LEVERAGING LARGE LANGUAGE MODELS FOR CRISIS DETECTION AND RESPONSE (CALIFORNIA WILDFIRE)

Pair - 10

Hamsalakshmi Ramachandran - 017423666

Sugandha Chauhan - 017506190

PROBLEM DEFINITION

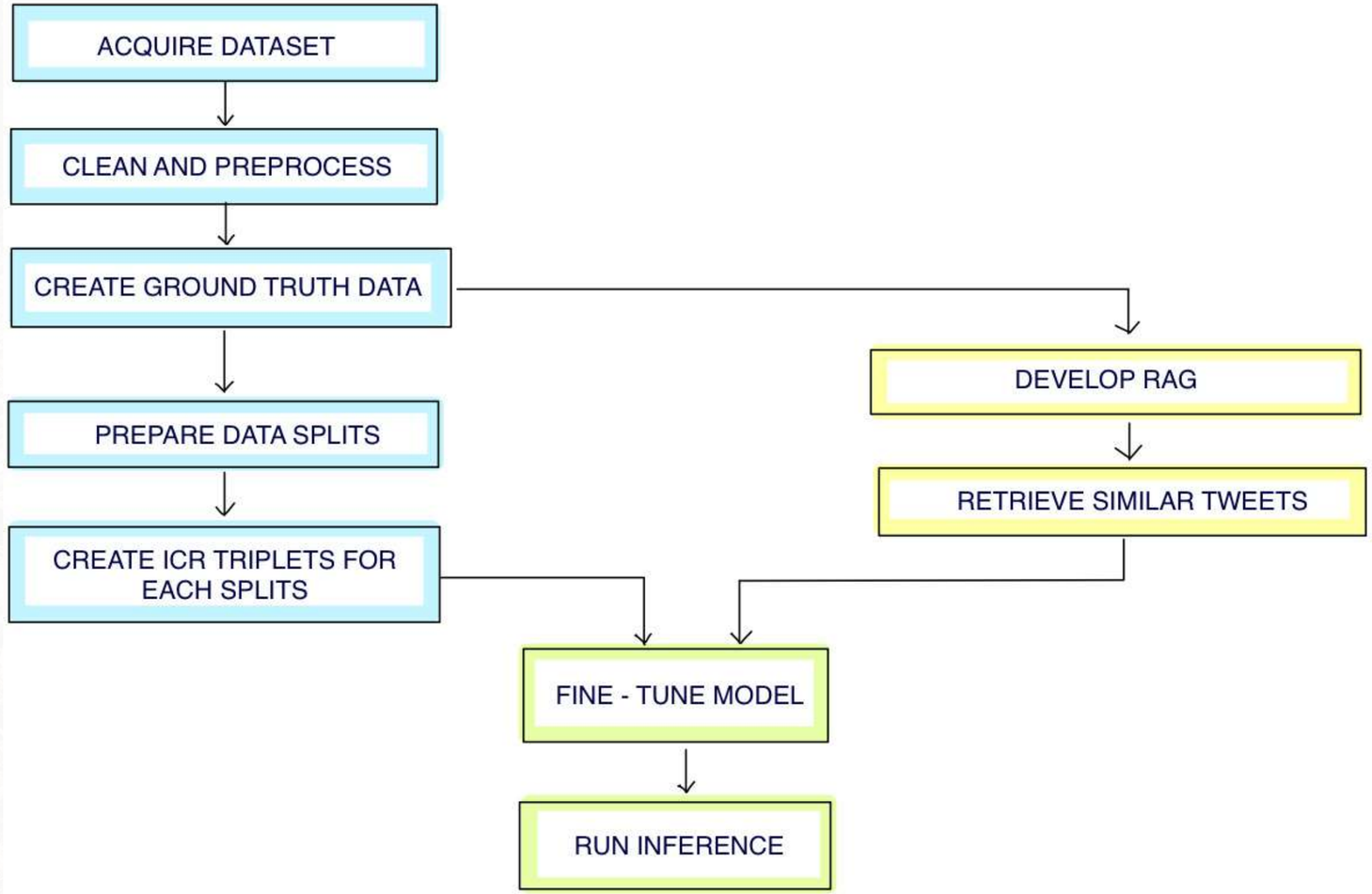
- In January 2025, catastrophic wildfires struck Los Angeles, particularly the Palisades and Eaton Canyon fires. These incidents destroyed more than 9,400 buildings and necessitated the evacuation of over 30,000 residents. Conventional emergency response systems face challenges in processing vast amounts of data from various sources, which can result in delayed decision-making.

- The LLM-Supported Crisis Management System utilizes Large Language Models (LLMs) to improve wildfire response and emergency decision-making.

- The system handles crisis information for precise, prompt, and organized reporting.

- The goal of this project is to enhance situational awareness, response effectiveness, and information precision for emergency teams managing wildfire events.

METHODOLOGY



DATASET DESCRIPTION

- The dataset consists of 16058 tweet texts and 18082 and images (multimodal).
- Annotated for informativeness, humanitarian categories, and damage severity.
- Supports multitask learning for disaster response and situational awareness.
- Used to train AI models for real-time crisis detection and decision support.
- Developed by Qatar Computing Research Institute (QCRI) to aid humanitarian efforts.
- https://crisisnlp.qcri.org/crisismmd

```
Disaster Type Distribution:
  - earthquake: 1504 (10.4%)
  - hurricane: 10029 (69.5%)
  - other: 1371 (9.5%)
  - flood: 679 (4.7%)
  - wildfire: 842 (5.8%)

Distress Distribution:
  - not distress: 12844 (89.0%)
  - distress: 1581 (11.0%)
```

Informativeness		
	Text	Image
Informative	11509	9374
Not informative	4549	8708
Total	**16058**	**18082**
Humanitarian		
Affected individuals	472	562
Infrastructure and utility damage	1210	3624
Injured or dead people	486	110
Missing or found people	40	14
Not humanitarian	4549	8708
Other relevant information	5954	2529
Rescue volunteering or donation effort	3293	2231
Vehicle damage	54	304
Total	**16058**	**18082**
Damage Severity		
Little or no damage	-	475
Mild damage	-	839
Severe damage	-	2212
Total	-	**3,526**

PREPROCESSING

- Convert tweet text to lower case
- Remove missing value
- Emoji & Unicode Symbol Removal
- Hashtag Symbol Removal
- Special Character Filtering

```
Text preprocessing complete.
                                                tweet_text
0  RT @MSN: Island of Barbuda 'literally under wa...
1  RT @Reuters: Hurricane Irma threatens luxury T...
2  RT @TheAnonJournal: BREAKING NEWS: Hurricane I...
3  JUST IN: 11PM #Hurricane #Irma update. @ABC7Ne...
4  RT @cnnbrk: Hurricane Irma destroys "upwards o...

                                               cleaned_text
0   rt island barbuda literally water hurricane irma
1  rt hurricane irma threatens luxury trump property
2  rt breaking news hurricane irma big enough cov...
3                    pm hurricane irma update weather
4  rt hurricane irma destroys upwards barbuda off...
```

```
--- Missing Values per Column --
tweet_id                    0
image_id                    0
text_info                   0
text_info_conf              0
image_info                  0
image_info_conf             0
text_human                  0
text_human_conf             0
image_human                 0
image_human_conf            0
image_damage            14455
image_damage_conf       14455
tweet_text                  0
image_url                   0
image_path                  0
crisis_type                 0
is_california_fire          0
has_image                   0
dtype: int64
```

```
Shape of DataFrame before dropping: (18082, 18)
Shape of DataFrame after dropping columns ['image_damage', 'image_damage_conf']: (18082, 16)
```

GROUND TRUTH

- We have used CrisisMMD dataset that has ~18000 tweets and images.
- Ground Truth data was created using Rule Based logics and Manual Screening.
- It has the following columns:

  tweet_id, image_id, raw_tweets, cleaned_tweets, hashtags, image captions, distress signal, take_action, state, sub_location, disaster type.

tweet_id	image_id	raw_tweet_text	tweet_text	tweet_hashtags	image_caption	distress	take_action	state	sub_location	disaster_type
917793158251077000	917793158251077632_0.jpg	RT @FoxNews: Southern California fire shrouds Disneyland Anaheim in dramatic, smoky skies https://t.co/nNxpY2kSmS https://t.co/6ripQFkIVh	southern california fire shrouds disneyland anaheim in dramatic smoky skies		mickey mouse pumpkin with a flag on top of it	0		California	southern disneyland anaheim	other
917794580728295000	917794580728295424_0.jpg	More than 100 missing persons reports made in CaliforniaÃ, wildfires https://t.co/614UWl30MO https://t.co/CBuNhzQ5xf	more than 100 missing persons reports made in california wildfires		a fire is seen burning through the air near a road	1	start missing person search	California		wildfire
917792930315821000	917792930315821057_0.jpg	Mass Evacuations in California as Wildfires Kill at Least 10 https://t.co/gyoKFWZuMB #CaliforniaWildfires https://t.co/KEFtjITetK	mass evacuations in california as wildfires kill at least 10 #californiawildfires	californiawildfires	a fire is seen in the distance behind trees and a building	1	send evacuation and shelter support	California	wildfires	wildfire

ICR TRIPLETS INPUT

Split the dataset into 4 splits

- First took out a sample of 50 wellbalanced cases for final evaluation using RAG.
- Then split rest of the data into train, validate and test set.
- Each of these split were converted to the ICR triplet format for instruction fine tuning.

```
### INSTRUCTION:
Classify this emergency tweet using the provided context.

You must return the result in the following JSON format with these exact keys:
- disaster_type: one of ["wildfire", "earthquake", "flood", "hurricane", "other"]
- distress: either "distress" or "not distress"
- location_info: extract the most specific location from the tweet (state and sub-location), or use "unknown"
- action: one of ["send immediate help", "send medical team", "send rescue team", "start missing person search", "send evacuati
elp and shelter support", "send security and trauma support"]

Return a valid JSON with all four fields.

### CONTEXT:
playing with my new friend chai shes a california fire evacuee starting with us for a few days i her

### RESPONSE:
{
    "disaster_type": "other",
    "distress": "not distress",
    "location_info": "unknown, california",
    "action": "unknown"
}
```

```
### INSTRUCTION:
Classify this emergency tweet using the provided context.

You must return the result in the following JSON format with these exact keys:
- disaster_type: one of ["wildfire", "earthquake", "flood", "hurricane", "other"]
- distress: either "distress" or "not distress"
- location_info: extract the most specific location from the tweet (state and sub-location), or use "unknown"
- action: one of ["send immediate help", "send medical team", "send rescue team", "start missing person search", "send evacuati
elp and shelter support", "send security and trauma support"]

Return a valid JSON with all four fields.

### CONTEXT:
sonoma and napa wineries damaged by california wildfires updated list

### RESPONSE:
{
    "disaster_type": "wildfire",
    "distress": "not distress",
    "location_info": "sonoma napa, california",
    "action": "unknown"
}
```

EXPERIMENTAL DESIGN

Experimental Design for Fine Tuning

- **Library:** HuggingFace Transformers library
- **Dataset:** custom instruction-context-response (ICR) dataset and ground truth dataset.
- **Model:** meta-llama/Llama-2-7b-chat-hf with 4-bit quantization
- **Fine Tuning Method:** Supervised Fine-Tuning (SFT) using Parameter-Efficient Fine-Tuning (PEFT) with LoRA.
- **Tokenizer:** HuggingFace AutoTokenizer
- **Trainer:** HuggingFace traine with gradient accumulated over 2 checkpoints.
- Mixed precision (fp16) is enabled for training.

Experimental Design for RAG inplementation

- **Sentence Transformer:** all-mpnet-base-v2
- FAISS index is built on top of the embeddings using L2 distance
- top-5 most similar documents (top k=5) for each tweet query based on semantic similarity.
- The fine-tuned LLaMA 2 model is loaded with LoRA adapters.

EXPERIMENTAL RESULTS - FINE TUNED MODEL

Metrics	Disaster_Type	Distress_detected	Location	Recommended_act
Accuracy	0.9856	0.9911	0.7988	0.9873
Macro Avg Precision	0.81	0.98	0.35	0.75
Macro Avg Recall	0.81	0.97	0.35	0.76
Macro Avg F-1 Score	0.81	0.98	0.34	0.75
Weighted Average Precision	0.99	0.99	0.79	0.99
Weighted Average Recall	0.99	0.99	0.80	0.99
Weighted Average F1-Score	0.99	0.99	0.79	0.99
Total Samples				1804

- Fine-tuning with LoRA drastically improved distress classification — from ~56% accuracy to over 99%.
- Fine-tuning enabled the model to align with domain-specific instructions and patterns in distress-related text
- This result validates the use of PEFT + instruction fine-tuning as critical for performance in real-world disaster response tasks.

Previous Base Model Results were :

Class	Recall	F1-Score
distress	0.5	0.48
not	0.6	0.62
accuracy		0.56

EXPERIMENTAL RESULTS - RAG SYSTEM + FINE-TUNED MODEL

- Disaster Type classification achieves an F1-score of 0.814 with balanced precision (0.849) and recall (0.820), indicating consistent detection across disaster categories.
- Distress Detection performs reliably with an F1-score of 0.8397 and accuracy of 0.84, confirming the model's effectiveness in identifying distress cues in tweet text.
- The system achieves an Average BLEU Score of 0.5818, reflecting good alignment in semantic structure between generated outputs and ground-truth references.

Distress detection	LLaMA-2 7B Base	LLaMA-2 Quantized	GPT-3.5 Turbo	Fine-Tuned LLaMA-2 7B LoRA
Accuracy	78%	46%	90%	84%
Recall	39%	46%	90%	84%
F-1 Score	79%	44%	90%	83.9%

Metrics	Fine-Tuned Llama-2 7B Lora
Disaster_Type	
Accuracy	0.820
Precision	0.849
Recall	0.820
F-1 Score	0.814
Distress_detected	
Accuracy	0.8400
Precision	0.8422
Recall	0.8400
F-1 Score	0.8397
Recommended_act	
Accuracy	0.8400
Precision	0.960
Recall	0.840
F-1 Score	0.8951
Average BLEU Score	0.5818

Tweet: 11 dead 100 injured wildfires ravage northern california with shocking speed

```
Direct analysis result:
{
  "disaster_type": "wildfire",
  "distress": "distress",
  "location_info": "northern california, california",
  "action": "send medical team"
}
✓ Correct disaster_type: wildfire
✓ Correct distress: distress
~ Partial match for location_info: predicted 'northern california, california', actual 'northern, california'
✓ Correct action: send medical team
```

Tweet: an inferno like youve never seen deadly wildfires ravage california via

```
Direct analysis result:
{
  "disaster_type": "wildfire",
  "distress": "not distress",
  "location_info": "california, california",
  "action": "send immediate help"
}
✓ Correct disaster_type: wildfire
✓ Correct distress: not distress
```

ACADEMIC BENCHMARKS

Study & Dataset	Model(s)	Task(s)	Accuracy	Citation
Performance evaluation of NLP & CNN models on CrisisMMD				
CrisisMMD (multimodal Twitter,	BERT-Base-Uncased, DistilBERT-Base, Twitter-RoBERTa,	Disaster-type classification	94%(language mo	(SpringerLink)
Leveraging LLMs for Enhanced Classification				
Wildfire tweets & reports	Instruction-tuned LLaMA-style LLM with LoRA	Fire vs. non-fire tweet classification	96.21% accuracy	(MDPI)
Multimodal Disaster-Tweet Classification				
CrisisMMD (text + image fusion)	Pretrained encoders + concatenation	Informative vs. non-informative & humanitarian categories	88% accuracy	(ACL Anthology)
Few-shot Disaster Tweet Classification				
Disaster-related tweets	DistilBERT (cross-entropy) vs. contrastive heads	Few-shot accuracy	~80.6%	(Stanford University)

INDUSTRY BENCHMARKS

Source	Focus	Reported Accuracy	Citation
Deloitte: AI in Emergency Management	Generative-AI frameworks for crisis response	Not reported	(Deloitte United States)
PrimerAI: Military AI for Wildfires	Predictive modeling & operational decision support	Not reported	(PrimerAI)
"CrisisMMD: Multimodal Twitter Datasets" (QCRI	Dataset creation & initial benchmarks	Informative detection: ~72%–82% (text only)[1]	(ResearchGate)

CONCLUSION

Our wildfire-detection system combines supervised fine-tuning on labeled fire reports, instruction fine-tuning for fluent and RAG indexing of historical data via FAISS to ground every answer in real-world evidence. This layered approach delivers decent results for precise location, severity, and action recommendations while minimizing hallucinations. Together, these methods form a modular, extensible pipeline ready for integration of new data streams and continual refinement.

SUGGESTED FUTURE WORK

- Could experiment with various prompting techniques to further improve the results.
- Try the system’s efficacy on real-time tweets and social media captions.
- Multimodal system development for analysing tweets and images as well.


Thank you!