Heaven's Light is Our Guide

Multimodal Transformer-Based Disaster Social Media Classification Using ViT and GPT-2 with Generative Summarization

Thesis Proposal

By

Nabila Ferdous
(2010035)

A Thesis Proposal submitted in partial fulfillment of the requirements for the Degree of Bachelor of Science

in

Electrical & Computer Engineering

Examination Committee: Prof. Dr. Md. Anwar Hossain (Head)
Moloy Kumar Ghosh (Supervisor)
Hafsa Binte Kibria (Supervisor)

Rajshahi University of Engineering & Technology
Faculty of Electrical & Computer Engineering
Department of Electrical & Computer Engineering
Rajshahi-6204

August 2025

1. Tentative Title: Multimodal Transformer-Based Disaster [1]Social Media Classification Using ViT and GPT-2 with Generative Summarization

2. Background and present state of the problem: Social media platforms like Twitter and Facebook have become crucial information sources during natural disasters. Millions of users post text, images, and videos in real time, which can help emergency agencies identify affected areas and prioritize rescue operations. However, a large proportion of these posts are redundant, irrelevant, or non-informative, making it difficult to extract meaningful insights efficiently.

Earlier studies provided annotated datasets for disaster-related tweet analysis[2]. Conventional models like CNNs and LSTMs[3] were used for classification tasks, but they struggled with contextual understanding and multimodal fusion. Recent transformer-based models—such as the Vision Transformer and GPT-2—achieved remarkable progress in image and text understanding[4].

A multimodal disaster response system using Vision Transformer and GPT-2 with Random Forest fusion improved classification accuracy significantly[5]. Despite this success, the model remained purely discriminative and lacked a generative capability to summarize key disaster information. Integrating generative AI can bridge this gap by producing concise disaster reports from informative posts, improving situational awareness for decision-makers.

3. Justification of the study: While transformer-based multimodal approaches have improved the classification of disaster-related tweets, current systems focus solely on labeling posts as informative or not informative. These models do not generate human-readable summaries or integrate outputs into actionable disaster reports. In real-world scenarios, emergency teams require summarized, structured information rather than isolated tweet labels.

Introducing a generative summarization module (e.g., Flan-T5) alongside ViT and GPT-2 can automatically create meaningful disaster summaries. This approach enhances interpretability, usability, and real-time response potential. By combining discriminative (classification) and generative (summarization) capabilities, this study addresses a key gap in existing multimodal disaster analytics. The research contributes toward a more human-centered, AI-assisted disaster response framework that transforms raw social media data into coherent, actionable insights.

4. Objective and Scope:

I. Objectives

1. To develop a multimodal transformer-based system combining **Vision Transformer (ViT)** and **GPT-2** for disaster tweet classification.
2. To integrate a **generative text summarization** model (Flan-T5) for producing concise disaster reports from informative tweets.
3. To evaluate the proposed model using metrics such as Accuracy, F1-Score, and ROUGE-L on the **CrisisMMD dataset**.
4. To visualize attention maps for model interpretability.

II. Scope

The study focuses on **text–image fusion** in social media posts during natural disasters (e.g., floods, earthquakes, wildfires). It uses publicly available datasets and pre-trained transformer models to ensure reproducibility and feasibility. The project's scope covers data preprocessing, model fine-tuning, multimodal fusion, generative summarization, evaluation, and visualization of outputs.

5. Points of Contributions:

- Implementation of a transformer-only multimodal pipeline using ViT for vision and GPT-2 for language understanding.
- Integration of a generative summarization module (Flan-T5) to convert classified tweets into coherent disaster summaries.
- Development of a fusion strategy combining visual and textual embeddings through Random Forest or transformer-based fusion.
- Quantitative comparison with existing CNN, BERT, and LSTM-based baselines.
- Contribution to the research on AI-assisted disaster management, offering a domain-specific generative model.
- Implementation on moderate hardware (RTX 3050 GPU) using efficient fine-tuning strategies, ensuring reproducibility for academic research.
- Providing a framework adaptable for future extensions such as multilingual analysis or multi-disaster classification.

6. Outline of Methodology:

Dataset: CrisisMMD tweet–image pairs labeled as informative or not, split into training (70%), validation (15%), and testing (15%).

Preprocessing:[2], [6]
Text: Lowercase, remove URLs/hashtags, tokenize via GPT-2 BPE.
Images: Resize to 224×224, normalize, and apply augmentations (flip, rotation).

Feature Extraction:
Visual: ViT-Base-16 generates embeddings from 16×16 image patches.
Textual: GPT-2 encodes tweets into contextual embeddings reflecting disaster cues.

Fusion & Classification:
Combine embeddings and classify with Random Forest or a small fusion transformer, optimized via validation accuracy and F1-score.

Generative Summarization:
Informative tweets are summarized using Flan-T5. Summaries (e.g., "Flooding in Dhaka, roads submerged") are evaluated with ROUGE-L and BLEU.

Interpretability:
Visualize attention maps from ViT and GPT-2 to identify key image regions and text tokens[7].

Tools & Training: Python (PyTorch, Hugging Face, scikit-learn), GPU (RTX 3050/T4), learning rate 2e-5, batch size 4–16, epochs 5–10.

Evaluation: Compare against CNN and BERT baselines using Accuracy, Precision, Recall, F1-score, ROC-AUC, and ROUGE-L.[8]

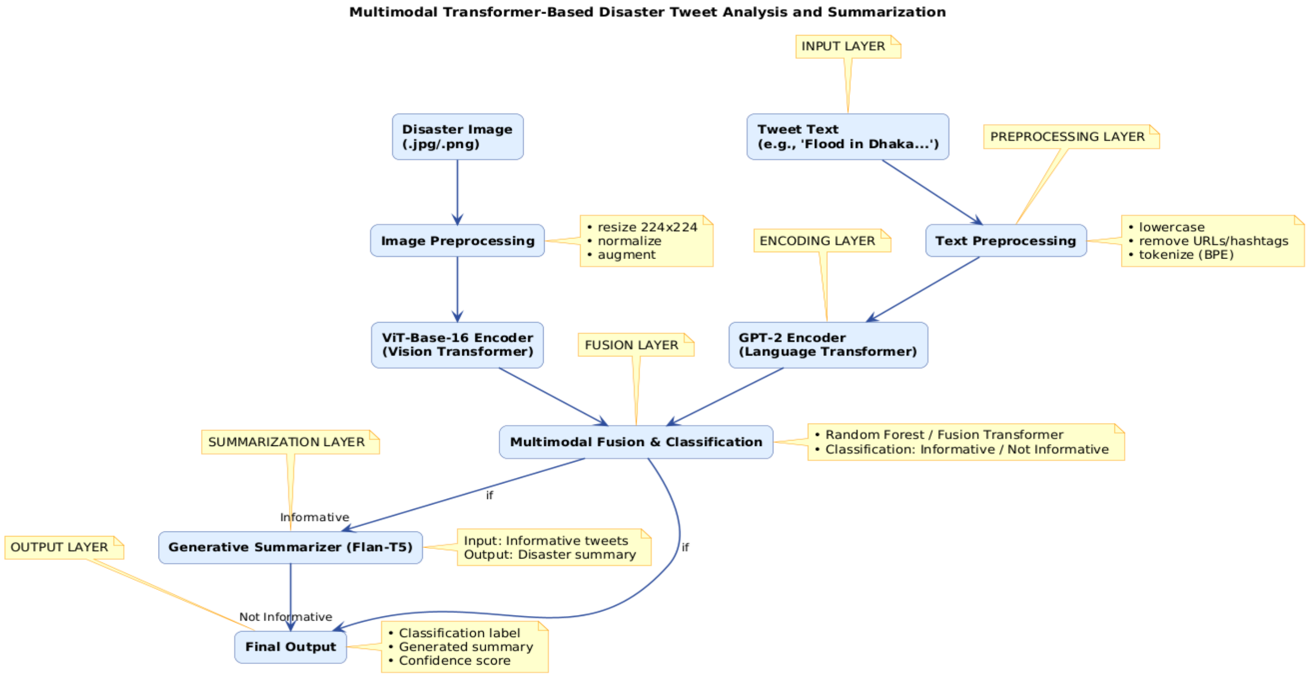
Multimodal Transformer-Based Disaster Tweet Analysis and Summarization
INPUT LAYER
Disaster Image (.jpg/.png)
Tweet Text (e.g., 'Flood in Dhaka...')
PREPROCESSING LAYER
Image Preprocessing
• resize 224x224
• normalize
• augment
ENCODING LAYER
Text Preprocessing
• lowercase
• remove URLs/hashtags
• tokenize (BPE)
ViT-Base-16 Encoder (Vision Transformer)
FUSION LAYER
GPT-2 Encoder (Language Transformer)
SUMMARIZATION LAYER
Multimodal Fusion & Classification
• Random Forest / Fusion Transformer
• Classification: Informative / Not Informative
if
Informative
if
OUTPUT LAYER
Generative Summarizer (Flan-T5)
Input: Informative tweets
Output: Disaster summary
Not Informative
Final Output
• Classification label
• Generated summary
• Confidence score

Figure 1:Tentative Methodology

7. Expected outcomes:

- A functional **multimodal transformer model** that classifies disaster-related tweets with ≥85% accuracy.
- Automatic **disaster summarization system** producing concise, readable reports.
- Demonstrated improvement over existing CNN/LSTM-based baselines.
- Visualization of model reasoning for greater interpretability.
- Contribution of a reproducible research framework using publicly available models and datasets.
- Potential foundation for future work on real-time disaster information dashboards.

8. Cost Estimate

Item	Estimated Cost (Tk.)	Notes
Cost of Material	0	Not required for this study
Fieldworks (if applicable)	0	No fieldwork involved
Conveyance / Data Collection	0	Data collected digitally
Drafting, Binding & Paper	3,000	Includes printing and binding
Miscellaneous	2,000	Contingency expenses

Total Estimated Cost: 5,000 Tk

9. Engineering & Society: This project contributes to social welfare[9] by enhancing disaster response efficiency. Automated analysis of multimodal social media data can help agencies detect critical incidents faster and allocate resources effectively. By transforming scattered social media posts into concise, actionable summaries, the system supports timely humanitarian interventions. It also encourages responsible AI use, emphasizing transparency and fairness through interpretable model outputs. In societal terms, the project showcases how engineering innovations in AI can directly save lives and improve public safety during emergencies.

10. Environment & Sustainability*:* The project promotes sustainability by using digital technologies to minimize the environmental footprint of disaster management. Early detection and fast reporting can reduce damage and prevent waste of rescue resources. The computational setup relies on energy-efficient models (small-scale ViT, GPT-2, and Flan-T5[10]) trained using low-power GPUs, reducing carbon emissions compared to large-scale AI systems. The methodology also encourages reuse of publicly available data and pre-trained models instead of creating new high-cost datasets, aligning with sustainable computing principles.

11. Engineering Ethics: All text and image data used in this project are publicly available and anonymized, respecting privacy and ethical data use standards. The work adheres to institutional research ethics guidelines. The similarity index of written content will be maintained below **15%** using originality checks (Turnitin or Grammarly). All sources are properly cited following academic norms. Model transparency and explainability are prioritized to ensure ethical deployment in real-world crisis systems, avoiding bias or misuse. Engineering practice principles such as responsibility, safety, and public welfare are upheld throughout the project.

12. References

[1] A. Vaswani *et al.*, “Attention Is All You Need,” p. 1, Jun. 2017, Accessed: Oct. 21, 2025. [Online]. Available: https://arxiv.org/pdf/1706.03762

[2] F. Alam, F. Ofli, and M. Imran, “CrisisMMD: Multimodal twitter datasets from natural disasters,” *12th International AAAI Conference on Web and Social Media, ICWSM 2018*, pp. 465–473, 2018, doi: 10.1609/ICWSM.V12I1.14983.

[3] M. R. Faisal *et al.*, “A Social Community Sensor for Natural Disaster Monitoring in Indonesia Using Hybrid 2D CNN LSTM,” *ACM International Conference Proceeding Series*, pp. 250–258, Oct. 2023, doi: 10.1145/3626641.3626932.

[4] S. Gite *et al.*, “Analysis of Multimodal Social Media Data Utilizing VIT Base 16 and GPT-2 for Disaster Response,” *Arab J Sci Eng*, 2025, doi: 10.1007/S13369-025-10314-7.

[5] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *ICLR 2021 - 9th International Conference on Learning Representations*, Oct. 2020, Accessed: Oct. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2010.11929

[6] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies - Proceedings of the Conference, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Oct. 21, 2025. [Online]. Available: https://arxiv.org/pdf/1810.04805

[7] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, vol. 21, pp. 1–67, Oct. 2019, Accessed: Oct. 21, 2025. [Online]. Available: https://arxiv.org/pdf/1910.10683

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners”, Accessed: Oct. 21, 2025. [Online]. Available: https://github.com/codelucas/newspaper

[9] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, “Processing Social Media Messages in Mass Emergency: Survey Summary,” *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*, pp. 507–511, Apr. 2018, doi: 10.1145/3184558.3186242.

[10] C. Raffel *et al.*, “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” *Journal of Machine Learning Research*, vol. 21, pp. 1–67, 2020, Accessed: Oct. 21, 2025. [Online]. Available: http://jmlr.org/papers/v21/20-074.html.

--------------------------------	--------------------------------	--------------------------------	-----------------------------
Signature of the Student	**Signature of the Supervisor**	**Signature of the External**	**Signature of the Head of the Department**