

DTCD-AFC: Disaster-Type Classification Dataset designed for Automated Fact-Checking

Yasuhisa Okumura, Kenji Oki, Shinya Kitajima

Fujitsu Limited, Japan
{o.yasuhisa, oki.kenji, kitajima.shinya}@fujitsu.com

Abstract

Natural disasters often trigger a surge of social media posts, unfortunately including misinformation that can disrupt response efforts, necessitating rapid fact-checking. To facilitate fact-checking, accurately identifying the specific disaster type mentioned in a social media post is crucial for effective evidence collection. However, a comprehensive multimodal social media dataset for disaster-type identification in fact-checking has been lacking. We address this gap by proposing the task of disaster-type classification from multimodal social media posts and introducing DTCD-AFC (Disaster-Type Classification Dataset designed for Automated Fact-Checking). This dataset, derived from CrisisMMD, features annotations for seven disaster types based on post claims, enabling evaluation of classification performance in automated fact-checking systems and supporting social media content analysis during disasters. Using GPT-4o as a zero-shot baseline, we found that evaluations grounded in DTCD-AFC labels better reflect post content than evaluations using event-derived CrisisMMD labels. Furthermore, we demonstrated that DTCD-AFC encompasses challenging tasks, such as identifying closely related disaster types (e.g., floods, hurricanes, and landslides), as well as tasks in which keywords within text exert significant influence.

1 Introduction

The popularization of the internet and social media has made sharing information easier, but it has also made it challenging to ensure the credibility and quality of that information. Particularly on social media, information spreads rapidly, and disinformation can easily spread. This disinformation has the potential to cause various negative impacts on society, necessitating its detection and countermeasures.

To address this issue, newspaper companies and online media conduct fact-checking. Fact-checking involves investigating whether specific information is based on facts and publishing the results as articles to provide accurate information to counter disinformation. While fact-checking plays a crucial role in enhancing information reliability, it faces challenges, such as the vast number of potential targets for investigation across the internet and the time-consuming, demanding nature of the work.

As a result, research to automate fact-checking has advanced in recent years, and our research group has also proposed an automated fact-checking method (Oki, Yamashita, and Kitajima 2025). Guo et al. suggest automating fact-checking by dividing the process into four steps: claim detection, evidence retrieval, verdict prediction, and justification production, and automating each step (Guo, Schlichtkrull, and Vlachos 2022). In automated fact-checking, web searches are primarily used to gather evidence for verdict prediction. However, web searches often include not only information from public institutions but also from individuals and machine-generated content, making it challenging to collect only credible information.

Therefore, we aim to collect credible evidence by identifying real-world events referenced in social media posts and obtaining information about those events from public institutions. For example, when fact-checking a social media post suspected of mentioning flooding, collecting information such as heavy rain alerts or river water level data issued by the National Weather Service¹ can provide evidence to assess the post’s veracity. Frequently referenced real-world events include natural disasters, political and economic issues, and social problems; however, this paper focuses specifically on natural disasters. During natural disasters, disinformation such as fake rescue requests or fabricated damage reports can spread, so fact-checking is crucial to prevent confusion in disaster response efforts. Here, weather information from government agencies and meteorological organizations is credible evidence for automated fact-checking related to natural disasters.

To gather information from public agencies, we must identify the type of disaster the social media post relates to. Therefore, we examine the disaster-type classification task, which determines the type of disaster.

In related work on disaster-type classification, Shetty et al. perform disaster-type classification from multimodal social media posts about disasters (Shetty et al. 2024). They use the CrisisMMD² (Alam, Ofli, and Imran 2018; Ofli, Alam, and Imran 2020) dataset, which contains multimodal X³ posts related to disasters for evaluation. However, Crisis-

¹<https://www.weather.gov/>

²<https://huggingface.co/datasets/QCRI/CrisisMMD>

³<https://x.com>

MMD primarily focuses on determining whether social media posts contain valuable information for humanitarian aid during disaster events. Therefore, disaster-type labels are not explicitly assigned; only labels based on the disaster events for which the posts were collected are available. For disaster classification in automated fact-checking systems, accurate identification of the specific disaster type from a social media post’s content is required. For example, even if a post concerns a hurricane, if its content expresses a political opinion, the evidence usable for fact-checking this post is likely to be political information rather than meteorological information about the hurricane. Consequently, CrisisMMD is insufficient as a dataset for disaster-type classification, and no suitable dataset exists specifically for determining disaster types from multimodal social media posts.

Therefore, we created a Disaster-Type Classification Dataset designed for Automated Fact-Checking (DTCD-AFC)⁴ based on CrisisMMD. In DTCD-AFC, we annotated each CrisisMMD entry with one of seven types based on the social media post’s claim. The types are: earthquake, flood, typhoon, fire, landslide, other disaster, or not a disaster-related post. DTCD-AFC helps to evaluate disaster-type classification performance in automated fact-checking evidence retrieval. Moreover, disaster-type classification performance is essential for analyzing social media posts during disasters and gathering supporting information in disaster response, and DTCD-AFC is also helpful for these applications.

2 Related Work

In this section, we describe datasets related to disaster-type classification, followed by a discussion of related work on disaster-type classification. Then, we explain related work on utilizing disaster-type classification in automated fact-checking.

2.1 Datasets related to Disaster-Type Classification

Although no dataset is specifically suited for disaster-type classification research targeting both text and images in social media posts, several datasets related to disasters are available. In this section, we describe CrisisMMD (Alam, Offi, and Imran 2018; Offi, Alam, and Imran 2020), created for the task of determining whether social media posts during disasters are beneficial for humanitarian aid, MEDIC⁵ (Alam et al. 2023), which provides disaster-type labels for images, and DisasterM3⁶ (Wang et al. 2025), composed of satellite images before and after disasters paired with Question-Answering (QA) data.

CrisisMMD is a dataset of text and images collected from X posts related to disasters, primarily targeting the task of determining whether posts contain valuable information for

humanitarian aid. It comprises 16,058 texts and 18,082 images collected from past natural disasters (Hurricane Irma, Hurricane Harvey, Hurricane Maria, California wildfires, Mexico earthquake, Iraq-Iran earthquake, Sri Lanka floods) by specifying keywords and time periods. The discrepancy between the number of text entries and the number of images is because a single post may contain multiple images.

Table 1 shows the keywords and collection periods for each disaster. In CrisisMMD, collected posts are labeled for both text and images, indicating whether they are helpful for humanitarian aid, the type of damage, and its severity. If we use CrisisMMD for disaster-type classification, one approach is to treat the disaster event targeted for collection as the correct label. For example, all posts collected for “Hurricane Irma” would be labeled “hurricane.” However, a problem arises because the content of these posts is not always directly related to the specific disaster event, given that they are collected solely based on keywords and time periods.

MEDIC is a dataset that addresses the shortage of disaster images for machine learning by annotating images in disaster datasets (Alam et al. 2021; Mouzannar, Rizk, and Awad 2018; Nguyen et al. 2017), including CrisisMMD, with labels for disaster-type classification. The disaster-type labels consist of seven categories: earthquake, fire, flood, hurricane, landslide, non-disaster, and other disasters. While MEDIC can be used for disaster-type classification, it has the limitation that labels are annotated solely from images, making it unsuitable for multimodal disaster-type classification that uses both text and images.

DisasterM3 is a dataset composed of pre- and post-disaster satellite images paired with QA pairs. It aims to enhance the performance of vision-language models for disaster damage assessment and response by leveraging remote sensing. Satellite images were collected for 36 past natural and man-made disaster events. It categorized disaster types into 10 types: fire, volcano, tornado, hurricane, tsunami, flood, explosion, landslide, earthquake, and conflict. The QA pairs consist of 9 questions covering 5 capabilities: recognition, counting, localization, reasoning, and report generation. Here, the recognition category includes questions for disaster-type classification. While DisasterM3 can be used for disaster-type classification, its focus on satellite imagery means it cannot be applied to general disaster images found in social media posts.

2.2 Disaster-Type Classification

In this section, we introduce several methods for disaster-type classification. While the proposed DTCD-AFC cannot be directly used to evaluate these methods, it demonstrates the importance of research focused on disaster-type classification.

Huang et al. proposed an emergency detection framework called SBEED (Huang et al. 2021). In SBEED, the classification phase uses machine learning models to determine whether social media posts relate to emergencies and categorizes them into four types: natural disasters, accidents, public health incidents, and public safety incidents. After that, the extraction phase extracts the time and the location,

⁴The dataset is available at <https://huggingface.co/datasets/offi/Disaster-Type-Classification-Dataset-for-Automated-Fact-Checking>

⁵<https://huggingface.co/datasets/QCRI/MEDIC>

⁶<https://github.com/Junjue-Wang/DisasterM3>

Crisis name	Keywords	Data collection period
Hurricane Irma	Hurricane Irma, Irma storm, Storm Irma, Irma Hurricane, Irma	Sep 6–21, 2017
Hurricane Harvey	Hurricane Harvey, Harvey, HurricaneHarvey, Tornado	Aug 25–Sep 20, 2017
Hurricane Maria	Hurricane Maria, Maria Storm, Maria Cyclone, Maria Tornado, Tropical Storm Maria, HurricaneMaria, puerto rico	Sep 20–Nov 13, 2017
California wildfires	California fire, California wildfire, Wildfire California, USA Wildfire, California wildfires	Oct 10–27, 2017
Mexico earthquake	mexico earthquake, mexicoearthquake	Sep 20–Oct 6, 2017
Iraq-Iran earthquake	kuwait earthquake, iran earthquake, halabja earthquake, Iraq earthquake	Nov 13–19, 2017
Sri Lanka floods	flood Sri Lanka, FloodSL, SriLanka flooding, SriLanka floods, SriLanka flood, typhoon mora, cyclone mora, mora, CycloneMora	May 31–Jul 3, 2017

Table 1: Keywords used for data collection and data collection periods in CrisisMMD.

and the clustering phase groups posts related to the same emergency.

Shetty et al. proposed a multimodal deep learning model (Shetty et al. 2024). Their model extracts features from both text and images of social media posts. By learning how to combine these features, they generate a model capable of more accurate disaster classification.

2.3 Integration of Disaster-Type Classification in Automated Fact-Checking

Automated fact-checking typically employs a framework that collects external information related to a claim and evaluates its veracity based on that evidence. Particularly in the context of disasters, evidence sources are highly diverse, ranging from official announcements to observational data; therefore, the methods used for information search and selection significantly influence verification accuracy.

Several methods have been proposed to incorporate contextual information into the process of searching for disaster information and collecting evidence. Priya et al. introduce Topic-Aligned Query Expansion to address the vocabulary variations and notation discrepancies unique to social media, enabling high-precision retrieval of damage-related posts (Priya et al. 2020). Additionally, multi-stage retrieval for web search during crisis (Tcaciuc, Rege Cambrin, and Garza 2025) performs explicit filtering based on topics, such as disaster types, in the initial stage of search, aiming to eliminate irrelevant information while balancing search efficiency and performance.

Furthermore, disaster type is also used as a guideline for selecting appropriate sources of evidence. Jayalakshmi et al. propose a system that dynamically selects and matches external data sources such as official APIs and satellite images based on disaster types determined from social media posts (Jayalakshmi, Kumar et al. 2025). This method extracts physical features such as flooding patterns and thermal anomalies according to disaster types, enabling more precise verification through cross-modal matching with textual information.

In addition, ontology-based research, such as the TERMINUS ontology (De Nicola and Villani 2025), constructs knowledge bases that semantically expand related concepts centered on disaster types. Such a knowledge base is positioned as a foundation supporting vocabulary-independent

information search and the integration of heterogeneous information sources.

These trends indicate that contextual information plays an important role in evidence collection for disaster scenarios. Disaster-type classification is a key component of this context; thus, the fine-grained dataset constructed in this study serves as a fundamental dataset for building classification models, ultimately facilitating the integration of disaster-specific context into automated fact-checking.

3 DTCD-AFC: Proposed Dataset

This section defines the disaster-type classification task and describes the details of dataset annotation.

3.1 Disaster-Type Classification Task

The disaster-type classification task involves determining which type of disaster a social media post is related to. The criterion for judgment is whether the main content of the social media post is related to the disaster’s impact. For example, a post conveying the damage situation of a hurricane is classified as `hurricane`, while a post using a hurricane to express political opinions is classified as `not_disaster`. Additionally, posts related to multiple disaster types are classified into the primary disaster impact expressed in the post. We adopted the same seven disaster categories for annotation as used in MEDIC: `earthquake`, `fire`, `flood`, `hurricane`, `landslide`, `other_disaster`, and `not_disaster`.

3.2 Annotation

In this study, we annotated a dataset for the disaster-type classification task. The annotation targets are 18,082 text-image pairs from the CrisisMMD dataset. When a single post contained multiple images, we created pairs by matching the same text to each image and added annotations to each one. We assigned six annotators with general knowledge and divided the data into six blocks, each assigned to an annotator, who annotated their assigned block. Annotators reviewed both the text and images of social media posts to determine which disaster type the post’s content was related to and selected the appropriate category from the seven options. Subsequently, each block was assigned to a second annotator who had not previously annotated it. As a result, each data point was annotated by two annotators. For data

Type	Number of annotations
earthquake	1,717
fire	1,477
flood	1,643
hurricane	8,192
landslide	37
other_disaster	332
not_disaster	4,684
total	18,082

Table 2: Number of annotations for each disaster type.

with annotation discrepancies, the annotators met to reach a consensus on the final annotation.

To evaluate the reliability of the annotation, we calculated Inter-annotator agreement using Cohen’s kappa κ (Landis and Koch 1977). There are six annotators, but each data point is annotated by two annotators, so we calculated κ assuming that the entire dataset was annotated by two annotators. The resulting κ score was 0.91, indicating almost perfect agreement according to the interpretation criteria for κ .

3.3 Dataset Structure

Table 2 shows the number of annotations for each category. Among the 18,082 annotated text-image pairs, those labeled `hurricane` were the most numerous. This is because, as shown in Table 1, three of the seven disasters collected in CrisisMMD were hurricane-related. Next most numerous were pairs labeled `not_disaster`. The reason for this is that CrisisMMD collects posts based solely on keywords and time periods, resulting in a large number of posts whose primary topic was not disaster-related. These include posts that do not contain information related to disasters or posts that contain disaster-related topics but are political opinions or jokes unrelated to the disaster’s impact.

In the following, we compare the distribution of annotations in our dataset with those of CrisisMMD and MEDIC explained in Section 2.

First, we compare the distribution of annotations in our dataset with those of CrisisMMD. Figure 1 shows the number of annotations for each category for disaster events targeted for collection in CrisisMMD. The primary disaster type for each collected event is frequently annotated, while posts related to Hurricane Harvey were also frequently annotated as `flood`. This is because there were many posts about hurricane-related flood damage. For Hurricane Harvey posts, many were also classified as `other_disaster` because tornadoes that occurred during Hurricane Harvey were classified as `other_disaster`.

In particular, many posts related to Hurricane Harvey, Hurricane Irma, Hurricane Maria, and the Sri Lanka floods are classified as `not_disaster`. The reason for this is that the names of hurricanes that caused various disasters are commonly used in other contexts, such as personal names. Consequently, many posts do not contain disaster-related information.

Next, we compare the distribution of annotations in our dataset with those of MEDIC. Figure 2 shows the distribu-

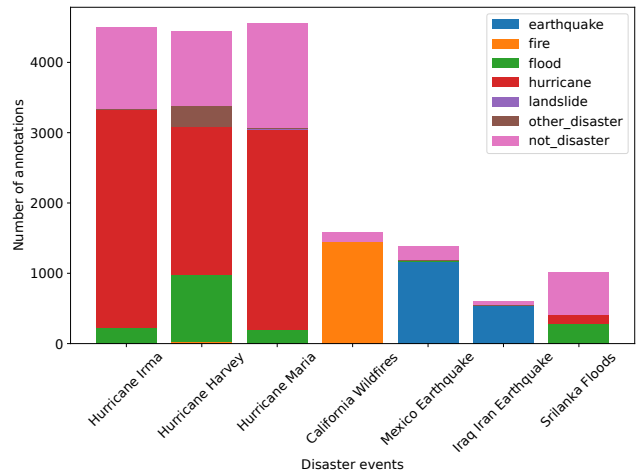


Figure 1: Distribution of annotation counts for each disaster event in CrisisMMD.

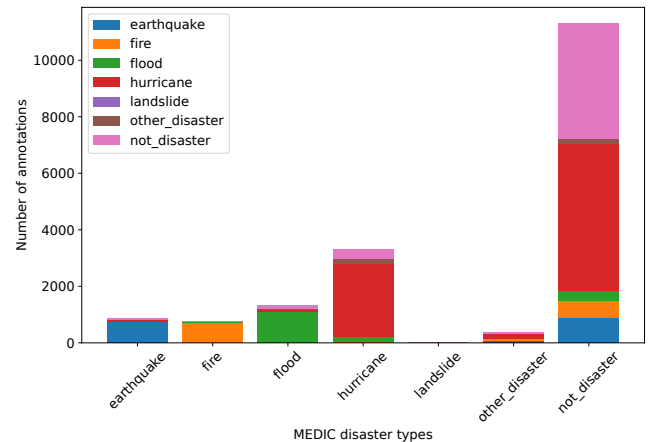


Figure 2: Distribution of annotation counts for each disaster type in MEDIC.

tion of annotation counts for each disaster type in MEDIC. MEDIC’s disaster-type labels have a very high number of `not_disaster` annotations. This is because many posts contain images unrelated to the content, or even if they are disaster-related photos, they cannot be judged as related to the disaster on their own, and MEDIC only annotates images, so all such posts are classified as `not_disaster`. Among these posts, some contain disaster-related content in the text, and our annotations classify them into disaster-related categories.

Figure 3 shows examples of posts annotated as `not_disaster` in MEDIC. Compared to MEDIC, our annotations can provide labels more suitable for multimodal disaster-type classification. Examples of annotated data are shown in Figure A.1.

Table 3 shows the confusion matrix of the annotations between the first annotator and the second annotator. The confusion matrix shows that most annotations are consistent, but there are some discrepancies between `hurricane` and

		The second annotator						
		earthquake	fire	flood	hurricane	landslide	other_disaster	not_disaster
The first annotator	earthquake	1,695	0	0	3	0	1	53
	fire	0	1,453	0	2	0	0	20
	flood	0	1	1,542	30	3	1	15
	hurricane	1	0	69	7,553	9	60	179
	landslide	0	0	0	5	25	0	0
	other_disaster	0	1	3	15	0	259	5
	not_disaster	28	22	29	582	0	11	4,407

Table 3: Confusion matrix of annotations. Rows represent annotations from the first annotator, and columns represent annotations from the second annotator.



Text: San Juan, Puerto Rico blackout post being hit by hurricane Maria (dailymail). Eerie2 gmn gitu ya... <https://t.co/4KSrZI2Gdu>

MEDIC label: not_disaster

DTCD-AFC label: hurricane

Figure 3: Examples of not_disaster annotations in MEDIC.

not_disaster. Many of these posts, while mentioning hurricanes, were not directly disaster-related; instead, they focused on political opinions or fundraising appeals. Annotators primarily evaluated these posts based on the usefulness of disaster information for fact-checking. These discrepancies likely arose because annotators’ judgments of usefulness varied.

4 Evaluation

In this section, we conduct experiments using an LLM for disaster-type classification to analyze the difficulty of the proposed dataset, DTCD-AFC, and to evaluate its effectiveness.

4.1 Evaluation Environment

We use an LLM, OpenAI’s GPT-4o⁷, to evaluate DTCD-AFC, since LLMs demonstrate high performance across various downstream tasks and can be readily adopted as a zero-shot baseline without additional training.

Prompt A.1 and Prompt A.2 show the disaster-type classification prompts used in the evaluations, and Figure 4 shows the flow of the prompts used for disaster-type classification.

⁷<https://platform.openai.com/docs/models/gpt-4o>

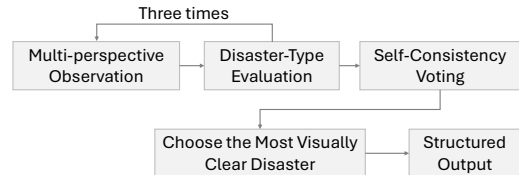


Figure 4: Flow of prompt processing for disaster-type classification.

The prompts consist of five steps: Multi-Perspective Observation, Disaster-Type Evaluation, Self-Consistency Voting, Choose the Most Visually Clear Disaster, and Structured Output.

First, in the Multi-Perspective Observation step, the disaster type relevant to both the post text and image is identified, and their consistency is assessed based on the LLM’s knowledge. Next, in the Disaster-Type Evaluation step, we determine whether the post content corresponds to each disaster type. Then, in the Self-Consistency Voting step, the processes of Multi-Perspective Observation and Disaster-Type Evaluation are repeated three times, and a majority vote determines the inferred disaster type. In the Choose the Most Visually Clear Disaster step, if the majority vote in Self-Consistency Voting is not unanimous, it checks the image content and selects the most visually clear disaster type. Finally, in the Structured Output step, the final disaster type is output in JSON format.

In the evaluation, we fed the LLM with social media posts collected in CrisisMMD and prompts, and compared the disaster types it output with the annotated ground-truth labels (*DTCD-AFC labels*). When a single post contained multiple images, we input each image as a pair with the corresponding post text. In addition, since CrisisMMD is a dataset collected by specifying seven crises, we also compared the results with those obtained by determining the disaster type based on the crisis name (*CrisisMMD labels*); posts related to Hurricane Irma, Hurricane Harvey, and Hurricane Maria are labeled *hurricane*, posts related to California wildfires are labeled *fire*, posts related to the Mexico earthquake and the Iraq-Iran earthquake are labeled *earthquake*, and posts related to Sri Lanka floods are labeled *flood*. Furthermore, we consider posts in CrisisMMD that have both the text and the image labeled as “Not informative for humanitarian aid” as *not_disaster* in the

CrisisMMD Labels	DTCD-AFC Labels
0.743	0.782

Table 4: Accuracy of disaster-type classification.

Disaster type	CrisisMMD Labels	DTCD-AFC Labels
earthquake	0.874	0.930
fire	0.893	0.954
flood	0.236	0.759
hurricane	0.821	0.826
landslide	N/A	0.357
other_disaster	N/A	0.336
not_disaster	0.540	0.550

Table 5: F1 score of disaster-type classification.

CrisisMMD labels because they lack relevant information about disaster damage.

MEDIC includes CrisisMMD; however, since it is annotated solely from images, it is considered unsuitable for comparison. Therefore, no comparison was performed in this evaluation.

4.2 Evaluation Results

Table 4 and Table 5 show the evaluation results (accuracy and F1 score, respectively) by using DTCD-AFC labels and CrisisMMD labels. In the evaluation, we were unable to obtain output for 98 text-image pairs due to errors caused by the LLM’s content filters and other factors. Therefore, these pairs were excluded from the experiment. The results show that accuracy with DTCD-AFC labels is higher than with CrisisMMD labels, and that F1 scores are also higher across all disaster types with DTCD-AFC labels than with CrisisMMD labels. These results indicate that using DTCD-AFC labels enables more accurate disaster-type classification.

The details of the results are as follows.

Disaster Types not Present in CrisisMMD We should note that for the CrisisMMD labels, the F1 score for landslides and other disasters is N/A because there are no posts with these labels as ground truth. In the results of CrisisMMD labels, the F1 score for flood is remarkably low. This is because many posts collected with a focus on hurricanes contain flood damage related to hurricanes, and these posts were classified as flood damage. The following example and Figure 5 show an example of a post of misclassification in CrisisMMD labels.

Example of misclassification in CrisisMMD labels.

Text: RT @stephentpaulsen: My street in SE #Houston is now a river. That light is from lightning; it’s 10pm #Harvey <https://t.co/cmlH5tained>
Image: Figure 5
CrisisMMD label: hurricane
DTCD-AFC label: flood
Disaster-type classification result: flood

This post was collected with a focus on Hurricane Harvey in CrisisMMD, so the CrisisMMD label is hurricane.



Figure 5: Image example of misclassification in CrisisMMD label.



Figure 6: Example of a misclassified post between hurricane and flood.

However, the post’s content strongly indicates flood damage, and thus it was classified as flood by the LLM. In DTCD-AFC label, this post is judged to be related to flood damage and is labeled as flood, which matches the LLM result.

Misclassification between Hurricane and Flood When checking the F1 scores for each disaster type using DTCD-AFC labels, we obtained high F1 scores of 0.9 or above for earthquakes and fires. In contrast, the F1 score remained low at 0.4 or below for landslides and other disasters, indicating a significant difference in scores across disaster types. Table 6 shows the confusion matrix for the results using DTCD-AFC labels. From the confusion matrix, we can see that there are many misclassifications between floods and hurricanes. This is because posts related to hurricane-caused flood damage often involve heavy rain, leading to many similar posts and misclassification. The following example and Figure 6 show an example of misclassification between hurricane and flood.

Example of misclassification of hurricane

Text: WATCH: Tropical Storm Harvey soaks Houston with heavy rain <https://t.co/DLKWIcV1wY> #US <https://t.co/g2NhZsh8RE>

Image: Figure 6

DTCD-AFC label: hurricane

Disaster-type classification result: flood

This post contains the text “Tropical Storm Harvey” and includes a weather image of a hurricane, so it is labeled as hurricane in the DTCD-AFC label. However, the LLM strongly captures the expression “soaks Houston with heavy rain” in the post text and classifies it as flood.

Regarding floods and hurricanes, they are highly related

		Predicted						
		earthquake	fire	flood	hurricane	landslide	other_disaster	not_disaster
True	earthquake	1,678	0	3	4	0	4	4
	fire	4	1,459	4	3	0	1	5
	flood	0	1	1,297	329	2	0	0
	hurricane	11	7	258	7,713	5	36	135
	landslide	0	0	4	22	10	0	0
	other_disaster	5	3	16	214	0	87	6
	not_disaster	217	113	206	2,236	3	59	1,820

Table 6: Confusion matrix of disaster-type classification (DTCD-AFC labels).



Figure 7: Image of a misclassification example of landslide.

in evidence collection for disaster fact-checking, so these misclassifications could be considered non-critical. However, depending on the specific requirements of downstream tasks, these misclassifications may become problematic; therefore, we treat them as a problem.

Misclassification between Hurricane and Landslide

Since hurricanes often trigger landslides and the post text contains hurricane-related words, misclassification is frequent. The following example and Figure 7 show an example of misclassification between hurricane and landslide.

Misclassification Example of landslide

Text: Damage from #HurricaneMaria in St. Thomas, #USVI. Photo credit: Conn Davis. <https://t.co/uLyCmCvzfg>

Image: Figure 7

DTCD-AFC label: landslide

Disaster-type classification result: hurricane

This post shows a landslide in the image, so we labeled it as `landslide` in DTCD-AFC. However, the LLM judged it to be related to a hurricane based on the text “Hurricane-Maria” and classifies it as `hurricane`.

Misclassification between Hurricane and Tornado

Many posts related to tornadoes were labeled as `other_disaster` in DTCD-AFC because the collection keywords for Hurricane Harvey in CrisisMMD included tornadoes. However, the LLM tends to classify tornadoes as `hurricane` labels, leading to frequent misclassifications. The following example and Figure 8 show



Figure 8: Image of a misclassification example for `other_disaster`.



Figure 9: Image of a misclassification example of `not_disaster`.

an example of misclassification between hurricane and tornado.

Misclassification example of `other_disaster`

Text: There was really a tornado crazy <https://t.co/yhryvidFIO>

Image: Figure 8

DTCD-AFC label: `other_disaster`

Disaster-type classification result: `hurricane`

We labeled this example as `other_disaster` because the text indicated it was related to a tornado. However, the LLM interprets a tornado as a type of hurricane and classifies it as `hurricane`. One possible cause is the lack of specific examples for `other_disaster` within the prompt.

Misclassification in Not Disaster

For non-disaster posts, many posts contained political content related to disasters or used collection keywords in a different context. These posts contained mentions of disasters, but their main content was unrelated to actual disaster events, which led to frequent misclassifications. The following example and Figure 9 show an example of misclassification between hurricane and `not_disaster`.

Misclassification Example of `not_disaster`

Text: Chelsea Clinton Blasts Trump For Helping Harvey’s Victims, Gets Eviscerated? Instantly <https://t.co/obmgJbFZjH> <https://t.co/hx3nvn7cxD>

Image: Figure 9

DTCD-AFC label: `not_disaster`

Disaster-type classification result: hurricane

We labeled this post as `not_disaster` in DTCD-AFC because the text is political and the image is unrelated to disasters. However, the LLM judges it to be related to Hurricane Harvey based on the text “Harvey” and classifies it as hurricane.

4.3 Discussion

Compared with the CrisisMMD dataset, the DTCD-AFC dataset achieved higher accuracy across most disaster types. This underscores the utility of DTCD-AFC, which was meticulously annotated to enable comprehensive disaster-type classification from both text and images, rather than relying on a single information source.

In addition, the evaluation results clarified the characteristics of DTCD-AFC. Details are as follows:

- In DTCD-AFC, there is significant skew in the number of posts across different disaster types.
- DTCD-AFC provides examples that demonstrate how over-reliance on text keywords can lead to misclassification.
- Floods, hurricanes, and landslides are closely related, so DTCD-AFC poses challenges that are prone to misclassification.

Furthermore, these results indicate that disaster-related keywords in posts’ text significantly influence LLM-based disaster-type classification. To improve classification accuracy on this task using DTCD-AFC, integrating comprehensive judgments that effectively leverage image information, in particular, suggests the potential for more precise disaster-type classification.

5 Limitations

DTCD-AFC is an effective dataset for disaster-type classification, but it has several limitations.

First, the X posts used in the dataset are limited to specific time periods and regions, potentially restricting the dataset’s diversity. The dataset collects data from seven disasters that occurred around the same time period and does not include information on disasters that occurred outside this timeframe. Furthermore, there is significant skew in the number of data points per annotation. For example, while hurricane events are annotated very frequently, the number of landslide annotations is low. Such data skew could potentially impact evaluation results.

Second, distinguishing between disaster types is difficult, and some posts may involve multiple disaster types. For example, if a flood occurs after a hurricane, the post could address both disasters. As a result, such posts could make accurate disaster-type classification evaluation particularly challenging.

6 Conclusion

In this paper, we proposed DTCD-AFC, a dataset for classifying disaster types from social media posts. DTCD-AFC includes social media posts collected for disasters and label information indicating which disaster type the post’s claim content corresponds to. We also conducted disaster-type classification evaluation using an LLM as a baseline for the dataset and presented the results. The evaluation results indicate that disaster-related keywords in posts’ text significantly influence LLM-based disaster-type classification. Moreover, floods, hurricanes, and landslides are closely related, so the dataset provides challenges that are prone to misclassification. This dataset will support research on identifying disasters mentioned in social media posts, leading to advancements such as automated disaster-related fact-checking. As future work, we aim to improve the dataset’s quality by increasing the number of data points in minority classes, as there is an imbalance in the number of data points per disaster type. Additionally, we plan to explore methods for accurately classifying similar disaster types in the disaster-type classification task.

7 Ethical Statement

The data included in DTCD-AFC contains links to X posts and may provide access to poster information. Therefore, when using this data, users must respect the poster’s privacy and avoid misusing the poster’s information.

While this dataset aims to promote research on automated fact-checking systems for disaster-related posts, users could also use it for malicious purposes. Using the dataset could enable the development of attack methods that misclassify disaster types, potentially adversely affecting disaster response efforts. To mitigate these risks, when developing disaster classification systems, it is necessary to utilize not only the text and images of posts but also other data sources, while also considering the true purpose of the posts.

The dataset aligns with the FAIR principles:

- **Findable:** The dataset is published at <https://huggingface.co/datasets/o-yas/Disaster-Type-Classification-Dataset-for-Automated-Fact-Checking>, with DOI.
- **Accessible:** The dataset is freely available for non-commercial use, with clear licensing information.
- **Interoperable:** The dataset is provided in standard formats (CSV and JPG images) to facilitate integration with various tools and platforms.
- **Reusable:** The dataset includes comprehensive documentation, data descriptions and CC-BY-NC-SA-4.0 license text to support reuse in future research.

Acknowledgment

This work is based on results obtained from a project, JPNP22007, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

- Alam, F.; Alam, T.; Hasan, M. A.; Hasnat, A.; Imran, M.; and Ofli, F. 2023. MEDIC: A Multi-Task Learning Dataset for Disaster Image Classification. *Neural Computing and Applications*, 35: 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.; Alam, T.; and Qazi, U. 2021. Deep learning benchmarks and datasets for social media image classification for disaster response. In *Proceedings of the 12th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '20, 151–158. IEEE Press. ISBN 9781728110561.
- De Nicola, A.; and Villani, M. L. 2025. Actionable Semantic Patterns in the Crisis Management Lifecycle: The TERMINUS Ontology. *Smart Cities*, 8(5).
- Guo, Z.; Schlichtkrull, M.; and Vlachos, A. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics*, 10: 178–206.
- Huang, L.; Liu, G.; Chen, T.; Yuan, H.; Shi, P.; and Miao, Y. 2021. Similarity-based emergency event detection in social media. *Journal of Safety Science and Resilience*, 2(1): 11–19.
- Jayalakshmi, R.; Kumar, S.; et al. 2025. Truth-Aware Disaster Response and Warning System using Linear SVC. *International Journal of Research Publication and Reviews*, 6(11): 1–8.
- Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1): 159–174.
- Mouzannar, H.; Rizk, Y.; and Awad, M. 2018. Damage Identification in Social Media Posts using Multimodal Deep Learning. In *ISCRAM*. Rochester, NY, USA.
- Nguyen, D. T.; Ofli, F.; Imran, M.; and Mitra, P. 2017. Damage assessment from social media imagery data during disasters. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, 569–576.
- Ofli, F.; Alam, F.; and Imran, M. 2020. Analysis of Social Media Data using Multimodal Deep Learning for Disaster Response. In *17th International Conference on Information Systems for Crisis Response and Management*. ISCRAM, ISCRAM.
- Oki, K.; Yamashita, K.; and Kitajima, S. 2025. A hybrid approach combining LLMs and web-based information for automated fact-checking. In *2025 IEEE International Conference on Web Services (ICWS)*, 960–962. Los Alamitos, CA, USA: IEEE Computer Society.
- Priya, S.; Bhanu, M.; Dandapat, S. K.; Ghosh, K.; and Chandra, J. 2020. TAQE: Tweet Retrieval-Based Infrastructure Damage Assessment During Disasters. *IEEE Transactions on Computational Social Systems*, 7(2): 389–403.
- Shetty, N. P.; Bijalwan, Y.; Chaudhari, P.; Shetty, J.; and Muniyal, B. 2024. Disaster assessment from social media using

multimodal deep learning. *Multimedia Tools and Applications*.

Tcaciuc, C. C.; Rege Cambrin, D.; and Garza, P. 2025. Multi Stage Retrieval for Web Search During Crisis. *Future Internet*, 17(6).

Wang, J.; Xuan, W.; Qi, H.; Liu, Z.; Liu, K.; Wu, Y.; Chen, H.; Song, J.; Xia, J.; Zheng, Z.; and Yokoya, N. 2025. DisasterM3: A Remote Sensing Vision-Language Dataset for Disaster Damage Assessment and Response.

A Appendix

A.1 Data Example

Figure A.1 shows examples of annotated data.

A.2 Prompt

Prompt A.1 and Prompt A.2 describe disaster-type classification prompts used in the experiments in Section 4.

earthquake



Mexico Rescinds Harvey Relief Offer To Deal With An? Earthquake <https://t.co/KxqWPGG4yf>
<https://t.co/ZfOQ99WMST>

hurricane



Hurricane Mora #damages #Rohingya #homes
<https://t.co/qGsB9rI3s1> <https://t.co/aTdiEoBmC7>

fire



First responders sue Arkema over injuries in Houston chemical plant fire after???
<https://t.co/UfhheYfJuH> <https://t.co/eJrY6iD8oV>

other_disaster



Tornado activity in Palm Bay. A roof blew off a home and at least six mobile homes were destroyed
<https://t.co/mQ9DM9SWD>

flood



'Catastrophic flooding' to hit Texas after Harvey
<https://t.co/w1WlpQQIgy> <https://t.co/P62stw6Mz2>

not_disaster



RT @BSO: Lil' Flip Defends Donald Trump, Calls Out His Haters Over Hurricane Harvey (Video)
<https://t.co/YrKT09IbgE> <https://t.co/1Uz8D9v5uT>

Figure A.1: Examples of annotated data.

Prompt A.1: Disaster-Type Classification Prompt (Part 1)

1: Your task is to analyze a social media post containing text and/or images and:
2: 1. Decide whether the post is related to a disaster.
3: 2. If it is, identify the **single most likely** disaster type from the list below.
4: 3. Output the result in the specified JSON format.
5: Allowed disaster types: **earthquake, fire, flood, hurricane, landslide, other_disaster**
6: ---
7: ## Follow 5 steps
8: ### **Step 1 | Multi-perspective Observation**
9: Inspect the post and note any disaster clues from each perspective.
10: ```
11: [Observation]
12: - Text: ... → possibly indicates ...
13: - Image: ... → suggests ...
14: - Prior knowledge: ... → (consistent / inconsistent)
15: ```
16: ---
17: ### **Step 2 | Disaster Type Evaluation**
18: For every disaster category that appears relevant, state **Yes / No** with a short justification.
19: ```
20: [Evaluation]
21: - Flood: Yes (river overflow described; submerged streets visible)
22: - Earthquake: No (no shaking or collapse)
23: - Fire: No (no flames or smoke)
24: ...
25: ```
26: ---
27: ### **Step 3 | Self-Consistency Voting**
28: Repeat **Steps 1-2 three times**. Record the disaster type predicted in each run and count the votes.
29: ```
30: [Voting]
31: Run 1: "flood"
32: Run 2: "earthquake"
33: Run 3: "flood"
34: → Provisional winner: "flood" (2 / 3 votes)
35: ```
36: ---

Prompt A.2: Disaster-Type Classification Prompt (Part 2)

38: ### **Step 4 | Choose the Most Visually Clear Disaster**
39: If two or more disaster types are tied, or the vote margin is 1, break the tie using the image:
40: 1. Re-examine all images (or captions).
41: 2. Choose the disaster type **most clearly depicted** visually.
42: 3. Declare this as the **final winner**.
43: If a single provisional winner already has a clear majority, skip this step.
44: ---
45: ### **Step 5 | Structured Output**
46: Return the result in perfectly-formatted JSON (no code block):
47: ```
48: [
49: {
50: "category": "disaster", // or "others" if not disaster-related
51: "type": "flood" // single disaster type, empty string if none
52: }
53:]
54: ```
55: ---
56: ## Example
57: ```
58: [Observation]
59: - Text: River burst its banks and flooded nearby houses → possible flood
60: - Image: Street under water up to car windows → supports flood
61: - Prior knowledge: Area has frequent flooding → consistent
62: [Evaluation]
63: - Flood: Yes (multiple supporting cues)
64: - Earthquake: No (no structural damage visible)
65: - Fire: No (no flames or smoke)
66: [Voting]
67: Run 1: "flood"
68: Run 2: "flood"
69: Run 3: "flood"
70: → Provisional winner: "flood"
71: → Final winner: "flood" (tie-break not needed)
72: [Final Output]
73: [
74: {
75: "category": "disaster",
76: "type": "flood"
77: }
78:]
79: ```
