Digilians at NakbaVirality Shared Task: Bidirectional Cross-Attention for Multimodal Virality Prediction

Noureldeen H. Mohamed[1], Ahmed E. Hassan[2], AbdelrhmanM. Fawzy[3], Mohamed A. Abdelghany[4], Ahmed A. Hassan[5], Ahmed S. Qassim[6], Fady A. Abd El Sayed[7], Mohamed H. Mohamed[8], Rahma M. Mohamed[9], Arwa M. Abou-Attia[10], Mayar M. Mohamed[11], Shahd A. Sawla[12], Shahd O. Mahmoud[13]

[1]Faculty of Computer Engineering, Al-Shorouk Academy [2]Alexandria Higher Institute of Engineering and Technology (AIET) [3]6th October University [4]Faculty of Computer Science, Modern Academy [5]Arab Academy for Science, Technology and Maritime Transport (AASTMT) [6]Higher Institute of Engineering and Technology, Kafr El-Sheikh [7]Faculty of Computer Science and AI, Ahram Canadian University (ACU) [8]Faculty of Computer Science, Canadian International College (CIC) [9]Modern Sciences and Arts University (MSA) [10]Delta University for Science and Technology [11]Modern Academy, Faculty of Computer Science [12]Faculty of Artificial Intelligence, Menoufia University [13]Computer Science, Sohag University

Cairo, Egypt; Alexandria, Egypt; Giza, Egypt; Kafr El-Sheikh, Egypt; New Damietta, Egypt; Al-Qalyubia, Egypt; Sohag, Egypt

noorhossam1995@gmail.com, ahmedeid553@gmail.com, abdo.elsaadny74@gmail.com, mabdelghany2000111@gmail.com, ahmed.aly.hassan.official@gmail.com, ahmedsedeqqassem@gmail.com, fadyatef392@gmail.com, mohamedhabdelazezfarhat@gmail.com, rahma.elsheikh98@gmail.com, arwaabouattia@gmail.com, mayar22mostafa@gmail.com, shahdsawla@outlook.com, shahdomar444466@gmail.com

Abstract

The NakbaVirality shared task involves multimodal virality prediction using 2,600 multilingual posts from X and Reddit. The researchers developed an architecture integrating XLM-RoBERTa and Vision Transformer (ViT) encoders with bidirectional cross-attention, proving superior to simple concatenation. By employing focal loss, class weighting, and layer-wise learning rate scheduling, the system achieved 0.6009 accuracy on the hidden test set.

Keywords: Multimodal learning, Virality prediction, Bidirectional cross-attention, NakbaVirality, Conflict discourse.

1 Introduction

This research addresses multimodal virality during armed conflicts (Ezzini et al., 2026), where text and images jointly shape perception across three engagement buckets. The study tackles structured cross-modal dependencies (Zhao et al., 2024), severe class imbalance (Shen et al., 2023), and cross-platform generalization gaps. The proposed model integrates XLM-RoBERTa (Conneau et al., 2020) and ViT (Dosovitskiy et al., 2021) via bidirectional cross-attention (Vaswani et al., 2017) to dynamically reweight modalities.

Key Performance Metrics:

- **Development Set:** Weighted-F1 of 0.7432 and Macro-F1 of 0.6845.
- **Hidden Test Set:** Accuracy 0.6009, Macro F1 of 0.498, ranking 4th among 29 teams (107 total submissions).

The results demonstrate that while structured fusion and imbalance-aware optimization improve minority-class detection, measurable performance degradation on the test set highlights the difficulty of cross-platform generalization in dynamic socio-political environments.

2 Related Work

This section situates virality prediction at the intersection of network science, psychology, and NLP. We categorize previous research into content-based, network-based, and timing-based studies.

2.1 Defining and Operationalizing Virality

There is a lack of consensus on defining "virality." We follow the precedent of Dogan et al. (2025), who used a Hybrid Score to normalize engagement. While models like ViralBERT achieved a 13% improvement in F1 by using text and user signals, they ignore visual content—a critical gap in conflict discourse where images carry high emotional weight.

2.2 Multimodal Architectures for Virality

Research shows that cross-attention mechanisms outperform simple fusion. We build on the Multi-Way Multi-Modal Transformer (MMT) and findings by Sanchez Villegas et al. (2024), which showed that explicit cross-modal attention can improve performance by up to 2.6 F1.

2.3 Multilingual Representation and Class Imbalance

XLM-RoBERTa is superior to mBERT for multilingual tweet-style text. We adopt Focal Loss and Flan-T5-based paraphrase augmentation to address the fact that high-virality posts occupy only the top 10% of the engagement distribution.

2.4 Research Gaps and Our Contribution

We identify three gaps: a lack of conflict-specific multilingual datasets, limited use of end-to-end bidirectional cross-attention, and a need for systematic treatment of class imbalance in virality tasks.

Proceedings of the 2nd International Workshop on Nakba Narratives as Language Resources (Nakba-NLP 2026) @ LREC 2026, pages 265–268

3 Dataset Details

The study uses 1,691 multimodal samples from the NakbaVirality shared task. The goal is to classify posts into Low, Medium, and High Viral categories. The data is split with 15% reserved for internal validation.

3.1 Class Distribution and Imbalance

The dataset suffers from pronounced class imbalance, where “High Viral” is the minority. This skew biases models toward majority classes, requiring specific mitigation strategies.

3.2 Multimodal Nature of the Data

Content includes Arabic, English, and code-mixed text. We use OCR to extract text from images, concatenating it with post captions to enrich the input signal.

3.3 Validation vs. Competition Test Performance

A significant generalization gap exists: the model achieved 74.4% accuracy on internal validation but approximately 50% on the hidden test set, likely due to distributional shifts.

4 Model Architecture

The architecture uses a dual-encoder setup with a cross-attention fusion module and a three-layer classification head.

4.1 Text Encoder

Uses xlm-roberta-base to process multilingual input. The [CLS] token representation (dimension 768) is extracted as the primary text feature vector.

4.2 Image Encoder

Uses ViT-B/16, which processes 224×224 images into 16×16 patches. The class token representation (dimension 768) serves as the global visual feature.

4.3 Projection Layers

Both encoders pass through linear layers, Layer Normalization, and GELU activation, aligning both modalities into a shared 768-dimensional space.

4.4 Bidirectional Cross-Attention Fusion

The fusion mechanism is the central contribution. Text (T) and image (V) features mutually condition each other via parallel 8-head Multi-Head Attention (MHA):

$$T' = MHA(Q=T, K=V, V=V)$$

$$V' = MHA(Q=V, K=T, V=T)$$

The resulting vectors are concatenated into a 1536-dimensional representatio and compressed back to 768 dimensions.

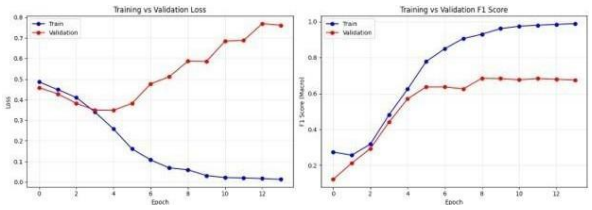

Figure 1: Multimodal Cross-Attention Architecture. Dashed bidirectional arrows indicate cross-modal attention flow.

Table 1: Architecture Component Summary

Component	Output Dim
Text Encoder (XLM-R)	768
Image Encoder (ViT-B/16)	768
Projection Layers	768
Cross-Attention (each)	768
Concatenated Fusion	1536
Compressed Fusion	768
Classifier	768→384→192→3

4.5 Validation Results

The model achieves strong performance on the High virality class (F1 = 0.938), indicating effective cross-modal alignment.

Table 2: Validation Performance Summary

Metric	Score
Accuracy	0.7440
Weighted F1	0.7432
Macro F1	0.6845
F1 – Low Viral	0.6920
F1 – Medium Viral	0.5880
F1 – High Viral	0.9380

5 Algorithms

5.1 Text Preprocessing and Translation

To unify Arabic, English, and code-mixed formats, a pipeline uses a heuristic-based detector and neural machine translation (Thakkar et al., 2024). A rule-based cleaner removes URLs and anonymizes user mentions.

5.2 Image Preprocessing and Augmentation

Images are resized to 224×224 pixels and normalized using ImageNet statistics. Training employs light stochastic augmentations, including random brightness and horizontal flipping.

5.3 Data Augmentation for Class Imbalance

Targeted augmentation is applied exclusively to the minority High Viral class:

- **Text:** Flan-T5 paraphrasing generates semantically equivalent restatements.
- **Images:** An aggressive pipeline varies brightness, contrast, rotation, and shearing.

5.4 Focal Loss with Class Weighting

The model addresses majority-class bias using Focal Loss:

$$FL(p_t) = -\alpha_t(1 - p_t)^r \log(p_t) \quad [\gamma = 2.0]$$

Class weights α_t are calculated inversely to frequency to amplify gradients for minority classes.

6 Experimental Design

6.1 Training Configuration

The model was trained for 20 epochs using a batch size of 16 on a single GPU with mixed-precision. Early stopping with patience of 6 epochs was employed.

6.2 Evaluation Protocol

Performance was measured using Accuracy, Macro F1, and Weighted F1. Macro F1 served as the primary optimization criterion.

6.3 Analysis of the Generalization Gap

The performance drop is attributed to distributional shift, data scarcity, and selection bias towards the validation set.

7 Results

Optimal model performance was reached at epoch 9, achieving 74.4% overall accuracy. Figure 2 plots training and validation metrics across epochs.

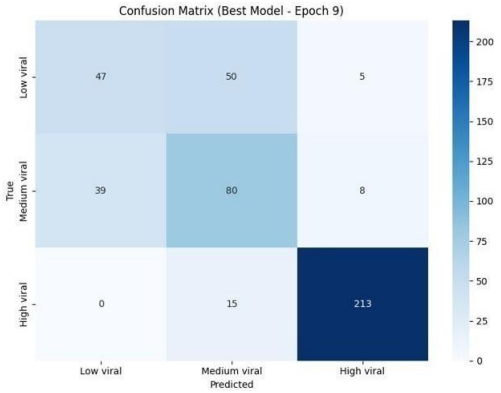

Figure 2: Training and validation performance metrics across epochs.

8 Architecture Illustration

Figure 3 illustrates the overall architecture. Text inputs are encoded using XLM-RoBERTa, images through ViT. The resulting representations are fused and passed to a classification head to predict virality category.

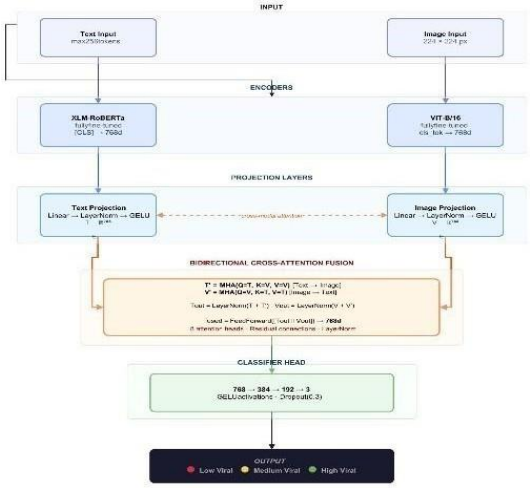

Figure 3: Proposed multimodal architecture integrating XLM-RoBERTa (text) and ViT (image) with a fusion layer and classification head.

9 Discussion & Conclusion

The multi-modal approach is highly effective for identifying viral content. By fusing text, images, and OCR content, the model improves minority class detection and validates multi-modal learning for real-world conflict-monitoring tasks. The bidirectional cross-attention mechanism enables richer modality interaction than late fusion baselines, and imbalance-aware training substantially improves minority-class recall. Future work should address cross-platform distributional shift through domain adaptation or additional data collection.

10 Bibliographical References

Chen, Yen-Chun, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. *In Proceedings of ECCV*, pages 104–120.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *In Proceedings of ACL*, pages 8440–8451.

Cottle, Simon. 2006. *Mediatized Conflict: Developments in Media and Conflict Studies*. Open University Press.

Cui, Yin, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. *In Proceedings of CVPR*, pages 9268–9277.

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

Jin, Zhijie, Cao Cao, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of ACM Multimedia*, pages 795–816.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *In Proceedings of ICCV*, pages 2980–2988.

Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *In Proceedings of NeurIPS*.

Papacharissi, Zizi. 2015. *Affective Publics: Sentiment, Technology, and Politics*. Oxford University Press.

Rameez, Rahmani, and Yilmaz. 2022. ViralBERT: A user-focused BERT-based approach to virality prediction.

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

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16×16 words: Transformers for image recognition at scale. *In Proceedings of ICLR.*

Villegas, Danae Sánchez, Daniel Preoțiuc-Pietro, and Nikolaos Aletras. 2023. Improving multimodal classification of social media posts by leveraging image-text auxiliary tasks.

Shen, Meng, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See. 2023. Towards balanced active learning for multimodal classification. *In Proceedings of ACM Multimedia.*

Tan, Hao, and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. *In Proceedings of EMNLP,* pages 5100–5111.

Tang, Jiajia, Li Kang, Hou Ming, Jin Xuanyu, Kong Wanzeng, Ding Yu, and Zhao Qibin. 2022. MMT: Multi-way multimodal transformer for multimodal learning. *In Proceedings of IJCAI,* pages 3458–3465.

Tedjasukmana, Jeffrey Junior, and Alexander Agung Santoso Gunawan. 2025. Classifying viral Twitter tweets with transformer models and multi-layer perceptron. *Engineering, Mathematics and Computer Science Journal (EMACS)*, 7(1), pages 80–88.

Thakkar, Gaurish, Sherzod Hakimov, and Marko Tadić. 2024. M2SA: Multimodal and multilingual model for sentiment analysis of tweets.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *In Proceedings of NeurIPS*, pages 5998–6008.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380), pages 1146–1151.

Zhao, Fei, Chengcui Zhang, and Baocheng Geng. 2024. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9).

Zhou, Xinyi, and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys*, 53(5), pages 1–40.