Analyzing Post-Catastrophe Social Media for Efficient Emergency Response

Xingchen Liu, Mengqi Jia, Xin He, Zack Zang

There is often an explosion of tweets on social media after a disaster, with topics ranging from the latest news, cries for help, moral support, and financial support. For decision-makers such as government agencies, the lack of appropriate tools for retrieving the most critical information can hinder efficient emergency response. In this paper, both supervised and unsupervised learning techniques in natural language processing are employed to classify tweets and extract key topics from disaster-related tweets, with the goal of helping relevant entities make informed decisions.

I. INTRODUCTION

Background and Motivation: In the era of big data, social media platforms like Twitter have become critical communication channels during natural disasters. Acting as distributed "social sensors," users provide real-time, on-the-ground updates that traditional monitoring systems may fail to capture. However, this massive influx of information creates a significant challenge: information overload characterized by a high noise-to-signal ratio. During crises, actionable intelligence—such as reports of infrastructure damage or urgent requests for help—can be buried among millions of messages conveying sympathy, general news, or redundant information. Manual processing of such data streams is computationally infeasible and inefficient for time-sensitive rescue operations.

Data and Methodology: To address this information bottleneck, the project leverages the HumAID (Human-Annotated Disaster Incidents Data from Twitter) dataset [1]. Building upon HumAID's large-scale annotations, we develop an intelligent processing framework that integrates two complementary approaches:

1) Supervised Learning: Using the annotated labels in HumAID, we train classification models—ranging from classical algorithms such as Logistic Regression (LR) to deep learning models such as Bi-LSTM and BERT to automate the filtration of critical updates from irrelevant content.

2) Unsupervised Learning: Clustering and topic modeling techniques are used to explore latent structures in the data. Since disaster scenarios evolve dynamically, unsupervised methods can help uncover emerging patterns or sub-events that fall outside predefined categories.

Humanitarian Impact: The broader goal of this project is to support humanitarian aid and decision-making processes. By transforming unstructured text into actionable intelligence, we aim to enhance situational awareness for organizations such as the UN Office for the Coordination of Humanitarian Affairs (OCHA) and the Red Cross. The data-driven approach may enable decision-makers to rapidly identify severely affected areas and determine specific resource needs (e.g., water, food, medical supplies), thereby optimizing resource allocation and improving response effectiveness.

II. DATASET DESCRIPTION

To address the information bottleneck described above, the project utilizes the HumAID dataset, which contains approximately 77,000 manually annotated tweets spanning 19 major natural disasters between 2016 and 2019. HumAID's core contribution is its transformation of chaotic social media streams into structured intelligence. Tweets are classified into several humanitarian categories, including (a) Caution and advice, (b) Displaced people and evacuations, (c) Don't know/can't judge, (d) Infrastructure and utility damage, (e) Injured or dead people, (f) Missing or found people, (g) Not humanitarian, (h) Other relevant information, (i) Requests or urgent needs, (j) Rescue, volunteering, or donation effort, (k) Sympathy and support.

Additional attributes—such as disaster location, disaster type, and year—were extracted.

The original tweet text underwent cleaning to ensure consistency and remove noise. Specifically, non-string inputs were converted to empty strings to avoid `NaN`-related errors; all text was lowercased (`text = text.lower()`); URLs, user mentions (@), and hashtags (#) were removed sequentially; the HTML entity `amp;` was replaced with `and`; and non-alphanumeric characters were filtered to retain only lowercase English letters, digits, and whitespace while converting all other symbols to spaces. Finally, redundant whitespace was collapsed and stripped from both ends of the text. The resulting dataset comprises six attributes, summarized in Table I.

TABLE I. ATTRIBUTES OF DATASET.

Attribute Name	Description	Data Type
id_str	Tweet ID	string
tweet_text	Content of the tweet	string
class_label	Category of the tweet content	string
place	Location of the disaster	string
disaster	Type of disaster	string
year	Year of disaster	int

III. SUPERVISED LEARNING

We first examine the topic classification task: classifying disaster-related tweets into one of ten human-annotated categories using the tweet text attribute alone. Three supervised learning algorithms are explored—logistic regression, Bi-LSTM, and BERT—each well-suited for natural language processing tasks. We acknowledge that random splitting may introduce optimistic bias, and leave event-level generalization as future work.

A. Logistic Regression

Logistic regression computes the softmax of the input feature vector to produce class probabilities. Since the algorithm requires numerical inputs, a TF–IDF vectorizer is constructed to transform tweet texts into token-frequency vectors.

We can then perform training based on these vectorized inputs. Despite being a traditional machine learning algorithm, logistic regression performs surprisingly well on this dataset, achieving a test accuracy of 73.8% and a macro-F1 score of 60.0%.

Training on these vectorized inputs reveals that logistic regression performs surprisingly well on this dataset, achieving a test accuracy of 73.8% and a macro-F1 score of 60.0%. Figure 1 visualizes the L2 weight norms of tokens for each topic, highlighting that only a small number of tokens strongly influence classification decisions. For example, under the "Urgent Needs" category, the most salient tokens—need, help, and pleas—are intuitively connected to the topic. This suggests that logistic regression effectively identifies key tokens associated with each class.

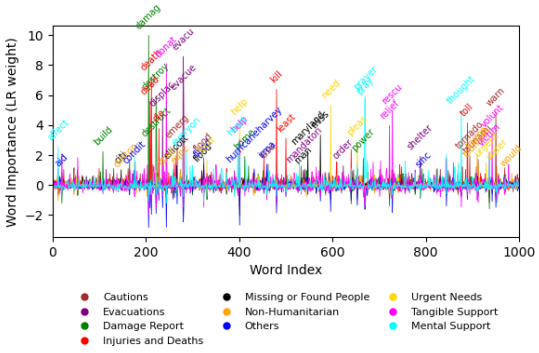

FIG. 1. L2 weight norms of the top six stemmed tokens to each topic

B. Bi-LSTM

Although logistic regression is efficient and highly accurate, it does not leverage token context for coherent language understanding. Recurrent neural networks such as Bi-LSTM overcome this limitation by remembering tokens across sequences, providing contextual awareness from both past and future inputs.

The Bi-LSTM architecture consists of an embedding layer, a Bi-LSTM layer, and a fully connected output layer. The model converges after only two epochs, achieving 75.3% accuracy and a macro-F1 score of 66.1%, marking a modest improvement over logistic regression.

FIG. 2 shows a 2D t-SNE visualization of word embeddings, where representative words

within the same topic form visible clusters. Tokens associated with damage and devastation cluster in one region, while those representing relief or fundraising cluster in another.

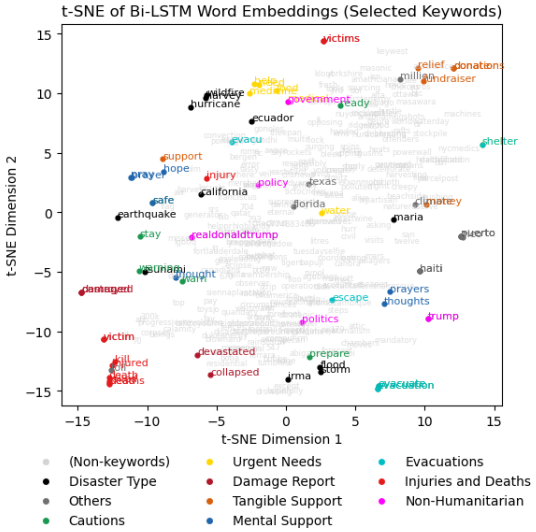

FIG. 2. Two-dimensional t-SNE plot of the word embeddings

C. BERT (Transformer-based Model)

To further leverage contextual and semantic information in disaster-related tweets, we adopt a Transformer-based model using a pretrained BERT encoder and fine-tune it on our self-constructed dataset. Unlike traditional bag-of-words models and recurrent architectures, BERT employs multi-head self-attention to model long-range dependencies and bidirectional context within a tweet. The `[CLS]` token representation from the final encoder layer is used as a global sentence embedding and fed into a linear classification head to predict the topic category.

The model is fine-tuned for two epochs under identical hyperparameter settings on two variants of the dataset: the original tweet text (`tweet_text`) and a cleaned version (`text_clean`) where punctuation and non-alphanumeric symbols are removed. When trained on the original tweet text, the model achieves a test accuracy of 79.1% and a macro-F1 score of 75.7%. Training on cleaned text yields a comparable accuracy of 78.8% and macro-F1 score of 75.4%. While text cleaning slightly improves inference throughput, it does not result in a consistent performance gain, suggesting that the BERT tokenizer and encoder are inherently robust to noisy social media text and are able to exploit punctuation and informal tokens as useful contextual cues.

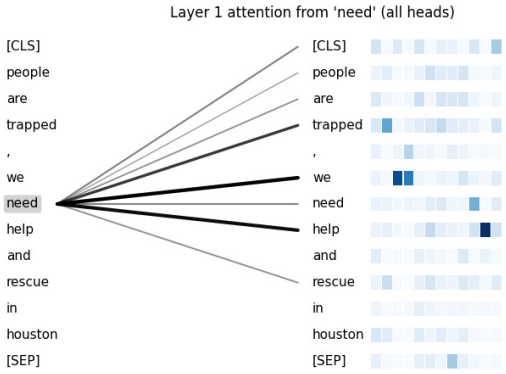

FIG. 3. Multi-head self-attention weights from the first Transformer layer of the fine-tuned BERT model, visualized for a representative tweet. Attention originates from the query token *"need"*.

To interpret how the Transformer model processes disaster-related language, we visualize token-level self-attention weights extracted from the first Transformer layer, as shown in Fig. 3. A representative tweet is selected, and attention originating from the query token *"need"* is examined across all attention heads. Each row corresponds to a token in the input sequence, while the multiple color blocks indicate attention weights from different heads.

The visualization reveals that semantically meaningful tokens such as *"we"*, *"help"*, and *"rescue"* consistently receive higher attention across multiple heads, whereas function words and punctuation receive relatively low attention. This indicates that different attention heads capture complementary aspects of disaster semantics, collectively emphasizing tokens that are highly relevant to urgent humanitarian needs.

The effect of fine-tuning on the learned representations is further illustrated in Fig. 4, where a two-dimensional t-SNE projection of the `[CLS]` embeddings is shown before and after fine-

tuning on the `tweet_text` variant. Before fine-tuning, embeddings from different categories are highly entangled, reflecting the fact that the pre-trained BERT encoder is not specialized for disaster-domain classification. After fine-tuning, samples belonging to the same category form noticeably tighter clusters with clearer inter-class separation. This demonstrates that supervised fine-tuning reshapes the embedding space to become more discriminative for the downstream classification task.

Overall, the Transformer-based model achieves the best performance among all evaluated methods on our dataset. In addition to strong quantitative results, attention visualization and embedding analysis provide qualitative evidence that BERT effectively learns meaningful disaster-related representations. These results highlight the advantage of Transformer architectures in capturing contextual and semantic information in short, noisy social media texts.

D. Comparison of Supervised Learning Methods

To better understand the strengths and limitations of different supervised learning approaches, we summarize the performance and characteristics of the three evaluated models: logistic regression, Bi-LSTM, and BERT in Table II. Given the strong class imbalance in humanitarian categories, macro-F1 is reported to better reflect performance on minority but operationally critical classes. These models represent increasing levels of model complexity and capacity for contextual understanding, ranging from linear classifiers with bag-of-words features to deep Transformer architectures.

From a performance perspective, logistic regression already achieves a strong baseline accuracy of 73.8%, despite relying solely on TF-IDF features and ignoring word order. This indicates that disaster-related tweets often contain highly discriminative keywords that can be effectively captured by simple linear models. Bi-LSTM further improves performance by modeling sequential context, achieving an accuracy of 75.3% and a higher macro-F1 score, demonstrating the benefit of incorporating contextual information across tokens.

The Transformer-based BERT model achieves the best overall performance, with test accuracy exceeding 79% and a macro-F1 score above 75%. Beyond quantitative gains, BERT provides richer interpretability through attention mechanisms and more discriminative sentence-level representations, as illustrated by the attention visualization and t-SNE analysis. Although BERT introduces higher computational cost, its superior representation learning capability makes it particularly suitable for complex and noisy social media text.

Overall, the comparison highlights a clear trade-off between model complexity, interpretability, and performance. While traditional models such as logistic regression remain competitive and efficient, deep contextual models offer consistent improvements and deeper insights into the semantic structure of disaster-related tweets.

IV. UNSUPERVISED LEARNING

Beyond supervised topic classification, extracting and summarizing the topics that emerge organically from disaster-related tweets can provide governments, NGOs, and local communities with a deeper understanding of what people actually care about during and after crises. Such unsupervised analysis helps reveal both the unique characteristics of individual events and the common patterns of needs, reactions, and support across different disasters.

In this section, we employ Latent Dirichlet Allocation (LDA) to discover the most frequently discussed topics from the tweet corpus. LDA is a generative statistical model in which each tweet is represented as a mixture of latent topics, and each topic is modeled as a distribution over words with Dirichlet priors. After tokenization and stop-word removal on the cleaned text field, we train an LDA model separately for each year from 2016 to 2018, using 5 topics per year, and select 6 representative words for each topic. The resulting topics are visualized in Fig

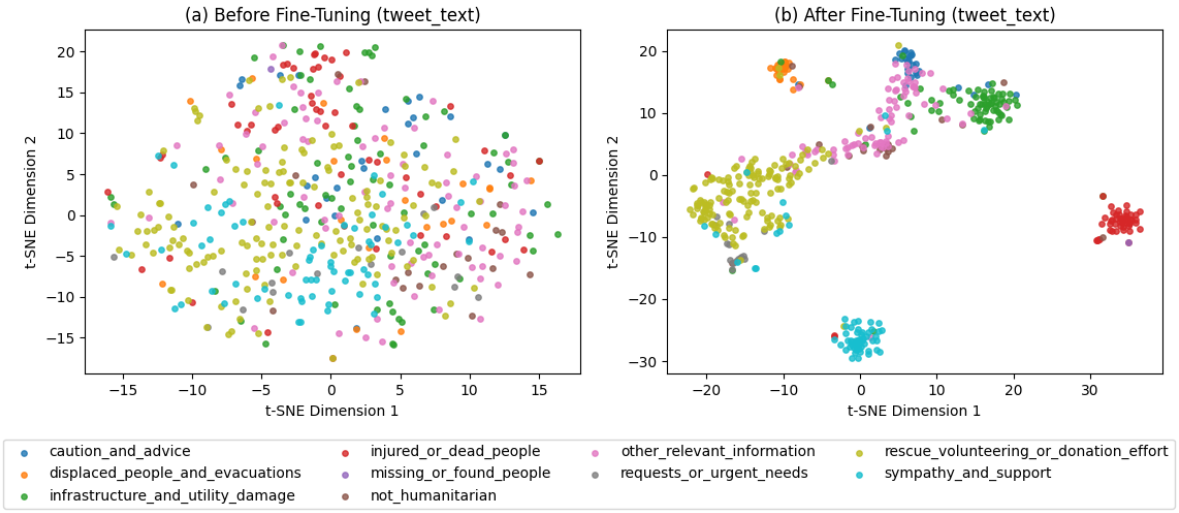

FIG. 4. t-SNE visualization of [CLS] embeddings before and after fine-tuning BERT on the original tweet text.

TABLE II. Comparison of supervised learning methods for disaster-related tweet classification

Model	**Feature Type**	**Context Modeling**	**Accuracy (%)**	**Macro-F1 (%)**
Logistic Regression	TF-IDF	No	73.8	60.0
Bi-LSTM	Word Embeddings	Yes (Sequential)	75.3	66.1
BERT	Contextual Embeddings	Yes (Self-Attention)	79.1	75.7

5

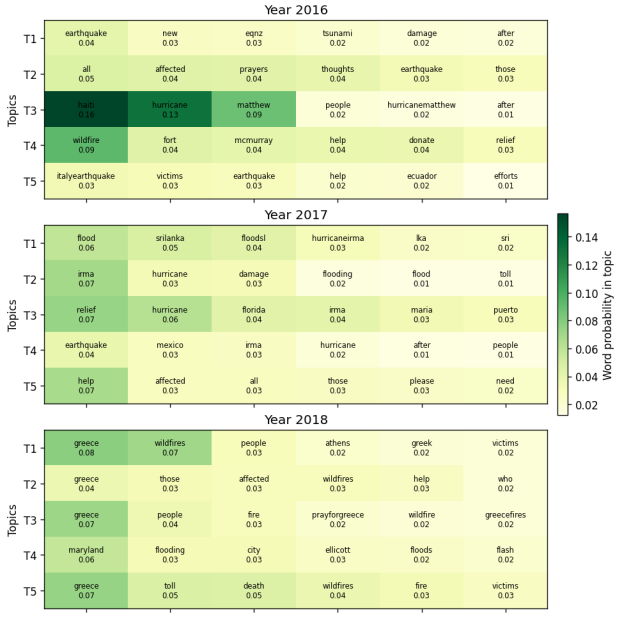

FIG. 5. Topic modeling results from Latent Dirichlet Allocation in years 2016-2018. Horizontal axis: 6 representative words for each topic. Vertical axis: mostly discussed topics in each year.

From the topic modeling results, several observations can be made. First, as the years progress, the prominent topics evolve dynamically with the major disasters of each period, but the changes are relatively mild and mainly driven by the appearance of new events. For instance, in 2016 the discussions are dominated by earthquakes, tsunamis, and wildfires. In 2017, hurricanes such as Harvey and Maria, as well as earthquakes in Mexico, become central. In 2018, large floods and California wildfires take more of the spotlight. Nevertheless, across the three years there is a strong consistency in the types of issues that attract attention on Twitter: earthquakes and wildfires repeatedly draw public concern, and similar patterns of damage, risk, and response recur over time.

Second, the topics exhibit a clear separation between event-specific and disaster-independent themes. Some topics focus on particular events and regions (e.g., specific hurricanes, earthquakes in certain cities, or local flooding), while others capture more general concerns that appear across many disasters, such as: calls for

help and donations, reports of casualties, property damage, and aftershocks, ongoing worries about safety, infrastructure, and recovery.

These more generic topics are relatively stable across years, even though the concrete disasters behind them change.

Third, within each individual year, LDA produces several well-separated and interpretable themes. For example, in 2017 the five sets of representative words can be interpreted as: Topic 1: Support and donations across multiple disasters. Topic 2: Disaster names and event identifiers. Topic 3: Floods and hurricanes, including heavy rainfall and storm surge. Topic 4: Rescues, evacuations, and shelters. Topic 5: Earthquakes in Mexico and discussions of U.S.–Mexico relations. These topics jointly capture both the physical impact of disasters and the social responses they trigger.

To further justify the validity of the LDA topics, we compare them with the actual use of hashtags in the tweets. We focus on tweets related to Hurricane Harvey in 2017, which caused severe flooding in Houston. We identify Harvey-related tweets as those labeled as hurricanes in the disaster field and whose raw text contains the string "Harvey" (case-insensitive). From this subset, we extract all hashtags of the form word, compute their frequencies, and select the 30 most common hashtags. These are manually grouped into six semantic categories: Support and donations, Disaster names, Floods and hurricanes, Rescues, evacuations, and shelters, Mexico-related and Others.

The bar plot of hashtag frequencies in Fig 6, plotted on a logarithmic scale, shows that most of the high-frequency hashtags fall into three major categories: support and donations, disaster names, and floods/hurricanes. These categories are highly consistent with the LDA topics described above. In contrast, hashtags corresponding to rescues/shelters and Mexico-related issues are less frequent, which is reasonable given the geographic and political context of Hurricane Harvey.

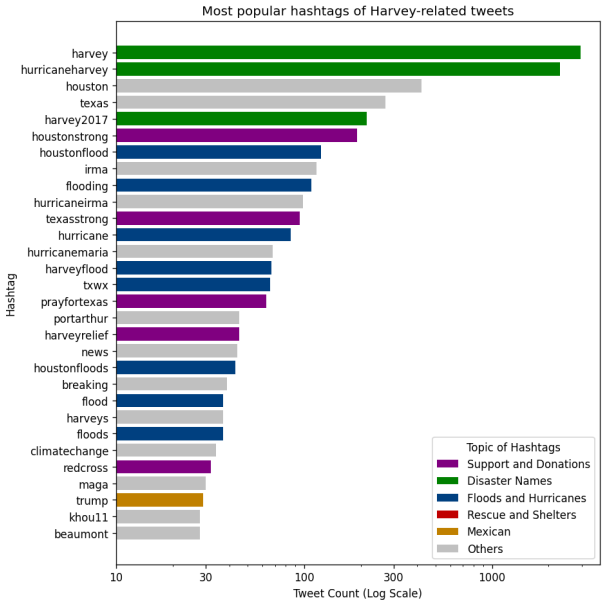

FIG. 6. Most popular hashtags of Harvey-related tweets.

The temporal dimension of topic evolution can also be examined around a specific event. For Hurricane Harvey, we approximate tweet timestamps by decoding the millisecond-level time information embedded in the Twitter snowflake tweet-id. Using a pivot date of August 29, 2017 as a turning point during the event, we split Harvey-related tweets into two groups: those posted during Harvey (on or before this date) and those posted after Harvey (after this date). For each group, we compute word frequencies from the preprocessed tokens and visualize the most salient words in the form of word clouds Fig 7.

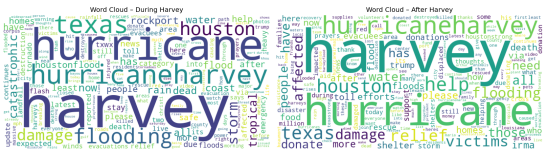

FIG. 7. Most popular words in the tweet corpus during and after Hurricane Harvey.

From these word clouds, we observe that during the hurricane, tweets mainly focus on floods, damages, and immediate impacts on victims, which aligns with the LDA topics on floods/hurricanes and rescues. After the hurri-

cane, the emphasis gradually shifts towards support, relief, and donation efforts, reflecting the transition from acute crisis to recovery and long-term assistance. Disaster names such as "Hurricane Harvey" remain prominent across both periods, serving as a persistent anchor for public discussion.

Overall, the topics extracted by LDA are consistent with the most frequent words and hashtags observed in the tweet corpus, both across years and around a specific event such as Hurricane Harvey. This confirms that unsupervised topic modeling can effectively capture the key dimensions of public attention and response in disaster-related social media data. In practice, such unsupervised topic monitoring can serve as an early-warning or situational-awareness tool, complementing supervised classifiers by highlighting emerging needs or previously unseen sub-events.

V. CONCLUSION

This study shows that both supervised and unsupervised NLP methods can effectively extract critical information from disaster-related tweets. Supervised models, particularly BERT, demonstrate strong performance in classifying humanitarian needs, while topic modeling reveals broader patterns in public attention across events and time. Overall, these approaches highlight the potential of social media analytics to enhance situational awareness and support more timely and informed emergency response efforts.

[1] Firoj Alam, Umair Qazi, Muhammad Imran, and Ferda Ofli. Humaid: Human-annotated disaster incidents data from twitter. In *15th International Conference on Web and Social Media (ICWSM)*, 2021.