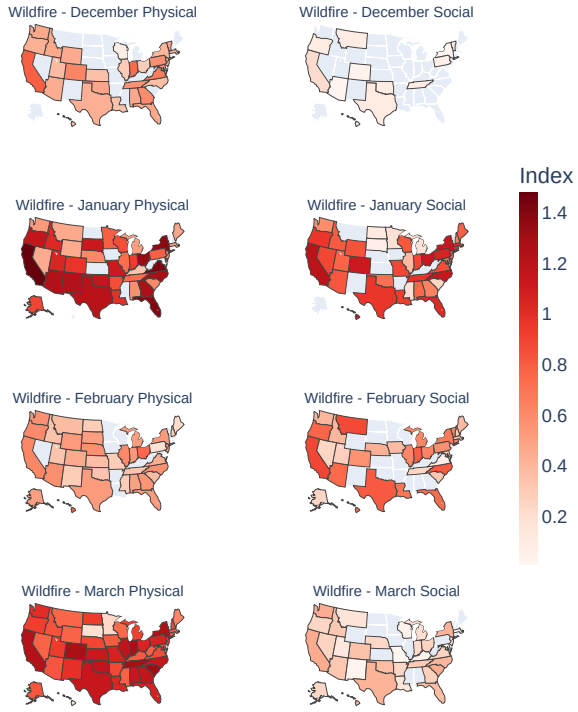

Figure 8: Physi-Social Impact Across U.S. during Wildfires

Prompt

Read the post, considering text, image, and video together. Determine whether the post is related to an actual hurricane disaster in North America (especially Hurricane Helene and Hurricane Milton) and give your reason.

Accepted examples:
- Hurricane disaster, even if it's not Helene or Milton

Some typical counter examples:
- Miami Hurricanes football team, Carolina Hurricanes hockey team, or other sports names
- WWE wrestler called 'Hurricane' or other people called 'Hurricane'
- Advertisement
- Use hurricane as a metaphor to describe other things
- The post contains hurricane-related hashtag, but the actual content is not related to the hurricane disaster
- Cyclone or Typhoon that does not happen in North America

Respond with:
{"Judgment": True or False }

Figure 9: Prompt for Hurricane Data Cleaning.

Prompt

Read the post, considering text and video together. Determine whether the post is related to a wildfire disaster in North America (especially LA wildfires) and give your reason.

Accepted examples:
- Wildfire disaster in North America, even if it is not in California

Some typical counter examples:
- Sports teams, movies, songs, or products with "Wildfire" in the name
- Use of "wildfire" as a metaphor (e.g., "the rumor spread like wildfire")
- Wildfire events clearly located outside North America (e.g., Australia, Greece)
- Other disasters such as hurricanes, floods, earthquakes, or tornadoes
- Posts with wildfire-related hashtags, but actual content not about the wildfire disaster
- Advertisements or promotions unrelated to wildfire disasters

Respond with:
{"Judgment": True or False }

Figure 10: Prompt for Wildfire Data Cleaning.

Prompt

You are a disaster-focused social media classification assistant. Your task is to analyze social media posts and determine which one of the predefined impact categories best describes the post, based on its text, image, and video content. If none of the categories apply, return 11 (Other / Not relevant).

Please select the single most appropriate category from the list below.

- Casualties & Injuries (1): Posts describing people or animals who are killed, seriously injured, missing, or experiencing immediate medical emergencies as a result of the disaster.
- Evacuations & Displacement (2): Posts describing people being forced to evacuate, or relocations to shelters caused by unsafe conditions during a disaster.
- Infrastructure & Utility Damage (3): Posts reporting physical damage to infrastructure or utility systems, such as roads, buildings, bridges, or disruptions to essential services such as electricity, water, or communication networks.
- Environmental Damage (4): Posts describing harm to the natural environment or resources caused by the disaster, such as damage to ecosystems, agriculture, or coastlines, as well as contamination of water, soil, or air.
- Resource Shortages (5): Posts requesting essential survival resources during a disaster, such as clean water, food, shelter, clothing, or other urgent supplies.
- Public Health (6): Posts describing public health consequences following a disaster, such as outbreaks of infectious diseases or mold-related illnesses, disruption of chronic illness care, and shortages of medical supplies or hospital services.
- Emotional and Psychological Distress (7): Posts describing psychological distress or emotional suffering experienced by oneself or others as a result of a disaster, such as trauma, anxiety, depression, grief, or cumulative emotional strain.
- Bias Narratives (8): Posts revealing social bias, discrimination, or unequal impacts of the disaster across different groups, such as political blame, hate speech, or highlighting the disproportionate suffering of marginalized or vulnerable populations.
- Assistance & Recovery (9): Posts describing efforts to support recovery and rebuilding after a disaster, such as formal aid from governments or NGOs, informal community-based assistance, long-term infrastructure repair, and expressions of resilience or adaptation by affected communities.
- Socioeconomic Disruption (10): Posts describing economic or social disruptions caused by the disaster, such as financial losses, business interruptions, housing insecurity, educational setbacks, job loss, or the decline of local industries such as tourism and retail.
- Other / Not Relevant (11): None of the above categories apply to this post. Return only the number of the most appropriate category (1-11). Do not include explanations unless asked.

Respond with:
{“Judgment”: Your Judgment (1-11 integer) }

Figure 11: Prompt for Physi-Social Classification.