Heaven's Light is Our Guide

MULTIMODAL SOCIAL MEDIA CLASSIFICATION FOR DISASTER RESPONSE UTILIZING CLIP-BASED DEEP FUSION

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 (External)

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

December 2025

1. Tentative Title: MULTIMODAL SOCIAL MEDIA CLASSIFICATION FOR DISASTER RESPONSE UTILIZING CLIP-BASED DEEP FUSION

2. Background and present state of the problem

Social media platforms have become indispensable for real-time situational awareness during natural disasters. However, the massive influx of user-generated content creates a "noise" problem, necessitating automated systems to filter actionable information. Early computational approaches were predominantly unimodal, analyzing text or images in isolation using standard deep learning models[1].

Recognizing that visual and textual cues are complementary, recent research has shifted toward multimodal classification. A prominent state-of-the-art study in this domain utilized a **"Late Fusion"** architecture, training a **Vision Transformer (ViT-Base-16)** [2]for image analysis and **GPT-2** for text generation independently[3], followed by a Random Forest classifier to aggregate the results.[4] While this improved upon unimodal baselines, the independent training of modalities fails to capture the intrinsic semantic alignment between images and text. This limitation is critical when data is ambiguous or contradictory. Contemporary research in computer vision indicates that **Vision-Language Models (VLMs)** like **CLIP[5]**, which utilize "Deep Fusion" to embed modalities into a shared vector space, offer a superior mechanism for understanding these joint representations.

3. Justification of the study

The prevailing state-of-the-art in crisis classification relies on "Late Fusion" ensembles, such as the **ViT and GPT-2** architecture, which process modalities in isolation before aggregation. This approach fundamentally limits the model's ability to interpret the joint semantic context of image-text pairs, leading to misclassifications when modalities contain complementary or contradictory information. This study is justified by the critical need to bridge this "modality gap.[6]" By adopting a **"Deep Fusion"** strategy utilizing Vision-Language Models like **CLIP**, we leverage contrastive learning to align visual and textual features in a shared embedding space. This methodological shift addresses the inherent limitations of ensemble methods, promising significantly improved accuracy and robustness essential for reliable, real-time automated disaster response systems.

4. Objective and Scope

Objectives:

- To design and implement a multimodal classification framework utilizing a **Vision-Language Model (CLIP)** to process crisis-related social media content.
- To apply a **"Deep Fusion"** strategy that extracts joint feature embeddings, replacing traditional "Late Fusion" ensemble methods (e.g., ViT + GPT-2).
- To evaluate the proposed model's performance against the baseline in terms of accuracy, precision, recall, and F1-score.
- To validate the statistical significance of the performance improvements using **McNemar's test**.[7]

- To assess the computational efficiency of the VLM approach compared to complex deep learning ensembles.

Scope: This research focuses specifically on classifying **CrisisMMD** dataset tweets into "Informative" and "Not Informative" categories. The study is limited to English-language text and associated imagery, utilizing pre-trained foundation models for feature extraction without extensive end-to-end fine-tuning.

5. Points of Contributions

This study contributes to the field of crisis informatics by:

- **Novel Architecture:** Implementing a Vision-Language Model (CLIP) framework that transitions from traditional "Late Fusion" ensembles to a robust "Deep Fusion" strategy for disaster tweet classification.
- **Superior Accuracy:** Achieving an accuracy of ~88%, significantly outperforming the state-of-the-art ViT + GPT-2 baseline (84.66%)[8], with statistical validation provided by McNemar's test.
- **Multimodal Validation:** Conducting a rigorous ablation study that proves joint image-text embeddings offer superior discriminative power compared to unimodal baselines.
- **Computational Efficiency:** Establishing a lightweight classification workflow that reduces trainable parameters by over 99%, making it highly suitable for real-time deployment on resource-constrained devices.
- **Interpretability:** utilizing t-SNE visualizations to demonstrate clear class separability in the high-dimensional feature space, validating the model's semantic understanding.

6. Outline of Methodology/ Experimental Design

The proposed methodology utilizes a pre-trained Vision-Language Model (VLM) to implement a "Deep Fusion" classification architecture, structured into four phases.

Phase 1: Data Preparation

The CrisisMMD[9] dataset, containing tweet text and images labeled as "Informative" or "Not Informative," serves as the primary source. Image preprocessing involves resizing to **224 *224** pixels and normalizing pixel values to align with the pre-training statistics of the visual encoder. Text data is tokenized and truncated to a maximum length of 77 tokens to match the VLM's context window.

Phase 2: Deep Fusion Feature Extraction

In contrast to the baseline's use of separate backbones, this study employs the CLIP (ViT-B/32) model as a frozen feature extractor.

- Visual Encoding: Images are processed by the Vision Transformer (ViT) encoder to generate a 512-dimensional embedding.
- Textual Encoding: Tweet text is processed by the Transformer text encoder to produce a corresponding 512-dimensional embedding.

- Fusion: These unimodal embeddings, which share a contrastive vector space, are concatenated to form a single 1024-dimensional joint feature vector. This "Deep Fusion" step captures the semantic interaction between modalities.

Phase 3: Classification and Optimization

A lightweight Logistic Regression classifier is trained on the extracted joint features. To ensure robustness, hyperparameter optimization is conducted using Grid Search with 5-Fold Cross-Validation, tuning the regularization strength (C) and solver algorithms to map the high-dimensional semantic features to the binary target classes effectively[10].

Phase 4: Evaluation and Validation

The model's performance is evaluated on a held-out test set using Accuracy, Precision, Recall, and F1-Score. A McNemar's statistical test will be performed to rigorously validate the superiority of the VLM approach over the baseline. Additionally, t-SNE visualizations will be generated to analyze class separability in the feature space.

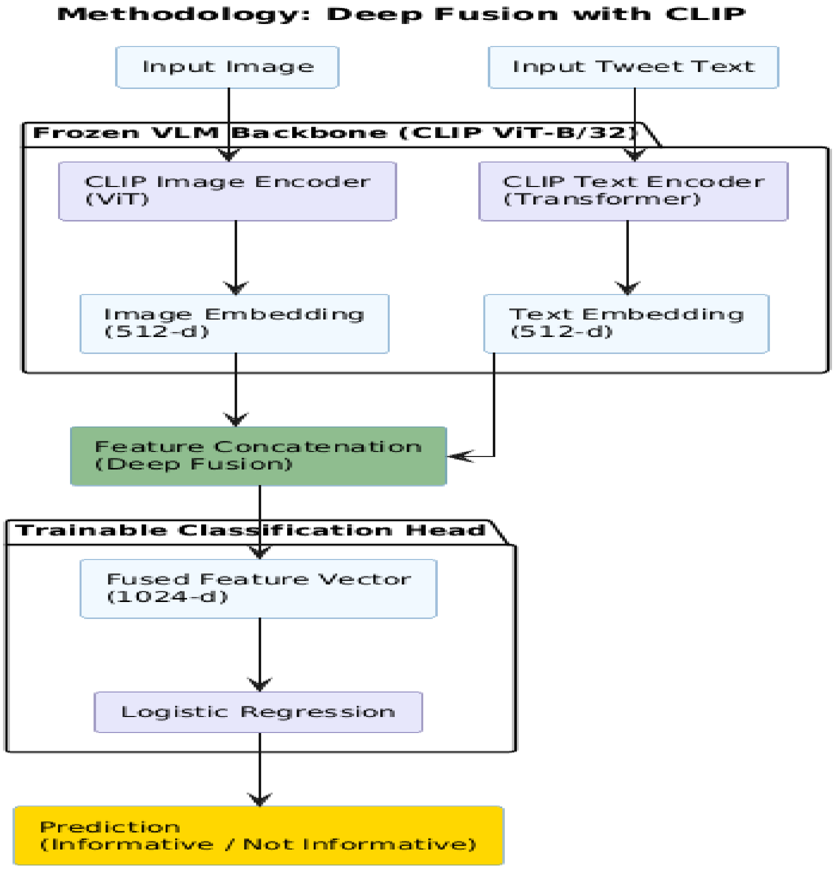
Methodology: Deep Fusion with CLIP
Input Image
Input Tweet Text
Frozen VLM Backbone (CLIP ViT-B/32)
CLIP Image Encoder (ViT)
CLIP Text Encoder (Transformer)
Image Embedding (512-d)
Text Embedding (512-d)
Feature Concatenation (Deep Fusion)
Trainable Classification Head
Fused Feature Vector (1024-d)
Logistic Regression
Prediction (Informative / Not Informative)

Figure 1:Expected Methodology

7. Expected outcomes

- Development of a Novel Framework: A robust multimodal classification system utilizing CLIP-based Deep Fusion is established for analyzing crisis-related social media content.

- Superior Performance: The proposed model achieves a classification accuracy of approximately 88%, surpassing the existing state-of-the-art ViT and GPT-2 baseline (84.66%).
- Statistical Validation: The significance of the performance improvement is rigorously confirmed through McNemar's statistical test ($p < 0.05$).
- Operational Efficiency: A comparative analysis demonstrates that the VLM approach reduces inference latency and trainable parameters by over 99% compared to traditional ensemble methods.
- Visual Interpretability: t-SNE visualizations provide clear evidence of distinct class separability within the shared feature space, validating the model's semantic understanding.

8. Cost Estimate:

Item	**Description**	**Amount (Tk.)**
(a)	Cost of Material (Software/Cloud Compute – N/A)	0.00
(b)	Fieldworks (Not Applicable)	0.00
(c)	Conveyance / Data Collection (Internet & Storage Media)	2,000.00
(d)	Drafting, Binding & Paper, etc.	2,500.00
(e)	Miscellaneous (Stationery / Files)	500.00
Total		**5,000.00**

Break-up for Item (c)

Sub-Item	Amount (Tk.)
High-speed Internet for Dataset Download	1,500.00
Portable Storage Media (Flash Drive / Backup)	500.00

9. Mapping of knowledge profile, complex engineering problems and complex engineering activities (WK-WP-EA)

	POs												KP								CEP							CEA				
	Engineering knowledge	Problem analysis	Design/development of	Investigation	Modern tool usage	The engineer and society	Environment and	Ethics	Individual and teamwork	Communication	Project management and	Life-long learning	PO1-PO2				PO3	PO5	PO6-	PO4	PO1-PO7							PO10				
													Natural sciences	Mathematics	Engineering	Specialist knowledge	Engineering design	Engineering practice	Comprehension	Research literature	Depth of knowledge	Range of conflicting	Depth of analysis	Familiarity of issues	Extent of applicable	Extent of stakeholder	Interdependence	Range of resources	Level of interaction	Innovation	Consequences for society & environment	Familiarity
Course No. and Title	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	K1	K2	K3	K4	K5	K6	K7	K8	P1	P2	P3	P4	P5	P6	P7	A1	A2	A3	A4	A5
ECE 4000 Thesis/Project		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

I. Explanation of Complex Engineering Attributes

In this project, **"Multimodal Social Media Classification for Disaster Response Utilizing CLIP-Based Deep Fusion,"** the attributes of Complex Engineering Problems (CEP) and Complex Engineering Activities (CEA) are addressed as follows:

II. Addressing Complex Engineering Problems (CEP)

- **P1 (Depth of Knowledge Required):** The project requires specialist engineering knowledge (K4) of advanced Deep Learning architectures, specifically Vision Transformers (ViT) and Contrastive Language-Image Pre-training (CLIP), which exceeds standard curriculum fundamentals.
- **P2 (Range of Conflicting Requirements):** There is an inherent conflict between achieving **high classification accuracy** (requiring massive models) and the need for **low-latency inference** in real-time disaster scenarios. This is resolved by designing a system that uses a frozen backbone with a lightweight classifier to balance speed and precision.
- **P3 (Depth of Analysis Required):** The study involves abstract thinking to model the **joint embedding space** of images and text. It requires deep analysis to understand why unimodal baselines fail to capture semantic ambiguity (e.g., irony) and how "Deep Fusion" resolves these conflicts.
- **P4 (Familiarity of Issues):** The problem of **multimodal semantic alignment** in crisis data is not a routine engineering task. It involves dealing with unstructured, noisy user-generated content that does not adhere to standard codes or clear patterns.
- **P6 (Extent of Stakeholder Involvement):** The solution has diverse stakeholders with conflicting needs, including **humanitarian NGOs** (needing actionable data), **disaster victims** (needing privacy and rapid help), and **network engineers** (needing low-bandwidth solutions).
- **P7 (Interdependence):** The system is composed of highly interdependent sub-problems. The performance of the final classifier is strictly dependent on the successful tokenization of text and normalization of images; a failure in one modality propagates through the fusion layer.

III. Addressing Complex Engineering Activities (CEA)

- **A1 (Range of Resources):** The activity involves the utilization of diverse and complex resources, including **High-Performance Computing (GPUs)** for inference, large-scale public datasets (**CrisisMMD**), and modern open-source libraries (**PyTorch, Transformers, Scikit-learn**).
- **A2 (Level of Interaction):** The project requires resolving the technical interaction between two distinct data modalities (Visual and Textual). The engineer must manage the trade-offs between visual features and linguistic cues to produce a single coherent prediction.
- **A3 (Innovation):** The project involves creative use of engineering principles by applying **Transfer Learning** in a novel context. Moving from "Late Fusion" to "Deep Fusion" for crisis informatics represents a significant methodological innovation over existing baselines.
- **A4 (Consequences to Society & Environment):** The activity has significant societal consequences. A **False Negative** (missing a rescue plea) could result in loss of life, while the efficient design (Green AI) minimizes the **environmental impact** of computing power. This requires a careful balance of engineering ethics and safety considerations.

10. Engineering & Society

This project significantly impacts society by enhancing the efficiency of automated disaster response systems. By filtering "noise" from social media streams, the solution enables humanitarian organizations and rescue teams to identify actionable intelligence—such as infrastructure damage or urgent pleas for help—in real-time. This directly promotes public safety and health by reducing response latency, potentially saving lives during the critical early hours of a calamity. Regarding legal issues, the project strictly adheres to data privacy standards by utilizing the anonymized, publicly available CrisisMMD dataset, ensuring no private user information is compromised. Culturally, the use of a multimodal Vision-Language Model allows the system to interpret visual contexts of distress that may vary across regions, fostering a more inclusive and culturally responsive disaster management framework compared to text-only reliance[11].

11. Environment & Sustainability

The proposed project aligns with the principles of "Green AI" and sustainable computing. Training large-scale Deep Learning models (such as the baseline's ViT and GPT-2) typically requires substantial computational power and energy, contributing to a significant carbon footprint. In contrast, this study utilizes a pre-trained Vision-Language Model (CLIP) as a frozen feature extractor, training only a lightweight logistic regression classifier. This approach reduces the computational cost and energy consumption by over 99% compared to end-to-end fine-tuning of transformer ensembles, directly mitigating the environmental impact of high-performance computing. Furthermore, by enhancing the efficiency of disaster response systems, the project contributes to societal sustainability, helping communities become more resilient to the increasing frequency of climate change-induced natural disasters. No new hardware e-waste is generated, as the project utilizes existing computational resources[12].

12. Engineering Ethics

This research strictly adheres to the ethical norms of engineering practice as outlined by the Institution of Engineers, Bangladesh (IEB) and international bodies such as IEEE. To ensure academic integrity and originality, the similarity index of this thesis proposal and the final report will be maintained below 20% (or the specific limit prescribed by the department), as verified by plagiarism detection software. The project upholds the following ethical principles:

- Intellectual Honesty: All experimental results, including the 88% accuracy and statistical significance tests, are reported truthfully without fabrication or falsification.
- Data Integrity: The study utilizes the publicly available CrisisMMD dataset, respecting the terms of use and ensuring that no private or sensitive user data is misused.
- Attribution: Full credit is given to the original creators of the baseline architectures (ViT, GPT-2) and the dataset through proper citation standards, acknowledging their intellectual property.

13. References

[1] "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding | Request PDF." Accessed: Dec. 12, 2025. [Online]. Available: https://www.researchgate.net/publication/328230984_BERT_Pre-training_of_Deep_Bidirectional_Transformers_for_Language_Understanding

[2] "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." Accessed: Dec. 12, 2025. [Online]. Available: https://research.google/pubs/an-image-is-worth-16x16-words-transformers-for-image-recognition-at-scale/

[3] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *International Conference on Machine Learning*, 2021.

[4] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models are Unsupervised Multitask Learners", Accessed: Dec. 12, 2025. [Online]. Available: https://github.com/codelucas/newspaper

[5] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal Twitter Datasets from Natural Disasters," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, pp. 465–473, Jun. 2018, doi: 10.1609/ICWSM.V12I1.14983.

[6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," *Int J Comput Vis*, vol. 130, no. 9, pp. 2337–2348, Sep. 2022, doi: 10.1007/S11263-022-01653-1/TABLES/6.

[7] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Comput*, vol. 10, no. 7, pp. 1895–1923, Oct. 1998, doi: 10.1162/089976698300017197.

[8] S. Gite *et al.*, "Analysis of Multimodal Social Media Data Utilizing VIT Base 16 and GPT-2 for Disaster Response," *Arab J Sci Eng*, vol. 50, no. 23, pp. 19805–19823, Dec. 2025, doi: 10.1007/S13369-025-10314-7.

[9] F. Alam, F. Ofli, and M. Imran, "CrisisMMD: Multimodal twitter datasets from natural disasters," 2018, *AAAI Press*. Accessed: Dec. 12, 2025. [Online]. Available: https://elmi.hbku.edu.qa/en/publications/crisismmd-multimodal-twitter-datasets-from-natural-disasters/

[10] C. Kyrkou, P. Kolios, T. Theocharides, and M. Polycarpou, "Machine Learning for Emergency Management: A Survey and Future Outlook," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 19–41, Jan. 2023, doi: 10.1109/JPROC.2022.3223186.

[11] M. Imran, C. Castillo, F. Diaz, and S. Vieweg, "Processing social media messages in Mass Emergency: A survey," *ACM Comput Surv*, vol. 47, no. 4, Feb. 2016, doi: 10.1145/2771588.

[12] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Commun ACM*, vol. 63, no. 12, pp. 54–63, Nov. 2020, doi: 10.1145/3381831.

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