Mapping Natural Disasters Using Social Media Posts with an Encoder-Decoder Model

Nima Ekhtari [1], Andrew Niccolai[2], Kevin Slocum [3], Craig Glennie [4]

[1] Civil Engineering Faculty, University of Houston, Houston, Texas – nekhtari@uh.edu
[2] Associate technical director, Cold Regions Research and Engineering Lab, Hanover, New Hampshire - Andrew.Niccolai@erdc.dren.mil
[3] Science, engineering, and technical advisor, Cold Regions Research and Engineering Lab - kevinslocum@geospatialscience.net
[4] Professor of Civil Engineering, University of Houston, Houston, Texas – clglenni@Central.UH.EDU

Keywords: Disaster Mapping, Geoparsing, Toponym Detection, Natural Language Processing, Encoder-Decoder.

Abstract

Real-time mapping of social media posts from users in the affected areas during a natural disaster can generate actionable intelligence for disaster response teams. Given the complexities of natural language used in such posts, modern approaches rely on Artificial Intelligence (AI) to improve accuracy. Extracting actionable geospatial intelligence from unstructured text requires a robust geoparsing pipeline comprising toponym detection and toponym resolution. While general-purpose Large Language Models (LLMs) can be utilized for toponym detection, their operational utility in high-volume, real-time workflows is constrained by high computational costs and a heavy reliance on intensive prompt engineering. To address these limitations, this study presents a highly efficient alternative utilizing custom encoder-decoder models fine-tuned specifically for toponym detection. Leveraging a dataset of 7,400 curated tweets from the 2024 hurricane Helene and another dataset of 50,000 tweets from the 2017 hurricane Harvey, we fine-tuned Google Research’s Flan-T5-base architecture twice. Both finetuned versions of the model demonstrated relatively good robustness, converging to F1 scores of 0.87 and 0.83 for hurricanes Helene and Harvey respectively. For the subsequent toponym resolution stage, we implemented a hybrid pipeline that categorizes extractions into four GIS-aligned classes. Regional-scale entities are matched against authoritative local GIS feature layers using a fuzzy string matching approach to map polygons, while localized features are resolved to point coordinates via GoogleMaps Geocoding API. The results of mapped tweets are examined as dynamic heatmaps that prove to be valuable in generating geospatial intelligence for disaster management.

1. Introduction

Recent advances in Artificial Intelligence (AI) have catalyzed renewed interest in the geospatial applications of Natural Language Processing (NLP) within the GIScience community, particularly regarding the near-real-time mapping of natural disasters via social media streams. Transforming unstructured text into actionable geospatial intelligence necessitates a modular geoparsing pipeline consisting of toponym detection and toponym resolution. While geoparsing has been studied for over a decade, the increasing volume and velocity of user-generated content during disaster events introduce significant challenges related to scale, noise, and temporal relevance.

Toponym detection, a specialized form of Named Entity Recognition (NER), has historically evolved alongside broader NLP methodologies. Early approaches relied heavily on gazetteer-based matching, which is computationally efficient but fundamentally limited by lexical ambiguity (e.g., distinguishing "Paris" as a person versus a place). To mitigate these limitations, statistical sequence models such as Conditional Random Fields (CRFs) became widely adopted due to their ability to model contextual dependencies (Patil et al., 2020). Subsequent work in the geospatial domain extended these methods to social media contexts, yet these models often struggle with the informal syntax and linguistic noise typical of microblogs (Sagcan and Karagoz, 2015).

The introduction of Transformer-based architectures marked a paradigm shift in detection performance. Encoder-based models such as BERT leverage bidirectional context and pre-training on large corpora to achieve substantial gains in entity recognition. Domain-specific adaptations, such as TopoBERT (Zhou et al., 2023), have demonstrated strong performance by fine-tuning on geospatial datasets. However, these approaches remain fundamentally token-classification frameworks, which can exhibit reduced robustness when applied to highly noisy, event-specific data like disaster-related social media posts. They also require a token alignment step in which labelled tokens have to be reconciled to word-level tokens for toponyms.

More recently, decoder-only Large Language Models (LLMs), such as GPT-4 and Gemini, have emerged as state-of-the-art solutions for complex language understanding (Hu et al., 2024). Their capacity for zero-shot and few-shot learning enables the extraction of rich contextual information; however, their application to large-scale disaster mapping introduces practical constraints. These models are computationally expensive and exhibit high inference latency, often proving unnecessarily resource-intensive for the near-real-time requirements of operational GIS workflows.

This research positions itself between these two extremes by proposing a fine-tuned encoder–decoder architecture based on Flan-T5, designed for efficient and scalable toponym detection. Unlike encoder-only models, the text-to-text formulation of Flan-T5 frames toponym extraction as a structured generation task, simplifying the learning objective while retaining deep contextual understanding. Furthermore, the model remains lightweight enough for rapid inference and potential local deployment, effectively addressing the latency and cost barriers associated with proprietary LLM-based approaches.

Critically, our study departs from conventional NER formulations by focusing on a geospatially grounded taxonomy of toponyms consisting of four distinct classes: GPE (geopolitical entities), LOC (natural features), FAC (facilities), and ADD (addresses/intersections). This distinction is vital for GIS workflows, where the entity type directly determines the geometric representation—such as polygons for administrative regions versus point features for granular locations—and influences subsequent spatial analysis. By explicitly modelling these categories and fine-tuning on disaster-specific data, our approach aligns more closely with the downstream requirements of situational awareness and disaster response systems.

By leveraging the contextual strengths of modern Transformers while prioritizing computational efficiency, our method provides a practical and scalable solution for real-time disaster mapping that bridges the gap between classical sequence models and large-scale generative AI.
Our contributions in this paper are as follows:

1. A task-specific reformulation of toponym detection as a sequence-to-sequence generation problem.
2. A four-class geospatial NER schema aligned with GIS workflows.
3. A scalable alternative to LLM-based geoparsing with competitive accuracy and significantly lower latency.

2. Datasets and Pre-processing

This research leverages two primary social media datasets focused on major natural disaster events: Hurricane Helene (2024) and Hurricane Harvey (2017). The Helene dataset was collected directly via the X API and served as the foundation for the initial fine-tuning of our encoder–decoder models, with toponyms manually validated and labelled. To further train and evaluate our best-performing model, we utilized the Hurricane Harvey dataset obtained from the University of North Texas, which was originally collected via the X API for academic research purposes. Both datasets were subjected to a rigorous multi-stage filtering and pre-processing pipeline to ensure linguistic consistency and temporal relevance.

2.1 Data Collection and Filtering

The Hurricane Helene dataset was curated through continuous X API monitoring throughout 2024. After removing duplicates (primarily retweets), we identified approximately 7,400 unique posts. Of these, 5,400 tweets were found to contain at least one toponym, which were subsequently manually labelled for training and validation.
The Hurricane Harvey dataset served as a secondary benchmark for evaluating model robustness and cross-event generalization. From an initial repository of approximately 7 million tweets, we filtered for unique content by cross-referencing retweet status, timestamps, and extended text fields, resulting in approximately 1.13 million unique posts. Temporal filtering was then applied to isolate tweets from August 24 to September 12, 2017, yielding approximately 1.08 million event-specific tweets.
From this subset, 500,000 tweets were processed using the Gemini 2.5 API to automatically extract toponyms. These LLM-generated annotations were used as proxy gold labels to enable large-scale experimentation, acknowledging that they may contain errors and should not be considered true ground truth (gold label). From the LLM-annotated dataset, we randomly sampled 50,000 tweets for this study. A subset of 10,000 tweets was used for model fine-tuning, while the remaining tweets were used for evaluation and cross-model comparison. Future work will extend this approach to the full dataset.
Timestamps for both datasets were normalized to local time zones—U.S. Eastern Time for Hurricane Helene and U.S. Central Time for Hurricane Harvey—to facilitate accurate temporal analysis of disaster progression.

2.2 Data Pre-processing Strategy

Short social media posts often contain significant linguistic noise, including hashtags, user mentions, emojis, and URLs, which can hinder the learning process of encoder–decoder models. While hashtags and mentions may carry semantic or toponymic information, emojis and URLs are typically non-informative for this task and can unnecessarily increase input complexity.
Our pre-processing pipeline was designed to preserve grammatical structure while reducing noise. Hashtags were converted to plain text by removing the “#” prefix; user mentions were anonymized using a generic [USER] token; URLs were replaced with a [LINK] token; and emojis were removed. Additionally, we applied language filtering and retained only English-language tweets.

3. Methodology

Our approach is based on geoparsing, the process of extracting place names (toponyms) from unstructured text and converting them into mappable geographic representations. Geoparsing is typically formulated as a two-stage pipeline consisting of (1) toponym detection and (2) toponym resolution. In this work, we implement a custom pipeline in which a fine-tuned encoder–decoder model is used for toponym detection, followed by a hybrid GIS-based and API-based resolution strategy. Our main contribution lies in the toponym detection phase, where instead of using on encoder-only models (such as TopoBERT) or decoder-only models (such as Google Gemini LLMs), we utilized an encoder-decoder model. These models have less complexities and parameter compared to LLMs and can be finetuned for specific tasks. The decoder part of the model is finetuned in our work. The training data effectively nudge the model towards generating a structured list as output that includes toponyms mentioned in the input. The next two subsections detail our approach.

3.1 Toponym Detection

We adopt the Flan-T5 model (Fine-tuned Language Net Text-to-Text Transfer Transformer), an instruction-tuned encoder–decoder Transformer architecture derived from the original T5 framework (Raffel et al., 2023; Chung et al., 2022). T5 models are designed under a unified text-to-text paradigm, where both inputs and outputs are represented as natural language sequences. Flan-T5 extends this framework through large-scale instruction fine-tuning on a diverse mixture of tasks, including classification, question answering, and information extraction, enabling improved generalization and stronger zero-shot and few-shot capabilities.

Architecturally, Flan-T5 consists of a Transformer encoder–decoder network that leverages both self-attention and cross-attention mechanisms. The encoder processes the input text using self-attention to capture contextual relationships among tokens, while the decoder generates outputs by attending both to previously generated tokens (via self-attention) and to encoded input representations (via cross-attention). This dual-attention structure allows the model to effectively align input sequences with structured outputs.

This architecture is particularly well-suited for toponym detection, which can be formulated as a structured sequence generation task rather than a token-level classification problem. The encoder captures global contextual cues necessary for resolving ambiguity in short and noisy text, while the decoder—through cross-attention—can selectively focus on relevant parts of the input when generating toponyms. This is especially beneficial in disaster-related social media posts, where identifying location entities often depends on subtle contextual signals distributed across the sentence. Additionally, the

instruction-tuned nature of Flan-T5 enables the use of lightweight prompts to guide the model toward producing outputs aligned with domain-specific entity categories.

Toponym detection is treated as a sequence-to-sequence generation task using a fine-tuned encoder–decoder architecture. Rather than performing token-level classification (as encoder-only models would), our encoder-decoder model is trained to generate a structured list of toponyms directly from the input text. We trained two models, T5-small and Flan-T5-base, on the Hurricane Helene dataset. As Flan-T5 consistently achieved superior performance, only its results are presented.

During fine-tuning, token embeddings were frozen while the remaining model parameters were updated to adapt to the domain-specific characteristics of disaster-related social media text. To optimize our text-to-text formulation without incurring the massive computational footprint of full-parameter tuning, we integrated Low-Rank Adaptation (LoRA) into the Flan-T5-base architecture. LoRA injects trainable rank decomposition matrices into the Transformer attention blocks, specifically targeting the query and value (W_q, W_v) projection layers of both the self-attention and cross-attention sub-layers. By freezing the original pre-trained weight matrices and only optimizing these low-rank matrices, we drastically reduced the number of trainable parameters by over 90%, preventing catastrophic forgetting while preserving the base model’s generalized semantic understanding.

In our fine-tuning data and the small appended prompt, we lead the model to use 4 predefined and prefixed categories of toponyms. By definition, our toponym categories are tailored to our desired geospatial analysis:

1. GPE (Geopolitical Entities): Larger administrative regions such as states, counties, and cities that have defined boundaries.
2. LOC (Natural Features): Bodies of water, parks, or mountains (e.g., Buffalo Bayou, Lake Conroe, Memorial Hermann Park).
3. FAC (Facilities): Man-made structures including airports, hospitals, or specific buildings (e.g., IAH Airport) that are mentioned by name.
4. ADD (Addresses/Intersections): Granular location data including street names, street addresses, and road intersections (e.g., Main St & Preston St).

Given the limited context window of the model (256 tokens), we appended a short instructional prompt to each input tweet during both training and inference. The prompt used in this study is: Extract toponyms (GPE: geopolitical entities, LOC: locations, FAC: facilities, ADD: addresses) from this tweet. This formulation enables the model to leverage contextual understanding while producing structured outputs aligned with geospatial analysis requirements.

3.2 Toponym Resolution

Toponym resolution converts extracted place names into geographic representations, either as coordinates (for fine-grained locations) or as polygons (for administrative regions). To achieve this, we employ a hybrid resolution strategy combining local GIS resources and external geocoding services.

Step 1: Local GIS Matching

Toponyms belonging to the GPE and LOC categories are matched against authoritative local GIS layers using fuzzy string matching. These entities are represented as polygons to reflect their spatial extent. In the case of Hurricane Harvey, the majority of detected toponyms correspond to Texas and Harris County, making local GIS layers particularly effective for resolving regional entities.

Step 2: API-Based Geocoding

Remaining entities, including FAC, ADD, and city-level GPE toponyms, are resolved to geographic coordinates using the GoogleMaps Geocoding API. This approach enables high-precision localization of fine-grained features such as addresses and facilities. To mitigate the high monetary cost associated with external API calls, our pipeline caches resolved queries locally to prevent duplicate lookups of frequently mentioned shelters or intersections. Furthermore, the API’s component filtering capabilities allow results to be biased toward the disaster-affected region (e.g., the Texas Gulf Coast), reducing ambiguity and Geocoding API costs as well as improving accuracy.

4. Results

Our fine-tuned Flan-T5 model achieved a micro-averaged F1 score of 0.87 for toponym detection on the Hurricane Helene dataset. It is noteworthy that the majority of detected toponyms in this dataset belong to the GPE category, reflecting the tendency of social media users to reference broader administrative regions during disaster events.

To evaluate the model’s generalizability, we further fine-tuned and tested it on the Hurricane Harvey dataset. A total of 50,000 tweets were used in this study, of which 10,000 were used for training and the remaining 40,000 for evaluation. The evaluation was conducted against LLM-assisted annotations, and the model achieved an F1 score of 0.83. Given the scale and noisiness of the dataset, this result demonstrates robust cross-event performance.

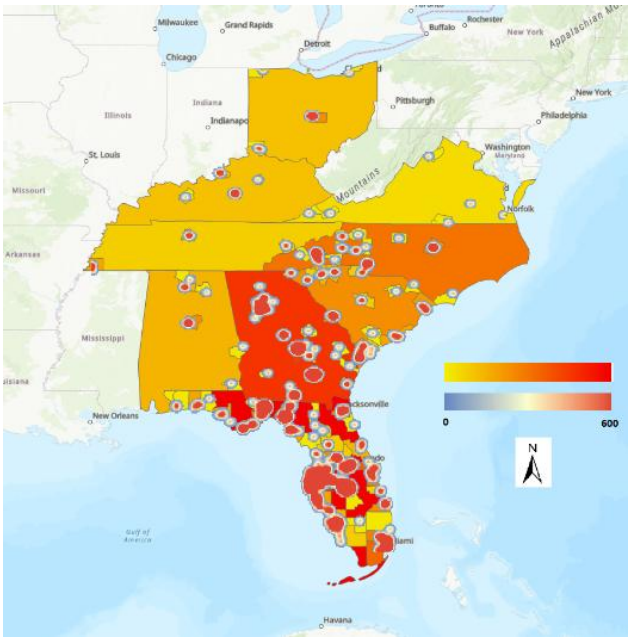

Figure 1- Heatmap generated from all downloaded tweets about hurricane Helene in a 48-hour window. Top color scale is used for polygon toponyms and the bottom color scale is used for clusters of point-based toponyms.

Beyond quantitative evaluation, we assessed the utility of the extracted toponyms for geospatial analysis. Approximately half of the tweets in the Hurricane Harvey dataset did not contain identifiable toponyms. Among those that did, the majority corresponded to GPE entities (e.g., Texas, Houston, Harris County), with relatively fewer fine-grained references (FAC and ADD).

Using the timestamps associated with each tweet, we generated temporal heatmaps to visualize the spatial progression of disaster-related activity. GPE and LOC entities were represented as polygons, while FAC and ADD entities were mapped as point features. Point-based features were visualized using kernel density estimation (KDE), while polygon-based features were aggregated by counting mentions within each time interval. Kernel density estimation provides an informative raster representation of point data. Normalization can be achieved through weighting schemes (based on retweet counts for example.)

Figure 1 illustrates the cumulative heatmap for Hurricane Helene, based on a 48-hour data collection window. By combining polygon-based representations of administrative regions with point-based representations of named granular toponyms, the resulting visualization provides a high-level overview of affected areas. It can be seen that states of Georgia and Florida have been heavily mentioned in the relatively small number of tweets we had for this experiment. This matches the impact zone of hurricane Helene specifically during the 48-hour window that our data was collected. At the regional-level representation of figure 1, point-type toponyms are shown as distinct clusters overlaid on state-level toponyms represented by color-symbolized polygons. We intentionally used two different color scales to distinguish the point-based toponyms.

For Hurricane Harvey, we only focused on mapping point-based toponyms (named facilities and addresses) due to the higher data volume and smaller geographic extent of the study area. This approach resulted in clearer visualization of localized impacts, as shown in Figure 2. In this figure, the final heatmap is shown that showcases the aggregate tweets from entire 21-day temporal window of this study. The color scale is between 0 and 50, with 50 being set as maximum for better visualization purposes. This kernel density estimated raster is weighted by retweet counts (which play a role into the importance of social media posts.) Cell sizes are set to 500 meters and the kernel radius is set to 2 kilometers.

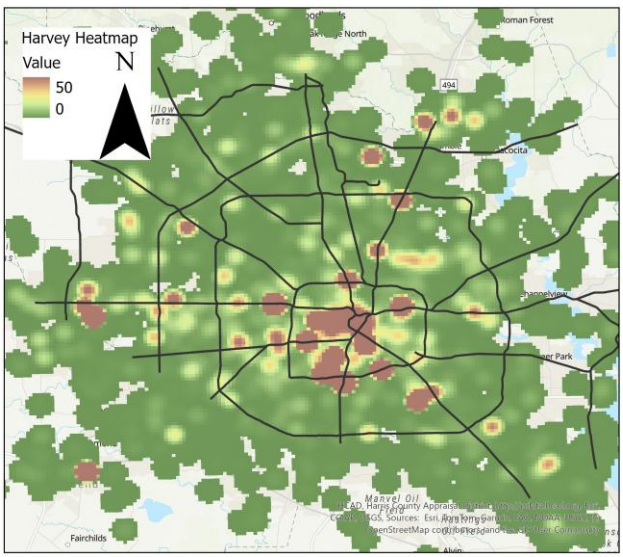

Figure 2 - Final heatmap showing concentration and extent of tweets within the entire 21 days. The raster map is a kernel density estimation of geocoded tweets that is weighted by retweet counts. Note that the symbology is capped at 50. The heatmap is overlaid by Houston major highways layer shown in black.

5. Discussion and Conclusion

5.1 Intrinsic Challenges in Social Media Data for Geoparsing

The efficacy of disaster-mapping systems built upon micro-blogging platforms such as X (formerly known as twitter) is fundamentally constrained by the chaotic and informal nature of the data. Social media posts, particularly from the period when Hurricane Harvey occurred (with a 140-character limit), are inherently brief, often resembling fragmented headlines rather than coherent sentences, and are usually used to share external content (links to videos, pictures, news websites, etc.). While the limit was later increased to 280 characters for the Hurricane Helene dataset, the brevity and lack of perfect sentence structure persist, making them challenging inputs for traditional NLP models. Even sophisticated LLMs such as Gemini 2.5 Flash were not always successful in detecting all the toponyms.

Another major limiting factor is the high level of linguistic noise. For instance, some users’ overuse of emojis in their tweeting styles significantly obscured our model's semantic understanding, given our stringent pre-processing pipeline which benefits more lingual tweets. We observed that such stylistic patterns degrade model performance across both the large language model and our smaller fine-tuned encoder-decoder models, particularly when textual cues were limited or ambiguous. Both an LLM and our small fine-tuned model are able to detect toponyms with significantly higher accuracy on longer, more structurally sound, and grammatically coherent texts.

Furthermore, the majority of posts in our dataset (containing a disaster related hashtag) were non-informative from a geolocating/geocoding perspective. More than half of the tweets either do not mention or are not about a specific geographic location. The overwhelming majority of tweets with toponyms simply mention large geopolitical entities (states, counties, major cities) as opposed to specific locations that can be mapped with point features. This limits the granularity of the resulting geospatial intelligence. For instance, from the entire 500,000 tweets of hurricane Harvey dataset that we labelled with Gemini 2.5, only about 5,000 tweets were detected that contained at least one ADD toponym, and only about 12,000 tweets were detected that contained at least one FAC toponym. While this number is about 25,000 for LOC, the total number of tweets with at least one detected GPE toponym is over 203,000. With such large differences in number of toponyms, it is apparent that social media posts can be used for larger scale disaster mapping (to map the toponyms at state, county, and city levels) rather than a more granular small-scale approach. However, our experiments show the potential of NLP approaches in creating GIS-based disaster monitoring systems. This limitation highlights an inherent constraint of user-generated data, where fine-grained geographic references are relatively sparse compared to broader geopolitical mentions.

Finally, we observed a substantial portion of content consisting of shared links, photos, or videos with minimal descriptive text, rendering them ineffective for purely text-based geoparsing approaches. This critical limitation suggests that achieving comprehensive coverage requires a future shift toward a multimodal system capable of leveraging visual and other non-textual media to determine location.

5.2 Comparative Analysis of Model Architectures

Our research provides a direct operational comparison between the resource-intensive, high-fidelity, decoder-only approach of proprietary Large Language Models (LLMs) and our small, optimized, and fine-tuned encoder-decoder architecture. It should be noted that while LLMs such as Gemini 2.5 Flash are capable of reasoning and therefore are expected to demonstrate human-like performance, the nuanced nature of short tweets and the writing style of various social media users present a challenge for these models. A manual examination of 100 labeled tweets by our team revealed that about 10 toponyms were not correctly detected. Therefore, it should be noted that LLM-generated toponyms cannot entirely replace manually labelled tweets yet. However, for the purposes of this study, we decided to treat LLM-generated labels as gold labels (ground truth) because manually labelling 50,000 tweets for hurricane Harvey was not feasible for our team. Therefore, our proposed fine-tuned Flan-T5 model for hurricane Harvey is trained and evaluated against LLM-generated gold labels. This can explain why the F1 score associated with the hurricane Harvey dataset shows a drop of 0.04.

5.2.1 The Operational Cost and Latency of LLMs

While generative LLMs, such as Gemini 2.5-flash, offer high potential for contextual understanding, their deployment for large-scale, near-real-time disaster mapping is severely hampered by computational and financial costs. Utilizing cloud-based LLM APIs involves a paid subscription model with usage billed per one million tokens (input and output), making high-volume inference both slow and expensive. Our extensive tests demonstrated that to maintain consistent performance and avoid rate-limiting errors, we had to adhere to highly constrained batch sizes and parallel request limits. Specifically, achieving reliable results necessitated a batch size of 50 tweets per request, with a maximum of five parallel batches sent per step to the Gemini API. Processing one million tweets under these optimized parameters still incurred approximately 20 hours of total processing time, averaging approximately 13–15 requests per minute. These parameters were determined empirically through iterative testing, balancing throughput and API rate limitations. Increasing batch size beyond 50 tweets per request resulted in higher latency per request and more frequent rate-limiting, while increasing the number of parallel batches beyond five led to unstable performance and failed requests. Conversely, smaller batch sizes reduced overall throughput. These observations highlight the sensitivity of LLM-based pipelines to batching strategies and suggest that suboptimal configurations can significantly increase total processing time.

Conversely, deploying smaller LLMs locally can be challenging due to hardware and energy requirements. Moreover, their small context windows restrict effective prompt engineering, batching, and detailed instruction sets.

In contrast, our fine-tuned model can be deployed on modest hardware (e.g., a single GPU or high-performance CPU), eliminating the need for large-scale cloud-based infrastructure and reducing operational costs. A larger encoder-decoder model could potentially perform better as apparent between our overall F1-score gain when switching from T5 to Flan-T5 model.

5.2.2 The Efficiency and Semantic Trade-Off of Fine-Tuned Models

Our proposed fine-tuned Flan-T5 model offers a stark contrast in efficiency. The model requires only about 15 hours for fine-tuning (using four epochs, a batch size of 80, and Low Rank Adaptation), and its subsequent inference process is exceptionally fast, processing 800 tweets in less than one minute (which is almost identical to the rate we experienced with our subscription tier to Gemini API).

However, this efficiency comes with an acknowledged trade-off in deep contextual relevance. While encoder-only and standard encoder-decoder models achieve high F1 scores in general toponym detection, they often lack deeper contextual reasoning capabilities. For example, in the sentence, "State lawmakers met in Austin to vote on a relief package for floods in Pearland," our system correctly detects both "Austin" and "Pearland" as toponyms. Yet, for disaster mapping, the tweet's core relevance is to Pearland, not Austin. Generative LLMs, through long and detailed instructional prompts, can be directed to filter and extract only disaster-relevant toponyms. While this is computationally expensive, it boosts the accuracy of the mapped entity. A potential direction to address this limitation is a hybrid or ensemble framework, where the encoder–decoder model performs rapid toponym detection, followed by a lightweight secondary model that ranks or filters entities based on contextual relevance.

It is important to note that the evaluation of the Hurricane Harvey dataset relied on toponyms generated by a Large Language Model (Gemini 2.5) as proxy reference labels. While this approach enabled large-scale experimentation that would otherwise be infeasible due to the high cost and time requirements of manual annotation, it introduces an inherent limitation. LLM-generated labels should not be considered ground truth, as these models can produce errors in toponym detection, particularly in short, noisy social media text. Consequently, the reported performance metrics for the Harvey dataset should be interpreted as relative to LLM-based annotations rather than absolute measures of accuracy. This highlights the need for future work incorporating human-validated benchmarks or hybrid annotation strategies.

Additionally, fine-tuned encoder–decoder models offer greater flexibility for domain adaptation compared to general-purpose LLMs. Once trained, these models can be efficiently re-deployed or further fine-tuned on new disaster events or geographic regions with relatively low computational overhead. In contrast, adapting LLM-based pipelines often requires repeated prompt engineering or continued reliance on external APIs, which may limit scalability and reproducibility.

5.3 Conclusion

The fine-tuned Flan-T5 model demonstrated strong performance, achieving an F1 score of 0.87 on the Hurricane Helene dataset and 0.83 on the larger Hurricane Harvey dataset, relative to LLM-based benchmarks. Notably, the model maintained robust performance when transferred across disaster events, indicating its ability to generalize beyond the initial training domain. While this approach entails a modest trade-off in accuracy, it offers substantial gains in computational efficiency, scalability, and inference speed.

It is important to contextualize these results within the evaluation framework. For the Hurricane Harvey dataset, Large Language Model (LLM)-generated toponyms were used as proxy reference labels due to the prohibitive cost of manually annotating large-scale social media data. Although this enabled comprehensive evaluation, LLM outputs are not error-free and

may introduce noise, particularly in short and linguistically irregular text. As such, the reported performance should be interpreted relative to these proxy labels rather than as an absolute measure of ground-truth accuracy.

Despite this limitation, the results indicate that fine-tuned encoder–decoder architectures provide a practical and scalable alternative to large-scale LLMs for operational geoparsing tasks. By explicitly modelling geospatially relevant entity categories and leveraging a text-to-text formulation, the proposed approach aligns closely with the requirements of GIS-based disaster response systems. In particular, it enables rapid, near-real-time extraction of geospatial information from high-volume social media streams, making it well-suited for time-critical disaster monitoring and situational awareness applications.

6. Future Works

Future work will focus on improving model performance and generalizability through targeted data augmentation and expanded evaluation. In particular, we plan to leverage synthetic data generation using Large Language Models to better represent underrepresented toponym categories (ADD and FAC), and to scale training and evaluation to the full Hurricane Harvey dataset.

We also aim to explore alternative pre-processing strategies, such as converting emojis into textual representations to preserve contextual information in short social media posts.

Instruct-tuning a small locally deployable LLM is another approach that is planned to be tested in the future. The advantage of such model over Flan-T5 would be that a larger more detailed prompt can be used along with the batched inputs.

Another key direction involves developing a multimodal geoparsing framework that incorporates visual and other non-textual data accompanying social media posts. Integrating multiple data modalities, as well as extending the framework to other disaster types such as wildfires and earthquakes, will further improve the robustness and applicability of the proposed approach.

Acknowledgements

This work is supported via a research grant by Cold Regions Research Laboratory within the Engineering Research and Development Center of the US Army Corps on Engineering. Efforts and critics of other collaborators Ovi Paul and William Forker is greatly appreciated.

References

Chung, H. W., et al. 2022. Scaling Instruction-Finetuned Language Models, Journal of Machine Learning Research 25 (2024) 1-53.

Hu, X., Kersten, J., Klan, F., & Farzana, S. M. (2024). Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge. International Journal of Geographical Information Science, 40(3), 670–697. https://doi.org/10.1080/13658816.2024.2405182

Patil, N., Patil, A., Pawar, B.V., 2020. Named Entity Recognition using Conditional Random Fields, Procedia Computer Science, Volume 167, Pages 1181-1188, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.03.431.

Phillips, Mark Edward 2025. Hurricane Harvey Twitter Dataset, dataset, 2017-08-18/2017-09-22; https://digital.library.unt.edu/ark:/67531/metadc993940/: accessed October 31, 2025), University of North Texas Libraries, Digital Library, digital.library.unt.edu

Raffel, C., et al. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer, Journal of Machine Learning Research 21 (2020) 1-67.

Sagcan, M., Karagoz, P., 2015. Toponym Recognition in Social Media for Estimating the Location of Events, IEEE International Conference on Data Mining Workshops, https://doi.org/10.1109/ICDMW.2015.167

Xuke Hu, Jens Kersten, Friederike Klan & Sheikh Mastura Farzana, (2024), Toponym resolution leveraging lightweight and open-source large language models and geo-knowledge, International Journal of Geographical Information Science, Volume 40, issue 3

Zhou, B., Zou, L., Hu, Y., Qiang, Y., Goldberg, D., 2023. TopoBERT: a plug and play toponym recognition moduleharnessing fine-tuned BERT, International Journal of Digital Earth, Volume 16, No. 1, 3045 - 3064, https://doi.org/10.1080/17538947.2023.2239794