# MASH: A Multiplatform and Multimodal Annotated Dataset for Societal Impact of Hurricane

**Ruichen Yao[1], Aslanbek Murzakhmetov[12*], Raaghav Pillai[1*], Aliya Maussymbayeva[15*], Zelin Li[1], Yifan Liu[1], Yaokun Liu[1], Lanyu Shang[3], Yang Zhang[4], Na Wei[1], Ximing Cai[1], Dong Wang[1]**

[1]University of Illinois Urbana-Champaign, Champaign, Illinois, USA
[2]M.Kh. Dulaty Taraz University, Taraz, Jambyl Region, Kazakhstan
[3]Loyola Marymount University, Los Angeles, California, USA
[4]Miami University, Oxford, Ohio, USA
[1]{ryao8, aslanbek, pillai7, aliyam, zelin3, yifan40, yaokunl2, nawei2, xmcai, dwang24}@illinois.edu
[2]an.murakhmetov@dulaty.kz, [3]lanyu.shang@lmu.edu, [4]zhang981@miamioh.edu, [5]maussymbayevaaliya@gmail.com

## Abstract

Natural disasters cause multidimensional threats to human societies, with hurricanes exemplifying one of the most disruptive events that not only caused severe physical damage but also sparked widespread discussion on social media platforms. Existing datasets for studying societal impacts of hurricanes often focus on outdated hurricanes and are limited to a single social media platform, failing to capture the broader societal impact in today's diverse social media environment. Moreover, existing datasets annotate visual and textual content of the post separately, failing to account for the multimodal nature of social media posts. To address these gaps, we present a multiplatform and **M**ultimodal **A**nnotated Dataset for **S**ocietal Impact of **H**urricane (**MASH**) that includes 98,662 relevant social media data posts from Reddit, X, TikTok, and YouTube. In addition, all relevant social media data posts are annotated in a multimodal approach that considers both textual and visual content on three dimensions: humanitarian classes, bias classes, and information integrity classes. To our best knowledge, MASH is the first large-scale, multi-platform, multimodal, and multi-dimensionally annotated hurricane dataset. We envision that MASH can contribute to the study of hurricanes' impact on society, such as disaster severity classification, public sentiment analysis, disaster policy making, and bias identification.

**Dataset** — https://huggingface.co/datasets/YRC10/MASH
**Online Platform** — https://hurricane.web.illinois.edu

## Introduction

Natural disasters have been a persistent threat to human societies, with hurricanes standing out as one of the most representative disasters that often cause severe destruction and significant societal impacts (Shang et al. 2024, 2025; Alam, Ofli, and Imran 2018; Alam et al. 2021). For example, Category 4 Hurricane Helene and Category 5 Hurricane Milton struck the United States in 2024, resulting in at least 250 casualties and over $300 billion in economic losses (Paciorek et al. 2024; Wilcox and Jacobs 2024). More importantly, beyond their immediate physical devastation, these hurricanes sparked extensive discussions across social media platforms, which played a crucial role in shaping public perception, disseminating timely information, and coordinating relief efforts. In this paper, we present a comprehensive dataset that consolidates multimodal posts (e.g., text, images, videos) from multiple social media platforms to study the societal impact of hurricanes. The dataset is annotated across multiple dimensions (e.g., humanitarian, bias, and information integrity) to facilitate research on the societal impacts of hurricanes. By providing labeled data, this dataset seeks to support future studies on disaster response, bias detection, truth discovery, public sentiment analysis, and disaster policy making during natural disasters.

Existing datasets often focus on hurricanes that formed several years ago. For instance, many widely used datasets such as CrisisMMD (Alam, Ofli, and Imran 2018), HIM-Twitter (Alam et al. 2018), and Eyewitness Messages (Zahra, Imran, and Ostermann 2020) focus on the 2017 hurricane season, which is 8 years ago. However, social media was far less developed at the time than it is today, with a much greater number of users and platform diversity than before. As a result, these datasets have limited effectiveness in studying the societal impacts of hurricanes in the current context. In addition, previous datasets predominantly collected posts from a single platform, X (previously known as Twitter), while neglecting other widely used platforms. By excluding posts from diverse platforms, such datasets overlook critical variations in user behavior and communication styles across different social media. This narrow focus fails to capture the broader spectrum of social media activities, leading to datasets that are insufficient for analyzing the comprehensive societal impact of hurricanes. Furthermore, existing datasets either focus exclusively on the textual components of social media posts or treat textual and visual elements independently (Imran et al. 2013; Alam, Ofli,

*These authors contributed equally.

| Dataset | Collection Time | Social Media | Multi Modal | Humanitarian Anno | Bias Anno | Info Integ Anno |
|---|---|---|---|---|---|---|
| Sandy 2012 (Imran et al. 2013) | 2012 | X (Twitter) | ✗ | ✓ | ✗ | ✗ |
| CrisisMMD (Alam, Ofli, and Imran 2018) | 2017 | X (Twitter) | ✓ | ✓ | ✗ | ✗ |
| HIM-Twitter (Alam et al. 2018) | 2017 | X (Twitter) | ✓ | ✓ | ✗ | ✗ |
| TweetDIS (Tekumalla and Banda 2022) | 2012 - 2017 | X (Twitter) | ✗ | ✓ | ✗ | ✗ |
| Eyewitness Messages (Zahra, Imran, and Ostermann 2020) | 2016 - 2018 | X (Twitter) | ✗ | ✓ | ✗ | ✗ |
| MEDIC (Alam et al. 2022) | 2017 - 2018 | X (Twitter) | ✗ | ✓ | ✗ | ✗ |
| Natural Hazards (Meng and Dong 2020) | 2012 - 2019 | X (Twitter) | ✗ | ✗ | ✗ | ✗ |
| HumAID (Alam et al. 2021) | 2016 - 2019 | X (Twitter) | ✗ | ✓ | ✗ | ✗ |
| **MASH (Ours)** | **2024 (latest)** | **Reddit, TikTok, X, YouTube** | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison with Existing Hurricane Disaster Datasets

and Imran 2018; Alam et al. 2018; Tekumalla and Banda 2022; Zahra, Imran, and Ostermann 2020; Alam et al. 2022, 2021). This fragmented treatment overlooks the inherently multimodal nature of social media posts, thereby impeding a comprehensive multimodal understanding of social media content. This gap highlights the need for more comprehensively annotated and up-to-date datasets that incorporate multimodal content from multiple platforms to provide a comprehensive understanding of hurricanes' societal effects.

Motivated by the identified gaps in existing datasets, we introduce a novel multiplatform and **M**ultimodal **A**nnotated Dataset for **S**ocietal Impact of **H**urricane (**MASH**). In particular, the MASH dataset consists of 98,662 relevant *social media posts* collected from four platforms: Reddit, TikTok, X, and YouTube. All relevant posts are annotated with 15 categories along three dimensions: *humanitarian classes*, *bias classes*, and *information integrity classes* in a multimodal approach that considers both textual and visual content, providing a rich labeled dataset for in-depth analysis. The data collection and annotation framework can also be deployed to other disasters like wildfires and earthquakes.

The development of the MASH dataset is the joint effort of an interdisciplinary research team from information science, computer science, hydrology, and environmental engineering. To the best of our knowledge, MASH is the first large-scale, multimodal, multi-platform, and multi-dimensionally annotated hurricane dataset. We envision that the comprehensiveness of platforms, the multimodality of the data, and the multidimensionality of the annotations will contribute to the understanding of societal impacts of hurricanes, the development of climate adaptation, and the improvement of disaster management strategies.

## Related Works

With the advancement of electronic devices and the ubiquity of the Internet, social media has become a rich resource for exploring the impact of specific events on human society. Recently, several datasets have been collected to study the societal impact of social media posts in the context of hurricanes. Sandy 2012 (Imran et al. 2013), TweetDIS (Tekumalla and Banda 2022), and NaturalHazards (Meng and Dong 2020) are datasets that cover tweets during the 2012 Hurricane Season, especially during Hurricane Sandy. In addition, CrisisMMD (Alam, Ofli, and Imran 2018), HIM-Twitter (Alam et al. 2018), TweetDIS (Tekumalla and Banda 2022), Eyewitness Messages (Zahra, Imran, and Ostermann 2020), MEDIC (Alam et al. 2022), Natural Hazards (Meng and Dong 2020), and HumAID (Alam et al. 2021) primarily capture social media discussions during the 2017 hurricane season, especially during Hurricanes Harvey, Irma, and Maria. Table 1 shows the comparison between the MASH dataset and prior datasets. The hurricanes that prior datasets focus on are all 8 years ago or even more than 10 years ago. With the progress of society and the development of technology, a single social media platform (X) at that time can no longer well reflect the impact of hurricanes on society. In contrast, in this study, we collected posts from four social media platforms (Reddit, X, TikTok, and YouTube) and focused on the recent 2024 hurricane season. In addition, prior datasets are limited to humanitarian class annotation and labeled textual and visual content separately (Imran et al. 2013; Alam, Ofli, and Imran 2018; Alam et al. 2018; Tekumalla and Banda 2022; Zahra, Imran, and Ostermann 2020; Alam et al. 2022, 2021), which restricts their ability to support comprehensive analyses of hurricanes' societal impact.

In comparison, we annotated social media posts in three dimensions (Humanitarian Class, Bias Class, and Information Integrity Class) by jointly analyzing textual and visual content, enabling more robust and multimodal machine learning models for disaster-related social media analysis.

## Data Collection

The MASH dataset contains social media data that focuses on recent hurricanes, especially the two major hurricanes that hit the United States during the fall of 2024: *Hurricane Helene* and *Hurricane Milton*, to obtain comprehensive information about the societal impact of hurricanes. According to National Hurricane Center [1], 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 from these two consecutive hurricanes is estimated to be more than $300 billion (Wilcox and Jacobs 2024). To ensure timely and relevant coverage of the hurricane-related discussions, the data collection period spans from September 1, 2024 to November 30, 2024.

We collected social media posts from four platforms: Reddit, TikTok, X, and YouTube with a total of 130,525 posts. These four platforms are widely used by diverse demographics and communities as they provide users with different ways to disseminate information in the multi-modality form of text, images, and videos (Sihag et al. 2023; Feng et al. 2023). The social media data collection adopted the Reddit API[2], Twikit Scraper[3], TikTok Research Tools[4], PyTok Wrapper[5], and YouTube API[6]. We utilized three keywords: *Hurricane*, *Hurricane Helene*, and *Hurricane Milton* to retrieve relevant content. In addition, we only consider posts written in English in this study. We note that social media data may contain users' personally identifiable information, raising potential concerns regarding user privacy. To ensure the ethics of the research and protect users' privacy, we only collected post information (i.e., ID, post content, and post time) without users' identity information. Table 2 presents the number of posts collected from various social media platforms. For text-centric platforms like Reddit and X, we not only collected the title and description of the post but also collected the attached images and videos. To reduce the cost and workload for further cleaning and annotation tasks, we only collected videos from TikTok and YouTube that are less than five minutes long.

## Data Annotation

The MASH dataset contains annotations for social media posts in three dimensions: *Humanitarian Class*, *Bias Class*, and *Information Integrity Class*. In natural disasters, the main purpose of humanitarian assistance is to save lives, reduce suffering, and rebuild affected communities (Alam et al. 2021). For the humanitarian class annotations, we defined 7 categories to classify the content of posts according to the textual and visual context of the post. The content of social media posts may contain bias that can affect user perception (Wessel et al. 2023; Liu, Li, and Wang 2024). The task of bias class annotation is to identify the presence of defined biases or discriminatory elements in 5 categories within the post. False claims arise in uncertain environments when people face a scarcity of needed information (Muhammed T and Mathew 2022). The information integrity category assesses the factual reliability of posts in the face of hurricane disasters, distinguishing between content with verifiable facts and incorrect content. Together, these annotations provide a comprehensive foundation for analyzing the societal impact of hurricanes on social media posts, enabling research in areas such as detection of false claims, bias, and critical events. Prior to annotation, we performed a cleaning step to filter out irrelevant posts, ensuring that the subsequent labeling process focuses only on meaningful content. In this section, we introduce the annotation methodology, as well as the label distributions and findings.

| | **Reddit** | **X** | **TikTok** | **YouTube** |
|---|---|---|---|---|
| Collected Raw Posts | 12,301 | 48,430 | 67,027 | 2,767 |
| Relevant and Annotated Posts | 9,928 | 39,055 | 47,915 | 1,764 |
| Avg. #Words per Annotated Post | 119 | 34 | 21 | 18 |
| #Images on Annotated Posts | 1,579 | 19,778 | – | – |
| #Videos on Annotated Posts | 878 | 7,294 | 47,915 | 1,764 |

Table 2: Distribution of Collected and Annotated Posts.

### Annotation Methodology

To reduce the time and cost of manual annotation, we leverage a Multimodal Large Language Model (MLLM) to support the labeling process. We recognize that a single judgment by the MLLM may contain errors and not be reliable. Therefore, we design a human-MLLM collaborative framework for both data cleaning and annotation, which includes consistency checks based on multiple rounds of MLLM-generated labels and human verification of MLLM-generated labels.

In the first stage, inspired by the work of Wang et al. (2023) and Chen et al. (2024), the MLLM is queried three times per post to generate three separate labels along with corresponding explanations. The input to the MLLM consists of task-specific prompts (available in the Appendix) and the full content of each social media post, which includes textual descriptions, and includes Images and videos if they are present. To protect user privacy, all tagged usernames, starting with "@", are replaced with "@user". This multi-modal input enables the MLLM to generate labels based on a comprehensive understanding of the post across all available modalities. To measure the consistency of the MLLM's outputs, we compute the entropy of the three MLLM-generated-labels. A zero entropy score, indicating all three predictions are identical, reflects a high degree of confidence in the MLLM's judgment. In such cases, the la-

[1]https://www.nhc.noaa.gov

[2]https://www.reddit.com/dev/api

[3]https://pypi.org/project/twikit

[4]https://developers.tiktok.com/products/research-api

[5]https://pypi.org/project/PyTok

[6]https://developers.google.com/youtube/v3

bel is considered high-confidence and directly adopted.

On the other hand, if the label entropy computed in the first stage is non-zero, indicating that the MLLM exhibits uncertainty regarding the task, we proceed to the second stage. In the second stage, we provide the MLLM with all three sets of labels and corresponding reasons generated in the first stage. The MLLM is prompted to compare, analyze, and determine which label and justification are more reasonable. This setup encourages the model to perform consistency resolution and preference selection, aligning with recent work demonstrating that MLLM benefits from self-consistency (Wang et al. 2023) and iterative self-refinement strategies (Madaan et al. 2023; Shinn et al. 2023) to improve label reliability and reason quality. The MLLM subsequently generates three new sets of labels and reasons, from which the new entropy is computed. If the new entropy value is zero, the new label is also considered as a high-confidence label and accepted accordingly. If the new entropy remains non-zero, indicating persistent uncertainty, the post is escalated to the third stage for human annotation. In this stage, human annotators from diverse disciplines in our team collaboratively review the post and judge the label. Through open discussion, annotators exchange perspectives and reach a consensus on the final label. This human annotation process ensures that difficult or ambiguous cases are carefully reviewed. We present several representative examples in the Appendix to illustrate how final labels were determined in challenging cases.

Furthermore, to verify the reliability of high-confidence labels, we randomly select 500 high-confidence labels from each social media platform in each task, and three human annotators in our team conduct independent verification. The verification results show high consistency and are presented in Section Annotation Consistency. We used Gemini-2.0-flash as the MLLM in both data cleaning and annotation tasks because of its cost efficiency and advanced capabilities over multimodal inputs (Jegham, Abdelatti, and Hendawi 2025; Hirosawa et al. 2024; Team et al. 2023, 2024).

## Data Cleaning

We observe that the collected raw social media posts contain irrelevant and off-topic posts, which could reduce the quality of the dataset. For example, using the keyword "hurricane" in API search may retrieve posts about the Miami Hurricanes football team or Carolina Hurricanes hockey team, which are not related to the hurricane disaster. Additionally, some posts use hurricane-related hashtags solely for promotional purposes, such as adding #hurricane to increase exposure and attract more attention to their advertisement. These practices often result in posts with irrelevant content being included in the dataset. To address this issue, we implement the human-MLLM collaborative annotation framework to keep posts related to actual North American hurricane disasters (particularly Hurricane Helene and Hurricane Milton) and exclude content that literally mentions "hurricane" but is not relevant to an actual disaster, such as sports, entertainment, and metaphors. The detailed prompt can be found in the Appendix. Table 2 presents the distribution of relevant posts for each social media platform after cleaning. A total of 98,662 posts were identified as relevant and subsequently annotated across the three dimensions.

| | Reddit | X | TikTok | YouTube |
|---|---|---|---|---|
| **Cslt** | 891 (9%) | 3,698 (9%) | 4,788 (10%) | 310 (18%) |
| **Evac** | 2,475 (25%) | 8,116 (21%) | 15,009 (31%) | 728 (41%) |
| **Dmg** | **4,340 (44%)** | 13,428 (34%) | 22,898 (48%) | **1,364 (77%)** |
| **Advc** | 2,079 (21%) | 4,844 (12%) | 14,461 (30%) | 334 (19%) |
| **Rqst** | 2,750 (27%) | 7,757 (20%) | 8,804 (18%) | 416 (24%) |
| **Aid** | 3,653 (37%) | 16,874 (43%) | **23,940 (50%)** | 1,015 (58%) |
| **Rcv** | 4,280 (43%) | **18,348 (47%)** | 22,437 (47%) | 1,122 (64%) |
| **OUI** | 2,511 (25%) | 10,614 (27%) | 8,991 (19%) | 145 (8%) |

Table 3: Distribution of Humanitarian Classes

## Humanitarian Class Annotation

The purpose of the humanitarian class annotation is to classify social media posts based on their themes before, during, and after the hurricane disaster, facilitating a semantic understanding of the data sample. This classification also supports the analysis of public communication and information dissemination during natural disasters like hurricanes. Based on previous datasets (Alam, Ofli, and Imran 2018; Zahra, Imran, and Ostermann 2020; Gebru et al. 2021; Alam et al. 2018; Imran et al. 2013; Tekumalla and Banda 2022; Alam et al. 2021), we classify the social media data samples into 7 categories: **Casualty** (Cslt), **Evacuation** (Evac), **Damage** (Dmg), **Advice** (Advc), **Request** (Rqst), **Assistance** (Aid), **Recovery** (Rcv). The detailed definition of these classes is available in the Appendix. If the content of the post does not fit into any of these categories we defined, it is classified as Other Useful Information (OUI).

We utilized the same MLLM-human collaborative framework introduced in Section Annotation Methodology to annotate the humanitarian class of each post. The prompt is presented in the Appendix and the label quality verification is presented in Section Annotation Consistency. Table 3 reports the distribution of humanitarian classes of posts from each social media platform, with the largest value in each platform highlights in bold and the second-largest value indicates with underline. We note that a social media post may contain information that belongs to multiple humanitarian classes. For example, a post describing disaster situations might mention both casualties and damage of infrastructure. Consequently, the cumulative percentages of posts across all humanitarian classes exceed 100% in Table 3. From Table 3 we observe that the categories Damage and Recovery have high prevalence across all platforms, especially for video-based platforms such as TikTok and YouTube, suggesting that users frequently upload and discuss content related to the physical destruction caused by hurricanes and the subsequent recovery efforts.

## Bias Class Annotation

Bias Class annotation aims to identify the existence of bias in social media posts. Gu, Guo, and Zhuang (2021) and Li-

| | Reddit | X | TikTok | YouTube |
|---|---|---|---|---|
| **LB** | **1,914 (19%)** | 10,578 (27%) | **13,371 (28%)** | **241 (14%)** |
| **PB** | 1,798 (18%) | **11,556 (30%)** | 5,405 (11%) | 238 (13%) |
| **GB** | 190 (2%) | 613 (2%) | 2,891 (6%) | 55 (3%) |
| **HS** | 106 (1%) | 1,127 (3%) | 1,249 (3%) | 17 (1%) |
| **RS** | 144 (1%) | 1,124 (3%) | 1,444 (3%) | 27 (2%) |
| **UB** | 7,141 (72%) | 22,332 (63%) | 31,508 (66%) | 1,334 (76%) |

Table 4: Distribution of Bias Classes

fang et al. (2020) point out that social media users tend to express more polarized and biased opinions during emergencies like natural disasters because of tension and anxiety. Based on previous studies (Wessel et al. 2023; Liu, Li, and Wang 2024; Spinde et al. 2023), we define 5 post-level bias types, **Linguistic Bias** (LB), **Political Bias** (PB), **Gender Bias** (GB), **Hate Speech** (HS), **Racial Bias** (RB). The detailed definition of these classes can be found in the Appendix. If the content of the post does not fit into any of the biases we defined, it is classified as Undefined Bias (UB).

Since the only difference between the bias and humanitarian annotations is the definitions of categories, we also adopt the human-MLLM collaborative framework to annotate bias classes. The prompt is presented in the Appendix and the label quality verification is presented in Section Annotation Consistency. Table 4 reports the distribution of bias classes of posts from each social media platform with the largest value in each platform highlights in bold and the second-largest value indicates with underline. As illustrated in Table 4, the majority of posts are free from defined bias, while a subset of posts exhibits one or multiple types of bias.

Table 4 demonstrates that all four platforms exhibit a relatively high ratio of posts classified as Linguistic Bias and Political Bias. This trend may be attributed to the strong emotional reactions triggered by hurricane disasters, which often prompt people to express their views in more intense or confrontational language, thus leading to linguistic bias. Additionally, posts about government relief efforts often reflect political leanings, especially given the close timing of Hurricanes Helene and Milton and the 2024 U.S. presidential election. This overlap in time could exacerbate political discussions and criticisms, further exacerbating the presence of political bias in these posts.

## Information Integrity Class Annotation

The primary goal of the information integrity class annotation task is to determine whether the content described in a post is factual or constitutes false arguments. The information integrity class includes three distinct labels: *True Information*, *False Information*, and *Unverifiable Information*. The inclusion of the Unverifiable Information label accounts for posts that share personal or subjective opinions, which cannot be objectively verified. Different from humanitarian and bias class annotation, information integrity annotation raises unique challenges as it requires factual verification that cannot be reliably performed by the MLLM alone. Additionally, the MLLM we adopted in the annotation framework, Gemini-2.0-flash, does not contain up-to-date knowledge beyond August 2024, making it difficult to verify the factuality of posts published between September to November 2024. To address this limitation, we enabled Gemini's online function (i.e., Grounding with Google Search[7]) to support truth discovery. This functionality allows the model to retrieve relevant external content based on the content of the input post, such as government announcements, truth discovery articles, and news media reports. The retrieved information serves as an external source of evidence, enabling the MLLM to assess the factuality of the post based on real-time and authoritative content. The prompt for the information integrity class annotation is present in the Appendix. Due to the strict rate limits imposed by the Grounding with Google Search, we limited the MLLM to a single round instead of three rounds. To ensure the reliability of the generated labels, human annotators also verify the labels and the results are presented in Section Annotation Consistency.

| | Reddit | X | TikTok | YouTube |
|---|---|---|---|---|
| **True Info** | 7,304 (74%) | 30,392 (78%) | 33,440 (70%) | 1,352 (77%) |
| **False Info** | 746 (7%) | 5,581 (14%) | 8,682 (18%) | 324 (18%) |
| **Unverif. Info** | 1,878 (19%) | 3,082 (8%) | 5,793 (12%) | 88 (5%) |

Table 5: Distribution of Information Integrity Classes

Table 5 reports the distribution of information integrity class for each social media platform. Across the four platforms, the percentage of posts labeled as factual information is consistently around 75%, demonstrating a notable level of uniformity. The proportion of posts labeled as false claims on X, TikTok, and YouTube is relatively consistent around 15%. In contrast, Reddit exhibits a lower proportion of false claims and accompanied by a corresponding increase in the proportion of posts categorized as unverifiable information. This pattern suggests that Reddit may contain fewer explicitly false claims and more content that lacks sufficient evidence for verification.

## Annotation Consistency

Similar to Manakul, Liusie, and Gales (2023), we invite three human annotators from different disciplines to manually verify the annotations by MLLM. Each human annotator is assigned 500 randomly sampled posts for each platform, resulting in 2,000 posts per annotator. For the information integrity class, we focus on true and false claims, thereby reducing all evaluation tasks to binary classification. We adopt Accuracy and Fleiss' Kappa Score for inter-annotator evaluation. Accuracy 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' Kappa is suitable for measuring agreement among three or more raters, making it appropriate for evaluating inter-annotator reliability (Fleiss 1971). Subsequently,

[7]https://ai.google.dev/gemini-api/docs/grounding

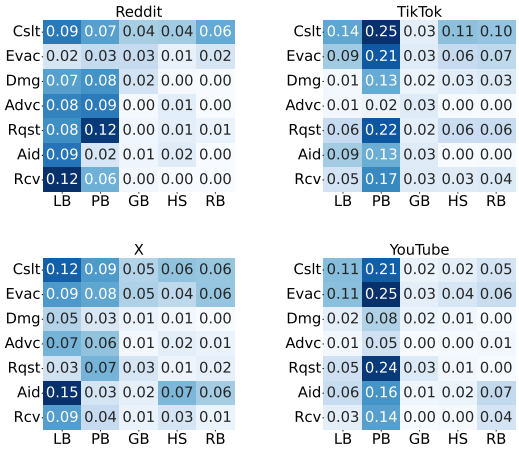
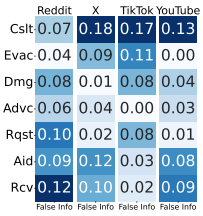
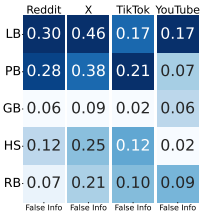



Figure 1: Humanitarian Class vs. Bias Class Correlation




(a) False Info vs. Humanitarian




(b) False Info vs. Bias

Figure 2: (a) False Info vs. Humanitarian Class Correlation, (b) False Info vs. Bias Class Correlation

we apply majority voting to aggregate human annotations. We compute Accuracy and Cohen's Kappa to evaluate the consistency between human consensus and MLLM's annotations. Different from Fleiss' Kappa, Cohen's Kappa quantifies agreement between two raters by correcting for the proportion of agreement that would be expected purely by chance, making it more appropriate for pairwise agreement assessment compared to Fleiss' Kappa (Cohen 1960). Table 6 shows the consistency of the annotations among the three annotators and the consistency between the MLLMs' annotations and the human consensus. We find that human inter-annotator agreement is strong, with Fleiss' Kappa consistently near or above 0.8. Similarly, the agreement between MLLM's label and human consensus is also high, with Cohen's Kappa above 0.80 and accuracy exceeding 0.90, which reflects the reliability of MLLM's annotations. We observe that categories such as Hate Speech and Gender Bias show relatively lower agreement. A plausible explanation is their highly imbalanced label distributions, which make the metrics more sensitive to small annotation disagreements.

## Preliminary Analyses

In this section, we comprehensively analyze the collected relevant social media data samples and their annotations across various dimensions. Specifically, we perform *Correlation Analysis* that explores the correlation relationship between different annotated labels and *Temporal Analysis* studying the time series distribution of annotated labels. In addition, we perform *Spatial Analysis*, investigating the geolocation distribution of the annotated posts; *Sentiment Analysis*, examining the sentiments expressed in relevant posts; and *Baseline Model Analysis*, evaluating labels using several baseline models in the Appendix.

### Correlation Analysis

In correlation analysis, we calculate and analyze the correlations between different annotated labels. In particular, we focus on understanding the connections between annotated labels in different dimensions to uncover potential patterns and interactions within the dataset. We utilize Cramer's $V$ to measure the correlations between the various annotated labels, where higher values indicate stronger correlations. Specifically, we calculate the correlation based on the co-occurrence of multiple labels in the same post, rather than relying on the overall distribution of posts. This approach ensures that correlation reflects the actual relationship between labels and avoids the bias caused by comparing a large number of posts in different categories.

Figure 1 indicates the correlation relationship between humanitarian classes and bias classes. We observe that the Request category is strongly associated with Political Bias across all platforms. One possible reason is that users may criticize government responses or call for policy-level actions to support victims. Similarly, the Casualty class tends to correlate with both Political and Linguistic Bias, implying that posts discussing human loss often incorporate emotional and polarized expression. In contrast, bias categories such as Gender Bias, Hate Speech, and Racial Bias demonstrate little correlation with humanitarian content, likely due to their relatively low prevalence in the dataset.

Figure 2 (a) illustrates the correlations between False Information and the various humanitarian classes. We observe that posts related to the Casualty class exhibit the strongest correlation with False Information across all platforms. This highlights the prevalence of False Information surrounding death and injury reports in the aftermath of disasters. Figure 2 (b) presents the correlations between False Information and the bias classes. We observe that Linguistic and Political Biases exhibit the strongest correlations with False Information across platforms, particularly on X and Reddit. This points out that False Information is often communicated through subjective or ideologically driven language. The presence of Linguistic and Political Bias can serve as an indicator for identifying false information in social media discussions. In contrast, Gender Bias shows consistently low correlation with false content across all platforms, indicating that gender-related content is largely absent from false information in the disaster context.

### Temporal Analysis

We conduct a temporal analysis of the social media posts, focusing on the distribution of annotation labels over time. Figures 3 to 5 illustrate the time series distributions of hu-

| Category | Inter-Annotator Consistency | | | | | | | | Human-MLLM Consistency | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Reddit | | X | | TikTok | | YouTube | | Reddit | | X | | TikTok | | YouTube | |
| | Acc | Fleiss'$\kappa$ | Acc | Fleiss'$\kappa$ | Acc | Fleiss' $\kappa$ | Acc | Fleiss'$\kappa$ | Acc | Cohen's $\kappa$ | Acc | Cohen's $\kappa$ | Acc | Cohen's $\kappa$ | Acc | Cohen's $\kappa$ |
| **Cslt** | 0.93 | 0.81 | 0.95 | 0.81 | 0.98 | 0.92 | 0.97 | 0.92 | 0.98 | 0.92 | 0.98 | 0.90 | 0.99 | 0.94 | 0.98 | 0.94 |
| **Evac** | 0.93 | 0.88 | 0.93 | 0.84 | 0.94 | 0.88 | 0.97 | 0.95 | 0.97 | 0.91 | 0.97 | 0.90 | 0.96 | 0.90 | 0.98 | 0.95 |
| **Dmg** | 0.95 | 0.94 | 0.94 | 0.91 | 0.96 | 0.94 | 0.96 | 0.93 | 0.98 | 0.96 | 0.99 | 0.97 | 0.98 | 0.95 | 0.97 | 0.92 |
| **Advc** | 0.92 | 0.84 | 0.94 | 0.83 | 0.96 | 0.94 | 0.97 | 0.93 | 0.98 | 0.94 | 0.98 | 0.89 | 0.99 | 0.96 | 0.98 | 0.95 |
| **Rqst** | 0.93 | 0.87 | 0.93 | 0.83 | 0.97 | 0.92 | 0.97 | 0.93 | 0.98 | 0.93 | 0.97 | 0.90 | 0.99 | 0.97 | 0.98 | 0.95 |
| **Aid** | 0.94 | 0.90 | 0.93 | 0.90 | 0.95 | 0.93 | 0.97 | 0.96 | 0.97 | 0.94 | 0.97 | 0.94 | 0.97 | 0.94 | 0.98 | 0.97 |
| **Rcv** | 0.93 | 0.91 | 0.93 | 0.91 | 0.96 | 0.94 | 0.96 | 0.95 | 0.97 | 0.93 | 0.98 | 0.96 | 0.97 | 0.94 | 0.98 | 0.96 |
| **LB** | 0.93 | 0.85 | 0.93 | 0.86 | 0.94 | 0.89 | 0.97 | 0.91 | 0.99 | 0.97 | 0.98 | 0.93 | 0.97 | 0.91 | 0.99 | 0.94 |
| **PB** | 0.93 | 0.85 | 0.95 | 0.91 | 0.99 | 0.95 | 0.98 | 0.94 | 0.99 | 0.96 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 | 0.93 |
| **GB** | 0.99 | 0.82 | 0.99 | 0.77 | 0.99 | 0.90 | 0.99 | 0.87 | 0.99 | 0.91 | 1.00 | 1.00 | 0.99 | 0.92 | 0.99 | 0.79 |
| **HS** | 0.99 | 0.66 | 0.99 | 0.86 | 0.99 | 0.95 | 0.99 | 0.83 | 0.99 | 0.86 | 0.99 | 0.94 | 1.00 | 1.00 | 0.99 | 0.86 |
| **RB** | 0.99 | 0.79 | 0.99 | 0.89 | 0.99 | 0.86 | 0.99 | 0.92 | 1.00 | 1.00 | 0.99 | 0.96 | 0.99 | 0.94 | 0.99 | 0.95 |
| **InfoInt** | 0.91 | 0.85 | 0.94 | 0.87 | 0.88 | 0.81 | 0.95 | 0.91 | 0.95 | 0.88 | 0.97 | 0.91 | 0.92 | 0.82 | 0.97 | 0.93 |

Table 6: Inter-Annotator and Human–MLLM Consistency by each Category

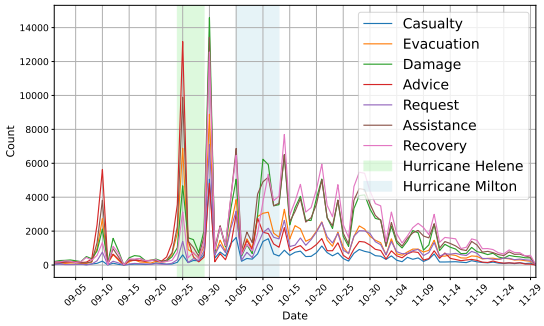



Figure 3: Distribution of Humanitarian Classes Over Time

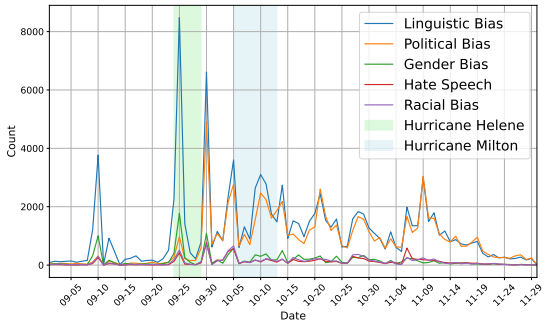



Figure 4: Distribution of Bias Classes Over Time

manitarian classes, bias classes, and information integrity classes respectively. We observe a significant increase in the number of posts during the periods when Hurricane Helene and Hurricane Milton occurred, reflecting an increase in user activity in response to these events.

For the humanitarian classes, the categories Request and Advice are most prominent during the early stages of the hurricane, reflecting how users turn to social media to seek and offer help in anticipation of the disaster. As the hurricane strikes and its impact becomes visible, posts increasingly shift toward the Damage category, capturing reports of destruction and loss. In the aftermath, the Recovery category becomes more dominant, indicating a collective focus on rebuilding efforts.

For the bias categories analysis, linguistic bias and political bias account for the majority of biased content. Notably, the distribution of both biases increases significantly during hurricanes. The increase in linguistic bias reflects that public discourse becomes more intense during disaster events, as individuals express stronger emotions in response to the crisis. Similarly, the increase in political bias is often driven by criticism of government relief efforts as public scrutiny of relief measures grows. These findings suggest that the impacts of hurricanes are not limited to physical dimensions, such as individual casualties and damage to infrastructure. Hurricane disasters also affect public discourse, highlighting the broad impact on society.

For information integrity classes, the amount of both True Information and False Information increases significantly during the hurricanes. We observe that during the period of Hurricane Helene and Hurricane Milton, the proportion of false claims was notably high, peaking at over 40%. In the aftermath, this proportion dropped and stabilized at around 20%. This increase shows that as the level of information dissemination increases, the amount of true and false claims also increases. This increase also highlights the urgent need to verify facts during disasters, because the urgency of sharing information during disasters often leads to the spread of false claims. The spread of false claims could further cause public tension about natural disasters.

## Conclusion and Discussion

This paper introduces a multiplatform and **M**ultimodal **A**nnotated Dataset for **S**ocietal Impact of **H**urricane (**MASH**) to facilitate the understanding of societal impact of hurricanes. Specifically, the MASH dataset includes 98,662

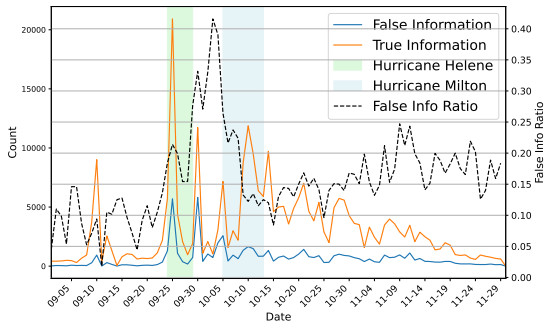



Figure 5: Distribution of Info Integrity Classes Over Time

relevant social media posts and is multimodal annotated in the three dimensions of humanitarian class, bias class, and information integrity class. By leveraging annotations in three dimensions, this dataset provides new opportunities for research on the social impact of hurricanes such as the spread of false claims during disaster events, the expression of public sentiment and bias, and the evolution of humanitarian needs over time. Moreover, the data collection and annotation framework introduced in this paper is generalizable to other disaster contexts across diverse social media platforms. However, this study mainly analyzes social media posts in English, which may lead to the loss of useful information in other languages. In future work, we plan to leverage the machine translation technology to better understand the impact of hurricanes on communities using different languages.

## References


Alam, F.; Alam, T.; Hasan, M. A.; Hasnat, A.; Imran, M.; and Ofli, F. 2022. MEDIC: a multi-task learning dataset for disaster image classification. *Neural Comput. Appl.*, 35(3): 2609–2632.

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.; Ofli, F.; Imran, M.; and Aupetit, M. 2018. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *Proc. of ISCRAM, Rochester, USA*.

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.

Chen, X.; Aksitov, R.; Alon, U.; Ren, J.; Xiao, K.; Yin, P.; Prakash, S.; Sutton, C.; Wang, X.; and Zhou, D. 2024. Universal Self-Consistency for Large Language Models. In *ICML 2024 Workshop on In-Context Learning*.

Clark, K.; Luong, M.-T.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR*.

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.

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.

Gu, M.; Guo, H.; and Zhuang, J. 2021. Social media behavior and emotional evolution during emergency events. In *Healthcare*, volume 9, 1109. MDPI.

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.

Imran, M.; Elbassuoni, S.; Castillo, C.; Diaz, F.; and Meier, P. 2013. Practical extraction of disaster-relevant information from social media. In *Proceedings of the 22nd international conference on world wide web*, 1021–1024.

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*.

Jiang, Z.-H.; Yu, W.; Zhou, D.; Chen, Y.; Feng, J.; and Yan, S. 2020. ConvBERT: Improving BERT with Span-based Dynamic Convolution. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12837–12848. Curran Associates, Inc.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Lifang, L.; Zhiqiang, W.; Hong, W.; et al. 2020. Effect of anger, anxiety, and sadness on the propagation scale of social media posts after natural disasters. *Information Processing & Management*, 57(6): 102313.

Liu, Y.; Li, Y.; and Wang, D. 2024. Intertwined Biases Across Social Media Spheres: Unpacking Correlations in Media Bias Dimensions. *arXiv preprint arXiv:2408.15406*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.; Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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*.

Mashkoor Siddiqui, S.; Sheikh, M. A.; Aleem, M.; and Singh, K. R. 2025. Comparative Analysis of Efficient Adapter-Based Fine-Tuning of State-of-the-Art Transformer Models. *arXiv e-prints*, arXiv–2501.

Meng, L.; and Dong, Z. S. 2020. Natural hazards Twitter dataset. *arXiv preprint arXiv:2004.14456*.

Muhammed T, S.; and Mathew, S. K. 2022. The disaster of misinformation: a review of research in social media. *International journal of data science and analytics*, 13(4): 271–285.

Naseer, M.; Asvial, M.; and Sari, R. F. 2021. An empirical comparison of bert, roberta, and electra for fact verification. In *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 241–246. IEEE.

Paciorek, E.; Zinnbauer, J.; Sweeney, K.; Schaaf, B.; and Sartori, N. 2024. 2024 Atlantic Hurricane Season.

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.

Shang, L.; Chen, B.; Vora, A.; Zhang, Y.; Cai, X.; and Wang, D. 2024. SocialDrought: A Social and News Media Driven Dataset and Analytical Platform towards Understanding Societal Impact of Drought. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 2051–2062.

Shinn, N.; Cassano, F.; Gopinath, A.; Narasimhan, K. R.; and Yao, S. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Spinde, T.; Hinterreiter, S.; Haak, F.; Ruas, T.; Giese, H.; Meuschke, N.; and Gipp, B. 2023. The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias. *arXiv preprint arXiv:2312.16148*.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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*.

Tekumalla, R.; and Banda, J. M. 2022. TweetDIS: A Large Twitter Dataset for Natural Disasters Built using Weak Supervision. In *2022 IEEE International Conference on Big Data (Big Data)*, 4816–4823.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.

Wessel, M.; Horych, T.; Ruas, T.; Aizawa, A.; Gipp, B.; and Spinde, T. 2023. Introducing MBIB-the first media bias identification benchmark task and dataset collection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2765–2774.

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

Zahra, K.; Imran, M.; and Ostermann, F. O. 2020. Automatic identification of eyewitness messages on twitter during disasters. *Information processing & management*, 57(1): 102107.

Zhang, W.; Deng, Y.; Liu, B.; Pan, S.; and Bing, L. 2024. Sentiment Analysis in the Era of Large Language Models: A Reality Check. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 3881–3906. Mexico City, Mexico: Association for Computational Linguistics.

## Reproducibility Checklist

---

**Instructions for Authors:**

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

**Example:** If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

### 1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) Yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) Yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) Yes

### 2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) No

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) NA

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) NA

2.4. Proofs of all novel claims are included (yes/partial/no) NA

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) NA

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) NA

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) NA

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) NA

### 3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) Yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) Yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) Yes

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) Yes

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) Yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) Yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

### 4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) Yes. The paper includes only a baseline analysis in the Appendix rather than full computational experiments.

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) NA

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) Yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) Partial

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) Yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) Yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/par-

tial/no/NA) Yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) Yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) Yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) Yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) No

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) NA

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) Partial

## Overview of Appendix

We perform *Spatial Analysis*, investigating the geolocation distribution of the annotated posts; *Sentiment Analysis*, examining the sentiments expressed in relevant posts; *Baseline Model Analysis*, evaluating annotated labels using several baseline models; definition and MLLM's prompt for each task; and examples of MLLM's uncertain annotation, providing examples of MLLM's uncertain posts and our human judgement with reason.

## Spatial Analysis

To examine the geographic distribution of posts in MASH dataset, we conduct a spatial analysis based on posts' location information. We employ two methods to extract geographic locations: (1) if a post explicitly discloses its location upon publication, we directly use this metadata; (2) if explicit location data is unavailable, we extract state-level information based on textual mentions of U.S. states within the post content. Posts for which neither method yields valid location information are excluded from this analysis. Figure 6 presents the percentage distribution of geolocated posts across U.S. states. Darker shades indicate a higher proportion of posts originating from each state. We observe that a substantial majority of geolocated posts are concentrated in Florida and North Carolina, which aligns with the primary trajectory of Hurricane Helene and Hurricane Milton captured in the dataset. This pattern highlights the spatial relevance of the social media posts collected in our dataset.

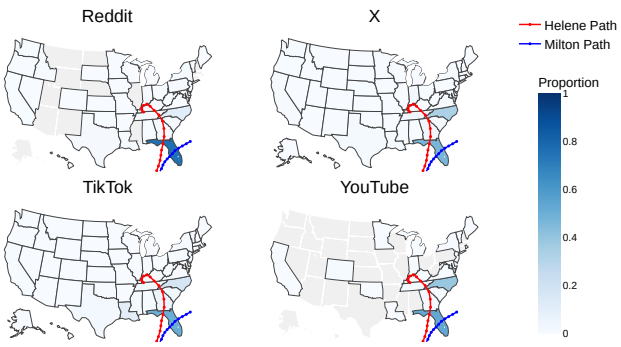



Figure 6: Spatial Distribution of Posts Across U.S. States

## Sentiment Analysis

We implement sentiment analysis on relevant posts from four social media platforms to examine the kinds of emotions expressed in the content. Following Zhang et al. (2024), we adopt their prompting strategy to guide the MLLM in analyzing the sentiment of each post in our dataset. Zhang et al. (2024) demonstrate that the MLLM achieves competitive performance in sentiment classification tasks, especially under few-shot or zero-shot settings, through a comprehensive test across 26 datasets. Therefore, we adopt the method because of the similarity of our task to the tested sentiment analysis task. Specifically, we let the

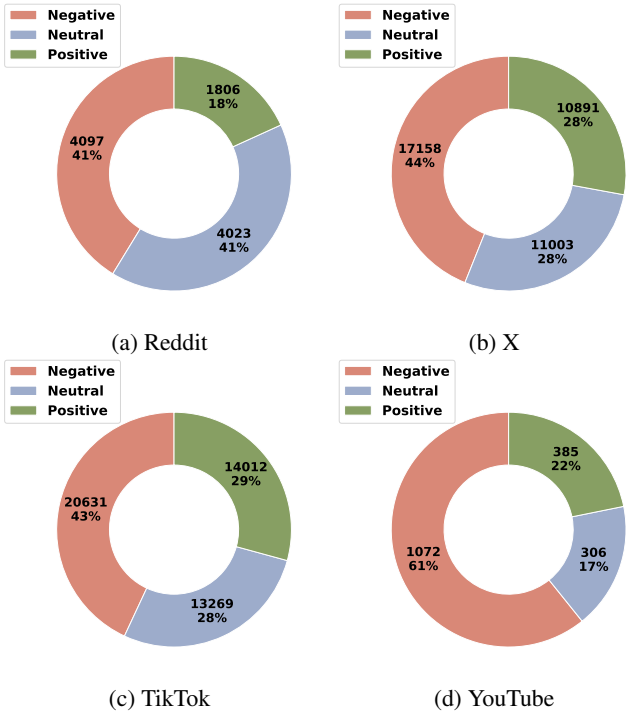



Figure 7: Sentiment Spectrum of Social Media Platform

MLLM consider the multimodal content of the post and classify the sentiment of the post into three categories: *positive*, *negative*, and *neutral*. Figure 7 illustrates the distribution of sentiments across all platforms. We observe that the majority of posts showed negative or neutral sentiment, with only a small portion reflecting positive sentiment. This distribution highlights the prevalent feelings of tension and pessimism among the public when facing disasters like hurricanes. On the other hand, posts on TikTok showed the highest proportion of positive emotions, indicating that although hurricane-related content tends to be negative, TikTok users are more likely to share uplifting or supportive data samples using the platform's visual and interactive features.

## Baseline Model Analysis

Finally, we evaluate the performance of four baseline models: RoBERTa (Liu et al. 2019), BART (Lewis et al. 2020), ConvBERT (Jiang et al. 2020), and ELECTRA (Clark et al. 2020) on the annotated labels. These models are selected due to their strong performance on a wide range of tasks, as well as their popularity and adoption as standard baselines in prior works (Naseer, Asvial, and Sari 2021; Mashkoor Siddiqui et al. 2025). Due to the substantial GPU resources and training time required for supervised multimodal models, we simplify the input to the textual content of social media posts in this evaluation. Similar to the work of Shang et al. (2024), we remove all URL links from the content. Additionally, to protect user privacy, all tagged usernames, starting with "@", are replaced with "@user". Subsequently, we combine the posts from all four platforms into a single dataset and divide them into 70% training, 15% validation, and 15% test-

| Category | RoBERTa | | BART | | ConvBERT | | ELECTRA | |
|---|---|---|---|---|---|---|---|---|
| | Accuracy | Macro F1 | Accuracy | Macro F1 | Accuracy | Macro F1 | Accuracy | Macro F1 |
| Casualty | 0.9306 | 0.7439 | 0.9258 | 0.7207 | 0.9303 | 0.7144 | 0.9315 | 0.7431 |
| Evacuation | 0.8009 | 0.6989 | 0.8116 | 0.7110 | 0.8032 | 0.6839 | 0.8040 | 0.6975 |
| Damage | 0.7732 | 0.7586 | 0.7664 | 0.7575 | 0.7732 | 0.7585 | 0.7710 | 0.7553 |
| Advice | 0.8344 | 0.6938 | 0.8382 | 0.7115 | 0.8365 | 0.6952 | 0.8326 | 0.6818 |
| Request | 0.8612 | 0.7345 | 0.8599 | 0.7441 | 0.8657 | 0.7463 | 0.8637 | 0.7474 |
| Assistance | 0.7884 | 0.7764 | 0.7888 | 0.7884 | 0.7871 | 0.7767 | 0.7874 | 0.7788 |
| Recovery | 0.8056 | 0.8035 | 0.7980 | 0.7924 | 0.8020 | 0.8001 | 0.8054 | 0.8026 |
| Linguistic Bias | 0.8395 | 0.7693 | 0.8314 | 0.7552 | 0.8423 | 0.7720 | 0.8417 | 0.7648 |
| Political Bias | 0.9195 | 0.8600 | 0.9216 | 0.8627 | 0.9232 | 0.8653 | 0.9208 | 0.8598 |
| Gender Bias | 0.9631 | 0.5712 | 0.9623 | 0.5776 | 0.9620 | 0.4903 | 0.9620 | 0.4903 |
| Hate Speech | 0.9769 | 0.6898 | 0.9783 | 0.6590 | 0.9785 | 0.6764 | 0.9746 | 0.4936 |
| Racial Bias | 0.9766 | 0.6947 | 0.9767 | 0.6826 | 0.9780 | 0.7085 | 0.9782 | 0.7014 |
| True/False Information | 0.8608 | 0.6934 | 0.8628 | 0.6801 | 0.8593 | 0.7156 | 0.8568 | 0.7201 |

Table 7: Accuracy and Macro F1 for Different Baseline Models

ing sets. We train every annotated category using NVIDIA L40S GPU and the evaluation results are presented in Table 7. We employ two metrics for evaluation: accuracy and macro F1 score, one focuses on the overall performance and the other focuses on the performance of the minority group. From Tables 3 to 5, we observe that most of the binary classification tasks exhibit data imbalance, with one class accounting for less than 40% of the labeled samples. To address the class imbalance problem, we use macro F1 score as one of the evaluation metrics for the annotations because it can focus on underrepresented classes.

For most categories, the four baseline models achieve test accuracy over 0.8 and macro F1 score over 0.75, showing strong model performance. However, for some categories, such as gender bias, hate speech, and racial bias, the macro F1 scores are significantly lower. This is mainly due to the small proportion of these categories in the dataset, which is only 2-3% of the total posts. The small proportion of these categories in the dataset limited the model's ability to receive sufficient training on these specific labels, thus affecting the macro F1 performance. These findings highlight the challenges of training bias detection models in imbalanced datasets.

## Definition and MLLM's Prompt for Each Task

Table 8 lists the definition for each category. In addition, Figures 8 to 11 illustrate the prompt for data cleaning, humanitarian class annotation, bias class annotation, information integrity annotation, and MLLM's second round annotation.

## Examples of MLLM's Uncertain Annotation

Table 9 lists the examples of MLLM's uncertain posts across different categories, with three examples provided for each. We also provide the label and reason through human annotation.

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,
"Reason": "your detailed explanation based on the post, the examples, and the comparative reasoning between the two viewpoints."
}

Figure 8: Prompt for Data Cleaning.

Prompt

You are reviewing a social media post related to a hurricane disaster. The post may include text, images, and videos. Carefully examine all available content together.

{Category Definition here}

Respond in strict JSON format:
{"Judgment": True or False,
"Reason": "Your detailed explanation based on the content of the post." }

Figure 9: Prompt for Humanitarian and Bias Class Annotation.

Prompt

You are reviewing a social media post related to a hurricane disaster. The post may include text, images, videos, and external links. Carefully examine all available content together. Your task is to determine whether the post contains false claims, describes truthful information, or is hard to identify due to a lack of verifiable evidence. To make your judgment, search online for reliable sources (e.g., government agencies, major news outlets, or verified fact-checking websites). Use these sources to verify any factual claims made in the post.

Label the post as one of the following:
- Flase Claims (-1): The post contains or promotes false or misleading information, even if only part of the content is inaccurate.
- True Claims (1): The post is factually accurate and supported by reliable sources.
- Hard to Identify (0): The post's truthfulness is difficult to determine (e.g., it describes personal experiences or no trustworthy sources can be found to verify it).

Please include the source(s) used to support your reasoning. If available, provide:
- A brief description of what the source says and how it supports your judgment.
- A URL link to the source (e.g., news article, official website, fact-checking report).

Respond in strict JSON format:
{ "Judgment": one of the following integer values: 1 = True Claims, 0 = Hard to Identify, -1 = False Claims
"Reason": Your explanation, including supported sources and their descriptions + url links in string format }

Figure 10: Prompt for Information Integrity Annotation.

| Category | Definition |
|---|---|
| **Casualty** | The post reports people or animals who are killed, injured, or missing during the hurricane.<br>For example:<br>- Deaths resulting from the hurricane<br>- Individuals or groups reported as missing or unaccounted for<br>- People or animals who are injured due to storm impact<br>- Reports of body counts, mass casualties, or disaster-related mortality statistics<br>- Posts seeking help to locate missing persons or pets |
| **Evacuation** | The post describes the evacuation, relocation, rescue, or displacement of individuals or animals due to the hurricane.<br>For example:<br>- People leaving their homes or being urged to evacuate from at-risk areas to safer locations (e.g., temporary shelters, higher ground, or other towns)<br>- Displacement of communities due to destruction of homes, rising water, or uninhabitable conditions<br>- Rescue operations carried out by emergency personnel, neighbors, or volunteers<br>- Reports of people or animals being trapped and awaiting rescue<br>- Use of boats, helicopters, or emergency vehicles to evacuate or rescue individuals |
| **Damage** | The post reports damage to infrastructure or public utilities caused by the hurricane.<br>For example:<br>- Destruction or severe damage to buildings, homes, or roads<br>- Power outages, downed power lines, or transformers exploding<br>- Cell towers, internet lines, or other communication infrastructure being down<br>- Disruption of water supply, sewage systems, or gas pipelines<br>- Damage to vehicles, boats, or public transit systems caused by wind or flooding |
| **Advice** | The post provides advice, guidance, or suggestions related to hurricanes, including how to stay safe, protect property, or prepare for the disaster.<br>For example:<br>- Safety tips for individuals, families, or pets during a hurricane (e.g., "stay indoors," "avoid floodwaters")<br>- Instructions on how to prepare emergency kits, store food/water, or stock up on supplies<br>- Advice on how to evacuate safely, choose a shelter, or avoid traffic routes<br>- Tips on securing property, such as boarding up windows, moving valuables to higher ground, or turning off gas/electricity<br>- Checklists or infographics about hurricane preparedness |
| **Request** | The post contains request for help, support, or resources due to the hurricane.<br>For example:<br>- People asking for rescue or evacuation assistance<br>- Requests for food, water, medical supplies, or basic necessities<br>- Posts seeking temporary shelter, housing, or safe relocation options<br>- Calls for volunteers, donations, or supplies<br>- Requests for transportation (e.g., fuel, vehicles, boat assistance) |
| **Assistance** | The post contains assistance and support to victims. This includes both physical aid and emotional or psychological support provided by individuals, communities, or organizations.<br>Physical and Material Assistance:<br>- Distribution of relief supplies (e.g., food, water, clothing, medical kits)<br>- Volunteers helping with debris removal, house repairs, or deliveries<br>- Government or NGO programs providing direct support (e.g., FEMA aid, Red Cross response)<br>- Organized donation drives, supply drop-offs, or logistical coordination<br>Emotional or Psychological Support:<br>- Expressions of solidarity, empathy, or moral support (e.g., "Praying for Florida," "Stay strong New Orleans")<br>- Messages offering comfort, hope, or encouragement to victims and survivors<br>- Posts acknowledging the suffering of affected communities and standing in support<br>- Public figures or organizations expressing concern or visiting victims |
| **Recovery** | The post describes efforts or activities related to the recovery and rebuilding process after the hurricane.<br>For example:<br>- Rebuilding homes, businesses, infrastructure, or public spaces after the storm<br>- Updates on clearing debris, restoring electricity, or repairing roads<br>- Reopening of schools, government services, stores, or public transport<br>- Support for mental health or trauma recovery following the disaster<br>- Resumption of economic activity (e.g., businesses reopening, job recovery) |

Table 8: Definition for each Category

| Category | Definition |
|---|---|
| **Linguistic Bias** | The post contains biased, inappropriate, or offensive language, with a focus on word choice, tone, or expression.<br>For example:<br>- Use profanity, vulgarities, or swear words (e.g., "f**k," "sh*t") in a way that is excessive, aggressive, or unnecessary<br>- Show emotionally charged language meant to belittle or ridicule victims, responders, or institutions<br>- Contain exaggerated, alarmist, or inflammatory wording (e.g., "this hellstorm wiped out everything!" when not supported by evidence)<br>- Use slang or informal language that may be perceived as inappropriate or disrespectful in the context of a disaster<br>- Feature mocking tones, insults, or dismissive phrasing (e.g., "Only idiots would stay behind") |
| **Political Bias** | The post expresses political ideology, showing favor or disapproval toward specific political actors, parties, or policies.<br>For example:<br>- Promote or criticize political leaders (e.g., mayors, governors, presidents) in relation to the hurricane response or preparedness<br>- Express support for or opposition to political parties or ideologies<br>- Frame the hurricane or its response through a politicized lens<br>- Use political slogans to express a viewpoint<br>- Criticize or endorse public policy, such as emergency spending, climate change legislation, infrastructure bills, etc. |
| **Gender Bias** | The post contains biased, stereotypical, or discriminatory language or viewpoints related to gender.<br>For example:<br>- Contain sexist language, slurs, or dismissive terms toward certain genders<br>- Reinforce traditional or harmful gender stereotypes<br>- Make generalizations about a specific gender<br>- Imply that certain responsibilities or behaviors are appropriate only for a particular gender<br>- Question or ridicule someone's capabilities or roles based on their gender |
| **Hate Speech** | The post contains language that expresses hatred, hostility, or dehumanization toward a specific group or individual, especially those belonging to minority or marginalized communities.<br>For example:<br>- Racial or ethnic minorities<br>- Immigrants or refugees<br>- Religious minorities<br>- People with disabilities<br>- Indigenous populations |
| **Racial Bias** | The post contains biased, discriminatory, or stereotypical statements directed toward one or more racial or ethnic groups.<br>Examples of racial groups commonly affected include: - Black or African American communities<br>- Latinx/Hispanic populations<br>- Asian or Pacific Islander groups<br>- Indigenous communities<br>- Arab, Middle Eastern, or South Asian populations |

Table 8: (Continued) Defination for each Category

Prompt

You are reviewing a social media post related to a hurricane disaster. The post may include text, images, and videos. Carefully examine all available content together.

{Category Definition here}

For this specific post, here are two sets of arguments:
- Arguments in favor (why the post may be related to a real hurricane disaster): {true reasons here}
- Arguments against (why the post may NOT be related to a real hurricane disaster): {false reasons here}

Carefully compare and contrast both sides of the arguments above, and critically assess which side presents a more credible and reasonable explanation based on the content of the post.
Consider the relevance, specificity, and plausibility of each argument in the context of the post, and use that analysis to form your final judgment.

Respond in strict JSON format:
{"Judgment": True or False,
"Reason": "Your detailed explanation based on the content of the post." }

Figure 11: Prompt for Second Round Annotation.

| Category | Post | Our Judgment and Reason |
|---|---|---|
| **Casualty** | *"The forecast storm surge with #Helene is something to be taken very seriously, and mandatory evacuations are already in effect for many Florida counties. Follow the guidance from local authorities. It is literally a matter of life and death."* | False. The post warns about the life-threatening potential of the storm surge but does not contain any reports of casualties (people or animals who are killed, injured, or missing). |
| | *"A helicopter appeared to deliberately damage supplies for victims of Hurricane Helene near Burnsville, NC last night. I hope they find the person(s) responsible and they lose their license to fly. They could've severely injured people. They should also be required to pay for damages of goods and property."* | False. While it mentions the potential for people to be severely injured, it does not report any actual injuries, deaths, or missing persons. |
| | *"We discovered this by doing our own recon work with a drone I just purchased to be used by my source who is a multi-tour infantry SOCOM veteran the sake of getting footage of the storm disaster and locating bodies."* | False. The post discusses search efforts and locating bodies, but it does not explicitly report any casualties (deaths, injuries, or missing persons/animals) resulting from the hurricane. |
| **Evacuation** | *"Neighbors: Cat 5 #Milton is now just 100s of miles away, pushing huge surge toward our coast, promising to be the most devastating storm here in 100+years. The exact landfall location will be elusive, wobbling up to landfall. Life is precious and fleeting. Be smart. Stay safe."* | False. The post is about a hurricane approaching the coast and urging people to stay safe, but it does not describe any specific evacuation or rescue efforts. |
| | *"Black Hawk helicopter crew has been grounded after officials say their flight thrashed a Hurricane Helene relief station."* | True. The post mentions a helicopter crew involved in Hurricane Helene relief efforts, implying evacuation and rescue operations. |
| | *"My thoughts are with all the people and animals going thru this stupid #HurricaneMilton. Stay safe and please take animals with you!!!!"* | True. The post mentions to 'take animals with you' implying evacuation is occurring. |
| **Damage** | *"Anderson Cooper struck by debris while reporting live on Hurricane Milton for CNN! Cooper okay, continues broadcasting. Brave journalism amidst extreme conditions."* | True. The video depicts a CNN reporter being struck by debris while covering the storm live, implicitly suggesting that the storm caused significant infrastructure damage. |
| | *"Just in video: The water has been sucked out of Tampa Bay by Hurricane Milton."* | False. The post describes water being sucked out of Tampa Bay due to the hurricane. This is a temporary effect of the hurricane's wind and pressure, not direct damage to infrastructure or utilities. |
| | *"Georgia shows appreciation to Alabama Power crews during Hurricane Helene restoration."* | True. The post discusses power crews restoring power after Hurricane Helene, which implies that there were power outages due to the hurricane. |
| **Advice** | *"This is a great write-up by North Carolina's debunking misinformation around Hurricane Helene. Highlights include: 1. Nobody can control the weather. 2. FEMA is not blocking donations. 3. Victims aren't receiving only $750 in assistance."* | True. The post provides guidance by debunking misinformation related to Hurricane Helene. While not direct advice, it clarifies facts and combats false information. |
| | *"I am seriously praying that this time, people listen and get out. The west coast already has lots of empty houses and piles of debris on the streets, reeling from the gut punch of Helene. "* | True. The post advises people to evacuate, which falls under guidance related to hurricane safety. |
| | *"Just remember, folks, Florida is a fafo state. Don't loot during Hurricane Milton. My governor isn't weak and isn't playing."* | False. This post is a warning not to loot in Florida during Hurricane Milton. It is not advice, guidance, or suggestions related to hurricanes. |

Table 9: Examples of MLLM's Uncertain Posts in each Category.

| Category | Post | Our Judgment and Reason |
|---|---|---|
| **Request** | *"Meanwhile, families were left waiting in desperate conditions, struggling to survive. In the midst of this, good Samaritans who put their lives on hold to help others were arrested, told to stand down, and had their supplies confiscated."* | True. The post describes families waiting in desperate conditions, struggling to survive after the hurricane, which indicates a need for help and support. |
| | *"NEWS: Starlink is providing a lifeline in East Tennessee, Western North Carolina, and all along the devastation path of Hurricane Helene."* | False. The post discusses Starlink providing cell service after Hurricane Helene. There are no direct requests for help, support, or resources in the post. |
| | *"Hurricane Helene devastates Bat Cave, NC; residents self-reliant due to FEMA's inability to navigate a road block."* | True. The post implies that residents of NC need help and support due to the devastation caused by Hurricane Helene and FEMA's inability to provide assistance. |
| **Assistance** | *"I feel so blessed that the weather got it together for this con. Devastating Cat 4 Hurricane Helene was last weekend and damaged the area and there is another predicted Cat 2 Hurricane they just named Milton projected to hit the Tampa area again on Tuesday. Everyone stay safe!"* | False. The post expresses concern about the hurricanes but does not offer any direct assistance or support to victims. It primarily focuses on the impact on an event and general safety advice." |
| | *"If you're in an evacuation zone, evacuate now, take your pets with you, and go to the supervisor of elections (SOE) office in your county and request your early ballot on the spot and vote now. You can request it and fill it out and turn it in now on the spot in person."* | False. The post primarily focuses on advising people to evacuate and vote early due to the impending hurricane. It does not contain elements of assistance and support to victims, either physical or emotional. |
| | *"#BREAKING: 'Open the freeway shoulders for MASSIVE hurricane evacuations' Governor Ron DeSantis is opening Florida highway shoulders for massive Hurricane Milton evacuations. 'No time like the present to put your evacuation plan into effect right now.' "* | True. The post describes a government action to facilitate evacuation, which constitutes direct support for people affected by the hurricane. Opening highway shoulders is a logistical measure to aid in a smoother evacuation process. |
| **Recovery** | *"Our amazing Red Cross volunteers are on the ground in GA & FL, supporting those hit by Hurricane #Helene. They've opened numerous shelters for thousands displaced. Help us continue this vital work."* | False. The post describes immediate relief efforts such as opening shelters, but does not mention any recovery or rebuilding activities. |
| | *"Wofford announces that the game at Western Carolina this Saturday will be played with no spectators in the wake of Hurricane Helene, which devastated the Carolinas last week."* | True. The post discusses the aftermath of Hurricane Helene and its impact on a football game, indicating recovery efforts are underway. The game is still being played, but with restrictions, suggesting return to normalcy. |
| | *"I wonder how much money foreign countries will send us to help with hurricane Helene relief."* | False. The post is wondering about foreign aid, not describing any actual recovery efforts. |
| **Linguistic Bias** | *"WATCH: Houses are now FLOATING AWAY in Steinhatchee, Florida as Hurricane Helene makes landfall Holy CRAP."* | True. The post uses the phrase Holy CRAP, which is considered inappropriate language, indicating linguistic bias. |
| | *"FORT MYERS, FLORIDA IS UNDER WATER This video released by the Lee County Sheriff is terrifying. Storm surge is DECIMATING Lee County, and Hurricane Helene hasn't even made landfall yet. PRAY FOR FLORIDA!"* | False. The post uses strong language such as DECIMATING and terrifying, but it reflects the severity of the situation. It is not inappropriate or offensive. |
| | *"Hurricane Helene DESTROYS Campaign After BOMBSHELL Announcement!"* | False. The title uses potentially charged words like DESTROYS and BOMBSHELL but not inappropriate or offensive. |

Table 9: (Continued) Examples of MLLM's Uncertain Posts in each Category.

| Category | Post | Our Judgment and Reason |
|---|---|---|
| **Political Bias** | *"BREAKING: Governor Declares State Of Emergency In 35 Counties Ahead Of Tropical Storm Milton, Which Is Forecast To Make Landfall On The West Coast Of Florida Tuesday As A MAJOR HURRICANE!"* | False. The post is reporting on a governor declaring a state of emergency in response to a hurricane. It does not express any political opinions or promote any political actors, parties, or policies. |
| | *"Senator offers disaster recovery resources for those impacted by Hurricane Helene."* | False. The post simply states that Sen. Blackburn is offering disaster recovery resources. It does not express any political opinions or biases. |
| | *"These Floridians Couldn't Flee Hurricane Milton. They're Incarcerated."* | True. The post implicitly criticizes the policy of incarcerating individuals who couldn't flee Hurricane Milton, framing it as a political issue related to incarceration and disaster response. |
| **Gender Bias** | *"It's been flooded before, but this is completely insane. It started at Wednesday, the pouring down rain. It rained Wednesday, Thursday. And she came Friday. So you can imagine all that rain plus that wind. I mean, everything she could move, she moved. That's some powerful water."* | True. The post refers to the flood as she and says everything she could move, she moved. This is an example of gendering an inanimate object in a way that is commonly associated with women. |
| | *"Where are all the man made hurricane counter experts?"* | False. The reference to man made is a conspiracy theory about the hurricane's origin, not a comment on gender. |
| | *"Yeah, we can control the weather. What are you gonna do about it? Your pretty boy muskrat can't even make a self-driving car properly, but we've got self-guided hurricanes. Just set em and forget em as they wipe out the opposition."* | True. The post uses the term 'pretty boy muskrat' to describe a person, which can be seen as gender-biased language due to its implication that someone is not strong or capable because of their appearance. |
| **Hate Speech** | *"Sorry, victims of Hurricane Helene. You're on your own. Communities of color need more. "* | True. The post dismisses the needs of hurricane victims, implies neglect and expresses hatred towards this specific group, and shows preference to minority communities. |
| | *"Until it looks like Hurricane Helene never occurred I don't want to see another MF cent go to a foreign country!"* | False. The post expresses frustration with foreign aid spending in the wake of a hurricane, but it does not contain any hate speech or target any specific group or individual with hatred, hostility, or dehumanization. |
| | *"Look at this worthless POS refusing to take questions about the disastrous Hurricane Helene. She's no leader."* | True. The post uses dehumanizing language such as worthless POS directed at a specific individual, which constitutes hate speech. |
| **Racial Bias** | *"Mexican immigrant families plagued by grief, questions after plant workers swept away by Helene."* | True. The post singles out Mexican immigrant families as being affected by the hurricane, which introduces a racial element to the reporting of a natural disaster. This could perpetuate stereotypes or create a perception of certain groups being disproportionately affected based on their ethnicity rather than the specific circumstances of the event (e.g., location of residence or workplace). |
| | *"The ultimate fact check of the lies this administration is peddling about their response to hurricane Helene. Stop underestimating the American people."* | False. Although the phrase mentions the American people, it refers broadly to the general public and expresses political criticism rather than targeting or excluding any racial or ethnic group. |
| | *"Not even a hurricane can destroy these fckers."* (With an image of cockroaches floating in flood water) | False. The post shows cockroaches floating in flood water after a hurricane. The post is vulgar but do not target any specific racial or ethnic group. |

Table 9: (Continued) Examples of MLLM's Uncertain Posts in each Category.