Xin1212 at NakbaVirality Shared Task: Frozen CLIP with Residual Adapter for Multimodal Virality Classification

Xinyan Zhang
Dalian University of Technology
2737717435@mail.dlut.edu.cn

Bingzhou Yang
Dalian University of Technology
ybz19596612@mail.dlut.edu.cn

April 16, 2026

Abstract

We describe our system for the NakbaVirality shared task on multimodal virality classification. Our final approach uses a frozen LAION CLIP backbone, a lightweight residual adapter over fused text–image embeddings, and a small MLP classification head. On our development split, the best configuration (V9) achieves Macro-F1 of 0.5492 and virality-weighted F1 of 0.5252. On our official test submission, our system obtains F1-score 0.4559 and accuracy 0.6089 according to the platform scorer. We provide implementation details, ablations across multiple versions (Baseline–V9), and practical error analysis for reproducibility.

Keywords: multimodal virality, CLIP, residual adapter, multimodal classification

1. INTRODUCTION

The NakbaVirality task studies multimodal virality prediction in high-stakes geopolitical discourse, where both textual framing and imagery influence reach. The task is a 3-class classification problem (*Low*, *Medium*, *High* virality) evaluated primarily with Macro-F1.

This setting is challenging for three reasons: small dataset, class imbalance, and multilingual noisy content.

Our objective is to improve minority-class sensitivity while preserving stable optimization. Starting from a CLIP-based baseline, we iteratively developed versions V1–V9 and selected V9 as the final system. The key strategy is to keep the CLIP backbone frozen and add a residual adapter.

Main contributions.

- We provide a reproducible versioned study (Baseline to V9).
- We show that frozen-backbone + residual-adapter improves Macro-F1.
- We provide practical error analysis for shared-task settings.

2. BACKGROUND

2.1. Task setup

Input consists of post text and an associated image. Output is one label from {`High viral`, `Medium viral`, `Low viral`}.

2.2. Data description and split protocol

Following the shared task description Ezzini et al. (2026), the dataset is collected from X and Reddit with multilingual content.

We performed a stratified train/dev split ($80/20$, seed 42). This produced $1,691$ training instances in total and a 339-example development split for model selection. The official blind test set used in our submissions contains 872 examples.

2.3. Related modeling direction

Recent multimodal sentiment research reports that naive feature fusion can introduce semantic noise across modalities, reducing classification quality. SECIF addresses this by explicitly combining semantic enhancement and cross-modal interaction fusion Mu et al. (2025).

A complementary direction uses hierarchical cross-modal attention to better capture inter-modal dependencies than text-only pipelines Vamsidhar et al. (2025).

Beyond sentiment analysis, multimodal fusion has also proven useful in social-media fake news detection. A recent BERT-based framework combines text and image information with OCR-derived visual text, showing that image-side textual cues can contribute important evidence Al-alshaqi et al. (2025).

For virality, findings show that community structure and social reinforcement affect diffusion dynamics. More recent multimodal work studies early virality prediction in cross-lingual Reddit settings Dogan et al. (2025).

Finally, conflict-domain discourse analysis across platforms highlights substantial temporal

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 144–146

and narrative variation, reinforcing the need for robust multimodal modeling Antonakaki and Ioannidis (2025).

3. SYSTEM OVERVIEW

3.1. Architecture

Our final model (V9) has four stages:

1. CLIP encoder: `laion/CLIP-ViT-B-32-laion2B-s34B-b79K`
2. Late fusion by concatenating text and image embeddings
3. Residual adapter (two-layer MLP) for task-specific correction
4. MLP classifier with dropout for 3-way prediction

Let $z_t, z_i \in R^d$ be CLIP text and image embeddings. We fuse and adapt features as:

$$h = [z_t; z_i] \in R^{2d}, \quad \tilde{h} = h + A(h), \qquad (1)$$

where $A(\cdot)$ is the residual adapter. Final logits are

$$\hat{y} = W_2\sigma\left(W_1\tilde{h}\right). \qquad (2)$$

3.2. Training objective and optimization

We train with weighted cross-entropy and label smoothing ($\epsilon = 0.1$). For a sample with target class y, the smoothed target distribution is $q_k = (1-\epsilon)\mathbf{1}[k=y] + \epsilon/K$, and the loss is $\mathcal{L} = -\sum_{k=1}^{K} w_k q_k \log p_k$, where w_k are class weights and $K = 3$.

4. EXPERIMENTAL SETUP

4.1. Data usage

We trained and selected models on a stratified $80/20$ split of the released training data (seed 42). The development split (339 samples) was used for all local ablations.

4.2. Training configuration (V9)

Key parameters: Batch size 16, Epochs 20, Learning rate $1\mathrm{e}^{-3}$, label smoothing 0.1, and cosine scheduler with warmup.

5. RESULTS

5.1. Version comparison on development split

Table 1 summarizes metrics logged in our study.

Version	Macro-F1	Weighted-F1	Acc.
Baseline	0.4798	0.3924	0.6401
V1	0.5398	0.5072	0.5870
V5	0.5429	0.5052	0.5929
V8	0.5023	0.4633	0.5457
V9	**0.5492**	**0.5252**	0.5959

Table 1: Results on development split.

5.2. Official test submission

The platform reported: F1-score 0.4559, Accuracy 0.6089, Precision 0.6000, Recall 0.6089.

6. CONCLUSION

We presented a frozen-CLIP + residual-adapter approach for NakbaVirality. V9 was the strongest dev configuration and improved minority-class behavior. On our official test submission, the system achieved F1 0.4559.

Mohammed Al-alshaqi, Danda B Rawat, and Chunmei Liu. 2025. A bert-based multimodal framework for enhanced fake news detection using text and image data fusion. *Computers*, 14(6):237.

Despoina Antonakaki and Sotiris Ioannidis. 2025. Cross-platform digital discourse analysis of the israel-hamas conflict: Sentiment, topics, and event dynamics. *arXiv preprint arXiv:2601.02367*.

Sedat Dogan, Nina Dethlefs, and Debarati Chakraborty. 2025. Early multimodal prediction of cross-lingual meme virality on reddit: A time-window analysis. *arXiv preprint arXiv:2510.05761*.

Saad Ezzini, Salima Lamsiyah, Shadi Abudalfa, Samir El-Amrany, and Walid Alsafadi. 2026. The nakbavirality shared task on multimodalvirality prediction in high-stakes discourse. In *Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Guangyu Mu, Ying Chen, Xiurong Li, Li Dai, and Jiaxiu Dai. 2025. Semantic enhancement and cross-modal interaction fusion for sentiment analysis in social media. *PLOS ONE*, 20(4):e0321011.

D Vamsidhar, Parth Desai, Aniket K Shahade, Shruti Patil, and Priyanka V Deshmukh. 2025.

Hierarchical cross-modal attention and dual audio pathways for enhanced multimodal sentiment analysis. *Scientific Reports*, 15(1):25440.

A. REPRODUCIBILITY DETAILS

The final submitted model uses a frozen CLIP backbone with a residual adapter and an MLP classifier. Optimization uses AdamW, cosine schedule with warmup, and label smoothing.