Crisis Observatory: Extracting Credible Signals During a Crisis in the Age of LLMs

Kuan-Chieh Lo *, Pranav Maneriker *, Sriram Sai Ganesh *, Dominik Winecki *, Kelly Garrett *, Ayaz Hyder *, Arnab Nandi *, Valerie Shalin †, Shannon Bowen ‡, Amit Sheth ‡, Srinivasan Parthasarathy *
* The Ohio State University † Wright State University ‡ University of South Carolina
Correspondence: lo.311@osu.edu, srini@cse.ohio-state.edu

***Abstract*—Systems for crisis response have required several different models for the analysis of unstructured text, such as identifying needs, locations, topics, routing, and matching of needs with available responders. Large Language Models (LLMs) have replaced task-specific models across various language processing tasks. However, LLMs are known to be limited by their training data, collected before the crisis. In this demo, we explore the use of LLMs for crisis response scenarios with rapidly evolving information environments. We show how augmentation of these models with external reliable sources of crisis-specific information can help build adaptive systems for response. The demonstration video can be found at: https://youtu.be/jKeU5WsG20o.**

***Index Terms*—Crisis management, Large language models, Visualization**

I. Introduction

Individuals affected by crises often turn to social media for emergency resources, providing valuable citizen-sensed insights. For decision-makers, it's crucial to distinguish genuine needs from misinformation. Current crisis response tools for such content focus on data collection, need identification, and location extraction [1], [2] but cannot provide key insights, especially regarding information trustworthiness and underlying themes necessary for sensemaking [3].

State-of-the-art Large language models (LLMs) potentially eliminate the need for extensive task-specific model building. However, such models often generate non-factual text, a fundamental liability during a crisis [4], [5]. Moreover, in a dynamic environment like a crisis, LLMs are unable to generalize to text generated after the training data cut-off period [6].

In this demonstration, we present *Crisis Observatory*, a system powered by LLM technology (while alleviating its limitations) that organizes social media content based on credibility and extracts actionable information, adapting to dynamic situations. This system supports emergency responders, crisis managers, and decision-makers in quickly assessing and addressing evolving crises.

Our contributions are threefold. We: (1) address the daunting task of extracting dependable, actionable information from social media during emergencies, especially when traditional communication channels are compromised. (2) provide a robust and all-encompassing solution by integrating credibility assessment and topic detection using LLMs. (3) display tweets with their geographic locations, topics, and credibility assessments in real-time, using a web-based dashboard that allows users to filter information temporally and spatially while visualizing trends and connected content.

II. Prior Work

Social media platforms are essential for gathering and sharing information during crises. Example efforts include AIDR [1], which uses machine learning to identify and classify crisis-related content, thereby enabling quicker responses. D-Record [7] is a multi-modal system that combines text, gazetteers, and imagery to match location-specific demands in disaster relief. It classifies needs expressed in tweets, geolocates them via OpenStreetMap, and filters out inaccessible resources using satellite flood mapping data. D-Record is powered by HUG-FM [8] and SEANO [9] leveraging a semi-supervised algorithm to map flood extents from imagery. While such systems excel in data collection, flood extent mapping and need identification, they lack the ability to assess and organize information along credible strands for effective crisis response. Our proposed system fills this gap to provide a comprehensive view of crises and enhance disaster response efforts.

III. Methodology

In this demonstration, we present the Crisis Observatory system using Twitter data from two significant crisis events: *Hurricane Ian*, a natural disaster causing widespread destruction; and the *COVID-19 pandemic*, a global disease outbreak affecting millions worldwide. The system employs multiple LLM agents built on an open-sourced LLM—Mistral-7B-Instruct-v0.2 [10]: one agent filters irrelevant data by identifying vague topics (§III-A), another agent detects the key discussion themes in the tweet corpus (§III-B), and the third agent identifies and verifies claims by retrieving factual information from external knowledge repositories (§III-C). Additionally, given the critical role of location-based information in crisis response, we incorporated geospatial analysis (§III-D) to enhance situational awareness and identify interdependencies between crisis events. Figure 1 illustrates the system framework, with detailed functionalities described in the following subsections.

A. Data Collection and Preprocessing

We obtained tweets related to COVID-19 from the COV19Tweets dataset [11], which is gathered by monitoring

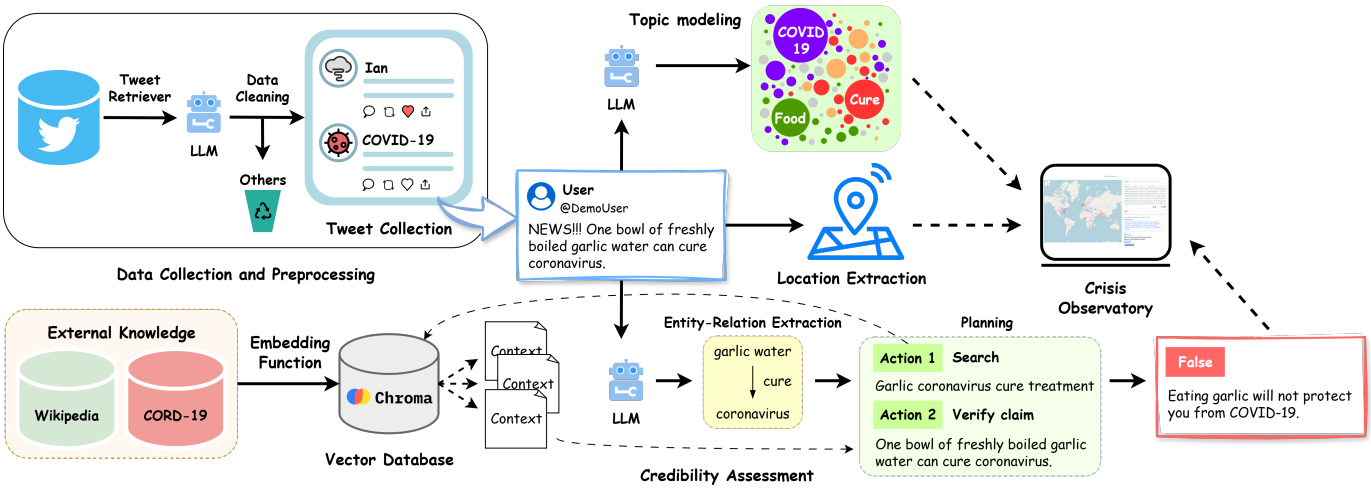


Fig. 1: Overview of the Crisis Observatory system.

a real-time Twitter feed for coronavirus-related keywords and hashtags starting from October 1, 2019. For Hurricane Ian, we used the twarc2[1] tweet retriever to gather relevant tweets mentioning the hurricane from September 23, 2022 to October 23, 2022. Any tweets that only included hashtags, mentions, or hyperlinks were removed. We developed prompts that included several examples to categorize each tweet into one of three predefined categories: 'Ian', 'COVID-19', or 'Others' (unrelated to either event). In this demonstration, we retain only the tweets assigned to 'Ian' and 'COVID-19', and drop tweets that are categorized as 'Others'.

B. Labeled Topic Modeling

Topic modeling algorithms can be beneficial for extracting latent semantic structures and discussion topics from large corpora of documents. We employed TopicGPT [12] to extract latent semantic structures and discussion topics from our tweet collection through a two-stage process: *topic generation* and *topic assignment*. In the topic generation stage, we first initialized seed topics using BERTopic [13], which leverages Sentence-BERT [14] to encode tweets into embedding clusters for interpretable topic extraction. These seed topics were then fed into TopicGPT, where the model either assigns a subset of tweets to existing topics or creates new topics with justification when tweets do not fit existing categories. In the topic assignment stage, we provided the TopicGPT with the extracted topic list and exemplar prompts demonstrating proper tweet-to-topic matching, enabling the model to assign topics that best align with given tweets.

C. Credibility Assessment and Explanations

Our approach for automated tweet credibility assessment employs a multi-stage pipeline that combines LLM with external knowledge retrieval and structured verification planning. We construct a comprehensive knowledge base by integrating *Wikipedia* for general domain knowledge and *CORD-19* [15] for COVID-19-related scientific literature. Documents from both sources are encoded using a pretrained embedding model (Instructor-xl [16]) and indexed in a ChromaDB vector database[2] using hierarchical navigable small world (HNSW) algorithm [17] for efficient similarity search, with metadata preserved for provenance tracking.

For credibility assessment, we implement a three-stage verification process: (1) *Entity-Relation Extraction*, where we parse tweet content to extract entities and their relationships, converting unstructured text into structured knowledge representations for targeted verification; (2) *Verification Planning*, where a planner identifies verifiable claims and generates structured search queries and validation actions; and (3) *Evidence Retrieval and Assessment*, where the system queries the vector database using the generated search terms to retrieve the top-k most semantically similar passages. The system then employs retrieval-augmented generation (RAG) [18] to synthesize the retrieved evidence to assess tweet credibility while generating transparent explanations that justify the credibility determination.

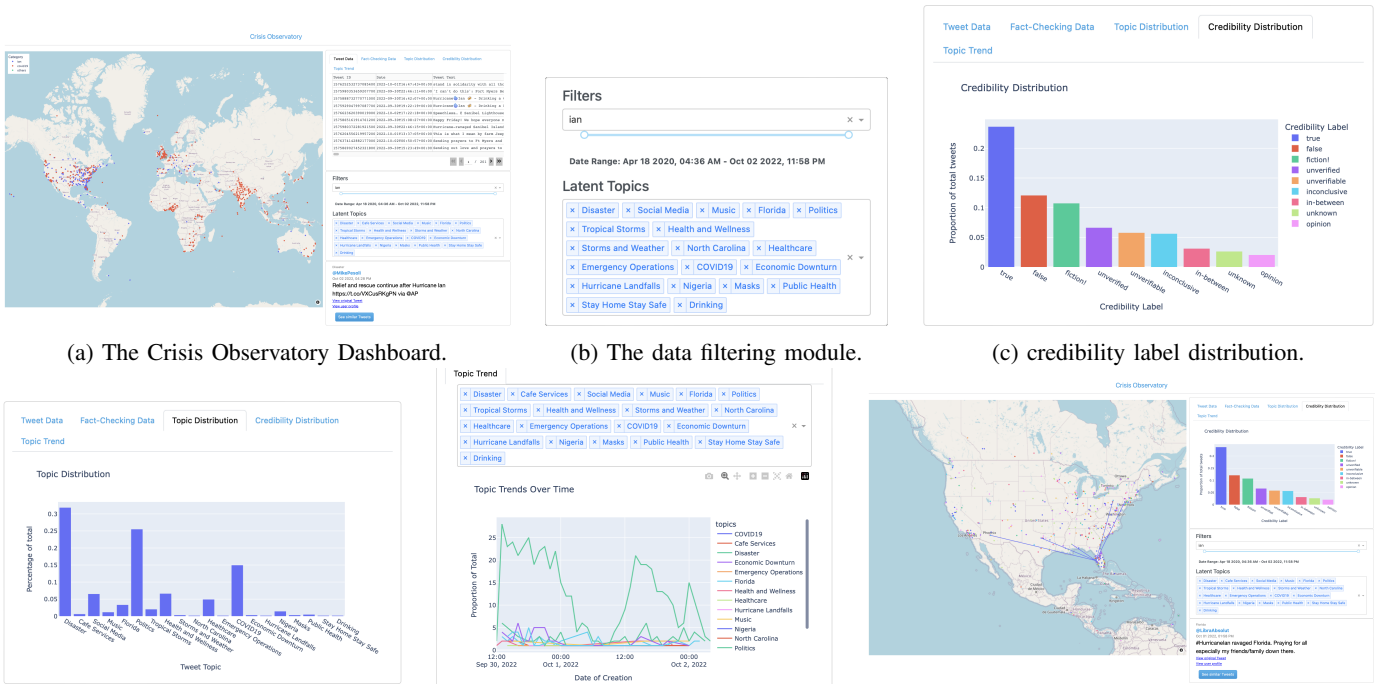
D. Location extraction and Geolocation

To achieve high-quality location detection, we utilized text entities from tweets, author location, and tweet geolocation when available. Next, we created bounding boxes corresponding to each of these entities using a geocoding API[3]. To solve a preference-weighted maximum overlapped area problem, we constructed a tree of bounding boxes with different sizes. Successively adding bounding boxes for each location data point about a tweet, we obtain a hierarchy of nodes based on differing location data resolutions. The root node denotes

[1]https://twarc-project.readthedocs.io/

[2]https://www.trychroma.com/

[3]We used OSMNames-sphinxsearch



(a) The Crisis Observatory Dashboard. (b) The data filtering module. (c) credibility label distribution.

(d) Latent topics distribution. (e) Visualization of the topic trends. (f) Connecting the tweets with similar content.

Fig. 2: The Crisis Observatory Dashboard showcases the system's interactive web-based interface for visualizing and exploring extracted insights from social media data during crisis events.

a bounding box corresponding to the entire planet. We then search for the deepest leaf corresponding to the heuristic of highly likely (all nodes on the path to a leaf overlap the box) and highly specific (as it is the most contained box) location information using the shapely python library. Finally, we tuned the geolocation by designating a preference order of entity node additions, prioritizing device and author geolocations (that tend to be less noisy) higher than location entities extracted from the tweet text.

IV. Demonstration

To present emergent crisis information efficiently and intuitively, we developed the Crisis Observatory system with an interactive web framework that displays tweets with extracted geolocation data and corresponding topic and credibility labels (Figure 2a). The user interface features three primary modules designed for utility and responsiveness: A) the map module, B) the data filtering module, and C) the information panel. These modules enable users to explore and filter data geospatially while deriving actionable statistical results.

A. The map module

This module provides a geographic map for viewing geometric information and interacting with the system. Users can manipulate the current pan, tilt, and zoom of the map by either clicking and dragging regions of the map or using the on-screen controls. Hovering the cursor over a specific tweet marker on the map presents further information, such as the corresponding topic and credibility label. The Map Module also enables geographic filtering of tweet data. Users can draw bounding rectangles or use a free-form lasso tool to subset regions of interest, subsetting the global tweet data to include only tweets from the selected region.

B. The data filtering module

This module enables users to quickly narrow down the most relevant information for a specific period (See Figure 2b). At a high level, users can select crisis themes, such as Hurricane Ian or COVID-19 in our demonstration, to view relevant tweets. The time interval of interest can be chosen using a slider. The module also presents a list of latent topics derived from the currently visible tweets. For fine-grained topic filtering, users can add or remove individual topics from this list. Any changes made to the subset of tweets of interest trigger automatic updates to all statistical and graphical visualizations in the system, providing users withthe most up-to-date and relevant information for their selected criteria.

C. The information panel

This panel provides comprehensive information including tweet details, credibility assessments, summary statistics, and visualizations. Users can access individual tweet content with corresponding latent themes and credibility assessments, where each assessment includes a label, explanation, and links to supporting evidence. The panel enables quick inspection of credibility label distributions (Figure 2c) and latent topic distributions (Figure 2d) within crisis information, allowing users to grasp overall credibility and main themes. Topic

TABLE I: Selected examples.

Topic	Tweet	Themes	Credibility	Explanasion
Ian	It's been a crazy week here in Florida.I got really lucky and feel really blessed to have escaped any kind of storm damage.Others have not been as fortunate.	Disaster; Disaster Relief; Damage	True	The statement accurately reflects the situation in Florida during the recent storm.
Ian	Amazon Worker Delivers to 172 People During Hurricane Ian: 'I Hate Y'all'.	Disaster Relief; Family; Love	Unknown	Without additional context or information, it is not possible to determine the accuracy of this statement.
Covid19	Someone that would have been reported dead from UK if only Africa was hit with corona.	Conspiracy Theories	False	There is no information in the context provided that supports the claim that someone would have been reported dead from the UK if only Africa was hit with corona.
Covid19	1st Major plan after the MCO and this whole COVID thing should be "a peaceful weekend in Penang." Or maybe a bit further, Singapore sound great too !!	Family; Leisure	Opinion	The statement is an opinion and does not contain any factual claims to be evaluated for accuracy.

trend visualizations (Figure 2e) show prevalence over time, helping identify emerging issues and track ongoing concerns. Additionally, users can find similar tweets, with connecting links displayed on the map module (Figure 2f) to explore related content and discover additional relevant information.

V. Guardrails

We considered several safeguards against LLM errors in this demonstration: (1) Automated grounding assessments rely on reputable external sources like the CORD-19 dataset and Wikipedia. (2) The RAG pipeline, a critical element in enhancing system trustworthiness, empowers us to generate relevant and verifiable outputs. (3) Each automated fact-checking claim is designed to be interactive, surfacing relevant links to grounding information sources for users to cross-verify. These mechanisms function as critical quality measurements within our system. While a thorough evaluation is outside the scope of our demonstration, we assessed the system's efficacy through selective sampling and manual verification of outputs for readability and accuracy. Table I presents an illustrative subset of examples from our system.

VI. Conclusion and Future Work

In this demonstration, we presented Crisis Observatory, a novel crisis response system that leverages LLMs to: (1) filter incoming data and detect emerging discussion themes as the crisis evolves. (2) validate new claims against trusted sources as they emerge, adapting its credibility assessments to evolving situations. (3) enable users to dynamically explore and analyze crisis information through temporal, spatial, and topical dimensions, supporting adaptive decision-making during crisis response.

Future work could scale Crisis Observatory to support live data streaming from user-defined sources. Surfacing grounded and trustworthy viewpoints from other social media websites will enhance the utility of our system. These enhancements will make the system a more robust and versatile crisis response and analysis tool.

References

[1] M. Imran, C. Castillo, J. Lucas, P. Meier, and S. Vieweg, "Aidr: artificial intelligence for disaster response," in *ACM WWW*, 2014.

[2] S. Ghaffarian, F. R. Taghikhah, and H. R. Maier, "Explainable artificial intelligence in disaster risk management: Achievements and prospective futures," *International Journal of Disaster Risk Reduction*, vol. 98, 2023.

[3] K. E. Weick, *Sensemaking in organizations*. Sage, 1995, vol. 3.

[4] A. Birhane, A. Kasirzadeh, D. Leslie, and S. Wachter, "Science in the age of large language models," *Nature Reviews Physics*, vol. 5, no. 5, pp. 277–280, 2023.

[5] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[6] T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar *et al.*, "Freshllms: Refreshing large language models with search engine augmentation," *arXiv preprint arXiv:2310.03214*, 2023.

[7] S. Kar, H. S. Al-Olimat, K. Thirunarayan, V. L. Shalin, A. Sheth, and S. Parthasarathy, "D-record: Disaster response and relief coordination pipeline," in *ACM SIGSPATIAL Workshop on Advances on Resilient and Intelligent Cities*, 2018.

[8] J. Liang, P. Jacobs, and S. Parthasarathy, "Human-guided flood mapping: From experts to the crowd," in *ACM WWW*, 2018.

[9] J. Liang, P. Jacobs, J. Sun, and S. Parthasarathy, "Semi-supervised embedding in attributed networks with outliers," in *Proceedings of the 2018 SIAM international conference on data mining*. SIAM, 2018, pp. 153–161.

[10] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. Chaplot, D. Casas, F. Bressand, G. Lengyel *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[11] R. Lamsal, "Coronavirus (covid-19) tweets dataset," 2020. [Online]. Available: https://dx.doi.org/10.21227/781w-ef42

[12] C. M. Pham, A. Hoyle, S. Sun, and M. Iyyer, "TopicGPT: A prompt-based topic modeling framework," 2023.

[13] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[14] N. Reimers, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[15] L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide *et al.*, "Cord-19: The covid-19 open research dataset," *ArXiv*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:216056360

[16] T. Wolf, "Transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2020.

[17] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 4, pp. 824–836, 2018.

[18] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *CoRR*, vol. abs/2005.11401, 2020.