CUET_SYNTHETICA@EEUCA 2026: Gated Cross-Modal Attention with Domain-Adapted Text Encoding for Vaccine-Critical Meme Detection

Sumaiya Zaman, Miftahul Jannat Rishta, Shiti Chowdhury
Department of Computer Science and Engineering,
Chittagong University of Engineering and Technology, Bangladesh
{u2104110, u2104019, u2004027}@student.cuet.ac.bd

Abstract

Vaccine-critical memes have emerged as a growing challenge for public health communication, combining images and text to spread misinformation in ways that are difficult to detect automatically. In this paper, we have described our system for the EEUCA 2026 Shared Task on Multimodal Vaccine-Critical Meme Detection, classifying memes from the VaxMeme dataset into Vaccine-Critical, Neutral and Pro-Vaccine categories. We have experimented with multiple text encoders and visual backbones, finding that Twitter-RoBERTa fused with CLIP ViT-L/14 through gated cross-modal attention has achieved a test macro F1 of 0.8357. We have further shown that domain-specific pretraining has outperformed larger general-purpose models, highlighting the importance of domain adaptation over raw model scale. Finally, our system has secured the 3rd position on the shared task leaderboard.

1 Introduction

Memes have evolved into powerful tools for spreading vaccine misinformation online, combining text and images in ways that are difficult to counter through traditional fact-checking methods. While researchers have made progress in detecting harmful memes broadly (Suryawanshi et al., 2020), vaccine-specific meme detection has remained largely underexplored (Naseem et al., 2023).
The shared task has provided the VaxMeme dataset, consisting of over 10,000 manually annotated vaccination-related memes labeled as Vaccine-Critical, Neutral and Pro-Vaccine. It has encouraged the development of multimodal systems that jointly leverage textual and visual information for fine-grained meme understanding (Naseem et al., 2023). In the EEUCA 2026 Shared Task on Multimodal Vaccine-Critical Meme Detection, we have experimented with multiple text encoders and visual backbones, finding that Twitter-RoBERTa paired with CLIP ViT-L/14 under gated cross-attention fusion has proven the most effective, achieving a macro F1 of 0.8357 on the official test set. The core contributions of our work are as follows:

- We have implemented a gated cross-attention fusion mechanism that has learned to balance text and image features for each meme individually.
- We have shown that Twitter-RoBERTa has outperformed larger general-purpose encoders, highlighting the value of domain-specific pretraining over raw model size.
- We have developed a multi-stage training pipeline incorporating weighted focal loss, a weighted ensemble and retraining on combined training and evaluation data.

Code is available at: https://github.com/Mif-taa/VaxMemeStance.

2 Related Work

Detecting harmful content in memes has become an increasingly important research problem as social media platforms have grown into primary channels for health-related discourse. Early approaches have tackled this problem using unimodal models, but these have consistently fallen short in capturing how the two modalities interact to construct meaning (Suryawanshi et al., 2020), motivating multimodal fusion strategies that have shown meaningful gains (Koutlis et al., 2023). Pretrained transformers such as BERT and RoBERTa have proven effective for social media text understanding (Devlin et al., 2019; Liu et al., 2019), while vision-language models like CLIP (Radford et al., 2021) and FLAVA (Singh et al., 2022) have pushed multimodal representation further. Twitter-RoBERTa (Barbieri et al., 2020) has demonstrated strong

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 133–140

performance on Twitter-oriented tasks through domain-specific pretraining and DINOv2 (Oquab et al., 2023) has expanded the toolkit for image feature extraction without text-image contrastive objectives. Most directly relevant to our work, Naseem et al. (2023) have developed a multimodal system for vaccine-critical meme detection on Twitter, providing the foundational benchmark for our work. Building on these works, our approach has addressed the challenge of capturing cross-modal interactions in memes by combining Twitter-RoBERTa, CLIP ViT-L/14 and OCR-extracted text, enabling more effective understanding of meme semantics for vaccine-critical detection.

3 Dataset & Task Description

This work has addressed the shared task on Multimodal Identification of Vaccine Content Stance on Social Media (Thapa et al., 2026b), organized as part of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA) (Hürriyetoğlu et al., 2026). The task has required classifying vaccine-related memes into one of three stance categories: Pro-Vaccine, Neutral and Vaccine-Critical, to support automatic detection of health misinformation and stance in multimodal social media content (Thapa et al., 2024, 2025).

The competition has been hosted on CodaBench[1] and participants have been evaluated on a held-out test set. The task has built upon prior work in multimodal vaccine-critical meme identification (Naseem et al., 2023) and concept-grounded detection of vaccine misinformation (Thapa et al., 2026a), with the annotation schema shared with the CrisisHateMM framework for hate speech analysis in crisis contexts (Bhandari et al., 2023).

3.1 The VaxMeme Dataset

The VaxMeme dataset has consisted of multimodal samples, each comprising a meme image, associated `post_text` and OCR-extracted `image_text`. The dataset has been drawn from prior WSDM and WebConf resources (Naseem et al., 2023; Thapa et al., 2026a) and has been partitioned into Train, Evaluation and Test splits. The label distribution across splits has been presented in Table 1.

The training set has contained 8,195 labelled memes, with a mild class imbalance toward the Pro-Vaccine category, followed by Vaccine-Critical and Neutral. The evaluation set has contained 1,024 labelled samples, while the test set has contained 1,025 labelled memes used for final leaderboard evaluation.

Labels	Train	Evaluation	Test
Vaccine-Critical	2,535	308	314
Neutral	2,461	327	316
Pro-Vaccine	3,199	389	395
Total	**8,195**	**1,024**	**1,025**

Table 1: Data distribution across Train, Evaluation, and Test sets.

4 System Overview

4.1 Problem Formulation

Each meme in the VaxMeme dataset has been assigned one of three stances: Vaccine-Critical (0), Neutral (1) or Pro-Vaccine (2). We have trained three classifiers and combined their predictions via weighted soft voting. The final label has been determined by the highest weighted average probability:

$$\hat{y} = \arg\max_{k} \sum_{i=1}^{3} w_i \, p_i(k \mid x), \quad w_i = \frac{f_i}{\sum_j f_j} \tag{1}$$

where f_i is the macro F1 score of model i, ensuring stronger models influence the final decision, especially for borderline Neutral cases.

4.2 Model Architecture

4.2.1 Text-Only Classifier

The text-only classifier has been built on the Twitter-RoBERTa model, a RoBERTa-base model pretrained on 124 million tweets and fine-tuned on sentiment, making it well-suited to the informal language of vaccine-related social media. A classification head has taken the average of the `[CLS]` token and the masked mean-pooled embeddings, has passed through dropout and a linear projection to three output classes:

$$\mathbf{h} = \frac{\mathbf{h}_{\text{CLS}} + \text{MeanPool}(\mathbf{H}, \mathbf{m})}{2}$$
$$\hat{p} = \text{softmax}(\mathbf{W}\,\text{Dropout}(\mathbf{h})) \tag{2}$$

where $\mathbf{H}$ is the full last hidden state and $\mathbf{m}$ is the attention mask.

4.2.2 Gated Multimodal Model

The multimodal model has fused the OCR-augmented text stream with a visual stream from

[1] https://www.codabench.org/competitions/12085/

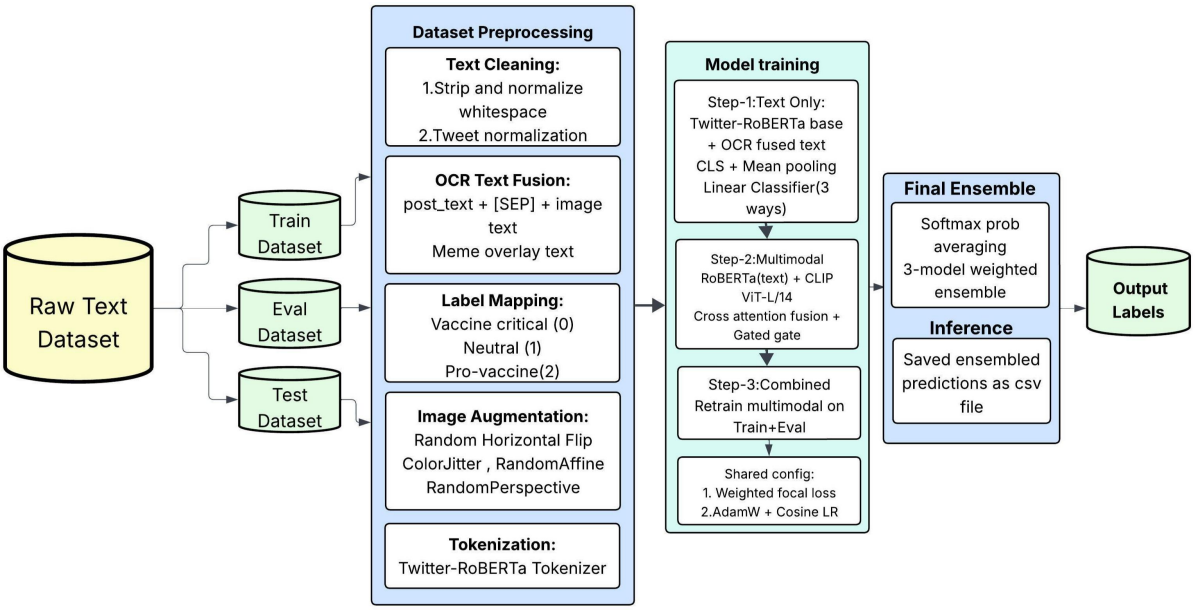

Figure 1: End-to-end pipeline of the 3-model ensemble for vaccine-meme detection.

the meme image. The text branch has reused the same Twitter-RoBERTa encoder and pooling strategy, while the visual branch has employed a CLIP ViT-L/14 encoder producing 768-dimensional image embeddings. For the first four epochs CLIP has been kept frozen to stabilise the fusion layers, after which CLIP parameters have been unfrozen with a conservative learning rate of 1×10^{-7} to allow gradual domain adaptation without catastrophic forgetting.

Both representations have been projected into a shared 512-dimensional fusion space via linear projections with LayerNorm and GELU activation. Cross-modal interaction has been modelled through 8-head cross-attention, where the text embedding has acted as query and the image embedding as key and value:

$$\mathbf{c} = \text{LayerNorm}\big(\text{CrossAttn}(\mathbf{q}_{\text{text}},\, \mathbf{k}_{\text{img}},\, \mathbf{v}_{\text{img}})\big) \tag{3}$$

A learned scalar gate $g \in (0, 1)$ has then blended the cross-attended representation with the raw text projection, allowing the model to down-weight the image when it has provided no additional discriminative signal:

$$\mathbf{z} = g \cdot \mathbf{c} \;+\; (1-g) \cdot \mathbf{h}_{\text{text}}, \quad g = \sigma\big(\mathbf{W}_g\,[\mathbf{h}_{\text{text}};\, \mathbf{c}]\big) \tag{4}$$

The fused representation $\mathbf{z}$ has been passed to a two-layer MLP classifier ($512 \to 256 \to 3$).

4.2.3 Combined-Data Model

The third model has been an additional instance of the gated multimodal architecture retrained on the union of the training and validation sets, exposing the model to every labelled example before test-time inference. Training has been run for a fixed 4 epochs with CLIP kept frozen throughout. Its logits have been included in the final ensemble using the validation F1 of the standalone multimodal model as a proxy weight, avoiding data leakage in the ensemble estimate.This proxy does not inflate ensemble performance because the standalone multimodal model's validation F1 was computed on a held-out evaluation set that was never used during combined-data training. The combined-data model thus contributes a conservatively weighted vote, and any overestimation of its weight would only marginally affect the final ensemble given that all three models produce similar probability distributions.

Figure 1 illustrates the pipeline.

5 Experimental Setup

5.1 Data Splits

The dataset has been divided into three splits: a training set of 8,195 samples, an evaluation set of 1,024 samples and a test set of 1,025 samples. During Stage 1 and Stage 2, the model has been trained on the training set and validated on the evaluation set. In Stage 3, the model has been retrained on the combined training and evaluation

set of 9,219 samples before running inference on the test set. The test set has been kept unseen throughout all training stages.

5.2 Preprocessing

5.2.1 Text

Each sample's textual content has been constructed by concatenating the post text with OCR-extracted meme overlay text separated by a special delimiter:

$$t = \texttt{post\_text} \parallel \texttt{[SEP]} \parallel \texttt{image\_text} \quad (5)$$

When no OCR text is available, only `post_text` has been used. This fusion has ensured that stance signals in meme overlays, like slogans and hashtags are preserved.

5.2.2 Image

Each meme image has been loaded, converted to RGB and preprocessed using the CLIP ViT-L/14 preprocessor, resizing and centre-cropping to 224×224 pixels with ImageNet normalization. During training, augmentations such as random flipping, colour jitter, affine transformation, and perspective distortion has been applied, with no augmentation during evaluation or inference. Missing images has been replaced with a zero tensor of the same shape.

5.3 Feature Extraction and Fusion

Textual features have been extracted using `cardiffnlp/twitter-roberta-base-sentiment-latest` with the final representation computed as the average of the `[CLS]` token and the mean-pooled token embeddings, both of dimensionality 768. Visual features have been extracted using the CLIP ViT-L/14 visual encoder, also producing 768-dimensional embeddings. Both representations have been mapped to a shared fusion space of dimension 512 via linear projection layers with LayerNorm, GELU activation, and dropout of 0.1.

Gated cross-modal attention has been used for fusion, where the text representation has served as the query and the image representation as the key and value in an 8-head multi-head attention layer. A learned gate has controlled the balance between the attended image features and the text features:

$$\mathbf{f} = g \cdot \mathbf{v}_{\text{cross}} + (1-g) \cdot \mathbf{f}_{\text{text}}, \quad g = \sigma\left(W[\mathbf{f}_{\text{text}}; \mathbf{v}_{\text{cross}}]\right) \quad (6)$$

The fused representation has then been passed through a two-layer classification head ($512 \rightarrow 256 \rightarrow 3$) with GELU activation and dropout.

5.4 Training

Training has proceeded in three stages. In Stage 1, a text-only model (Twitter-RoBERTa) has been trained for 5 epochs on the training set. In Stage 2, the full multimodal model has been trained for 8 epochs, with the CLIP encoder kept frozen for the first 4 epochs and gradually unfrozen with a low learning rate of 1×10^{-7} thereafter. In Stage 3, the multimodal model has been retrained for 4 epochs on the combined training and evaluation set to make use of all labelled data before running inference on the test set.All models have been optimised with AdamW using a cosine learning rate schedule with 10% linear warmup and gradient clipping at 1.0. The text encoder has used 2×10^{-5}, fusion and classification layers 5×10^{-5} and the CLIP encoder once unfrozen 1×10^{-7}. To address class imbalance, all models have been trained with a Weighted Focal Loss:

$$\mathcal{L} = -\sum_{k} w_k \, (1 - p_k)^{\gamma} \log p_k, \quad \gamma = 2 \quad (7)$$

The Neutral class has been up-weighted by 1.4 while other classes have been set to 1.0, with the focal term concentrating gradient on hard misclassified examples.

5.5 Ensemble and Inference

At inference, the best validation checkpoints of the text-only, multimodal and combined models have been used to produce softmax probability vectors. These have been blended as per Equation (1) and the `argmax` of the weighted average has determined the predicted stance label.

5.6 Parameter Setting

Table 2 lists the key hyperparameters for training the multimodal systems. All models have used the AdamW optimizer with a text encoder learning rate of 2e-5, cosine scheduling, and gradient clipping at 1.0, ensuring stable training and reducing overfitting.

5.7 Tools and Reproducibility

All experiments have been implemented in PyTorch using the Hugging Face Transformers library. A fixed random seed of 42 has been set across Python, NumPy and PyTorch for reproducibility. Training has been conducted on a GPU with mixed-precision via `torch.amp` and a batch size of 8.

Model	Text LR	Head LR	Optimizer	Batch Size	Epochs
RoBERTa-base + CLIP ViT-B/32	2e-5	5e-5	AdamW	16	5+6
BERT-base + CLIP ViT-B/32	2e-5	2e-4	AdamW	16	5+6
Twitter-RoBERTa + CLIP ViT-L/14	**2e-5**	**5e-5**	**AdamW**	**8**	**5+8+4**

Table 2: Key hyperparameters for multimodal model training.

5.8 Evaluation Metrics

All models have been evaluated using macro-averaged F1, precision and recall, treating each class equally regardless of frequency. The best-performing validation checkpoint has been retained for each model.

6 Results and Discussion

6.1 Task: Multimodal Vaccine-Critical Meme Detection

Table 3 has presented the performance of all models on the VaxMeme dataset. Among text-only models, RoBERTa-Large has achieved the highest macro F1 of 0.8165, followed by Twitter-RoBERTa at 0.8125 and RoBERTa-base at 0.8009. Image-only models have underperformed compared to text-only models, with CLIP ViT-B/32 achieving the best visual F1 of 0.7479, ahead of CLIP ViT-L/14 at 0.7179 and DINOv2-base at 0.7034.

Among multimodal models, Twitter-RoBERTa fused with CLIP ViT-L/14 has achieved the highest multimodal F1 of 0.8079 and accuracy of 0.8105. The weighted ensemble has further improved this to a macro F1 of 0.8090 and accuracy of 0.8115, with a final test F1 of 0.8357. To address classification challenges with the Neutral class, a weighted focal loss with a 1.4× Neutral upweight has been applied.

We acknowledge that Table 3 does not isolate the gate's individual contribution from the choice of ViT-L/14 backbone or the cross-attention structure itself. A standard cross-attention baseline (same backbone, no gate) was not included due to compute constraints, and we leave this ablation to future work. However, the gate's theoretical motivation dynamically suppressing weak visual signals is supported by the LIME and Integrated Gradients analyses in Section 8, which confirm that the model attends to task-relevant cues rather than uniform image features.

7 Error Analysis

7.1 Confusion Matrix

Figure 2 has shown the confusion matrix for our best system (Twitter-RoBERTa + CLIP ViT-L/14). The model has correctly identified 237 Vaccine-Critical, 258 Neutral and 336 Pro-Vaccine samples. The Vaccine-Critical class has had the most misclassifications, with 61 samples misclassified as Neutral and 10 as Pro-Vaccine.

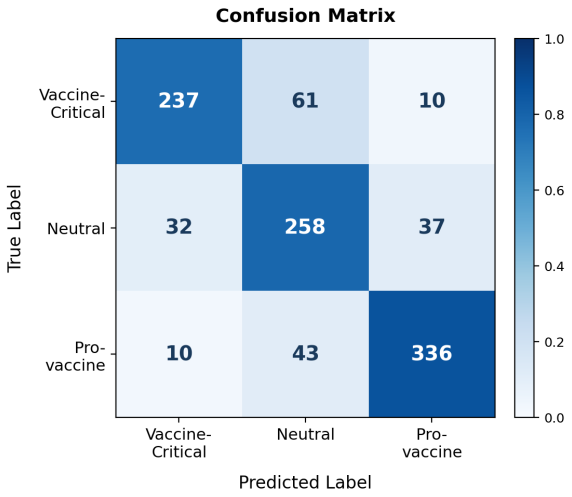

Figure 2: Confusion matrix for Twitter-RoBERTa + CLIP ViT-L/14

7.2 Failure Cases

An analysis of the misclassified samples has shown a clear pattern of errors across all three classes. The most difficult cases have involved the Vaccine-Critical class, where 61 samples have been misclassified as Neutral and 10 as Pro-Vaccine. This has suggested that the model has struggled to tell apart vaccine-critical content from ambiguous or sarcastic memes that look neutral on the surface.

The Neutral class has also shown confusion in both directions, with 32 samples misclassified as Vaccine-Critical and 37 as Pro-Vaccine. This has in-

Model	Classifier	P	R	F1	Acc
Unimodal (Text)	RoBERTa-base	0.8019	0.8003	0.8009	0.8047
Unimodal (Text)	Twitter-RoBERTa-base	0.8121	0.8138	0.8125	0.8145
Unimodal (Text)	RoBERTa-large	0.8174	0.8159	0.8165	0.8200
Unimodal (Image)	DINOv2-base	0.7033	0.7045	0.7034	0.7000
Unimodal (Image)	CLIP ViT-L/14	0.7191	0.7204	0.7179	0.7200
Unimodal (Image)	CLIP ViT-B/32	0.7496	0.7499	0.7479	0.7400
Multimodal	RoBERTa-base + CLIP ViT-B/32	0.7845	0.7834	0.7838	0.7881
Multimodal	BERT-base + CLIP ViT-B/32	0.7933	0.7926	0.7924	0.7969
Multimodal	Twitter-RoBERTa + CLIP ViT-L/14	0.8146	0.8058	0.8079	0.8105
Ensemble	**Twitter-RoBERTa + CLIP ViT-L/14**	**0.8132**	**0.8074**	**0.8090**	**0.8115**

Table 3: Performance comparison of unimodal, multimodal and ensemble classifiers on the validation set.

dicated that neutral memes have frequently shared image or text cues with both opposing stances, which has made boundary cases hard to classify correctly.

The Pro-Vaccine class has achieved the best recall of 86.38%, yet 43 samples have been misclassified as Neutral and 10 as Vaccine-Critical. These errors have mostly occurred in cases where pro-vaccine messaging has used mild or understated language, which has reduced the strength of the positive stance signal.

Overall, the confusion matrix has shown that most misclassifications have occurred between neighboring categories (Vaccine-Critical ↔ Neutral and Neutral ↔ Pro-Vaccine), rather than between the two opposing extremes (Vaccine-Critical ↔ Pro-Vaccine), which have accounted for only 20 of the 193 total misclassifications.

8 Explainability Analysis

We have applied two explainability techniques to our gated multimodal model for vaccine stance classification on the VaxMeme dataset:

- Integrated Gradients on token embeddings
- LIME superpixel explanations

Integrated Gradients has revealed token-level importance by computing attribution scores along the path from a baseline to the input embeddings. The method has consistently highlighted vaccine-critical keywords such as *antivax* and *side effect* as the strongest predictors, as has been shown in Figure 3. LIME has identified key visual regions by perturbing image superpixels while keeping the text input fixed, showing that symbolic imagery and text overlays have driven CLIP ViT-L/14 predictions, as has been shown in Figure 4. Together, both methods have confirmed that the model has learned task-relevant multimodal cues for vaccine stance classification.

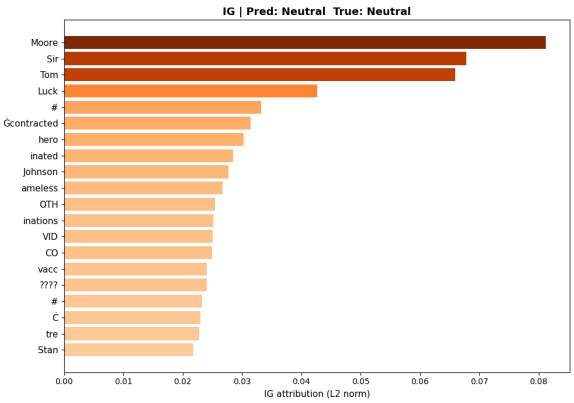

Figure 3: Integrated Gradients Token Attribution

9 Conclusion

Twitter-RoBERTa paired with CLIP ViT-L/14 under gated cross-modal attention has proven the strongest combination for vaccine-critical meme detection. Among the design choices, domain-specific pretraining, selective visual fusion and weighted focal loss have contributed the most to overall performance gains. The Neutral class has remained the most difficult category to classify correctly. Future work could explore larger vision-language models and multilingual training data to

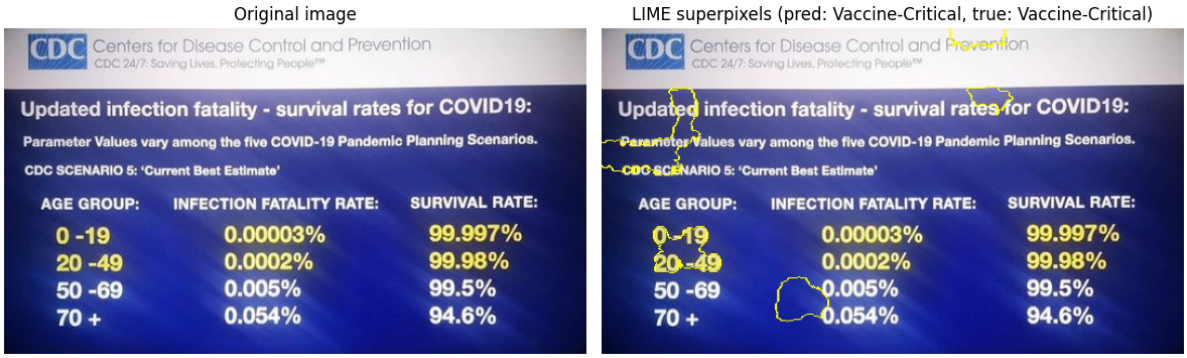

Figure 4: LIME Superpixel Importance Map

improve the system's robustness beyond English social media content.

Limitations

Despite strong overall results, the system has faced limitations due to CLIP's partial fine-tuning, which has caused the model to miss key visual cues in image-heavy memes. Furthermore, Twitter-RoBERTa has struggled with multilingual content and the Neutral class has remained the weakest performing category, which has reduced the system's applicability beyond Twitter-style content.

Ethical Considerations

All data has been drawn from the publicly available VaxMeme dataset and no personal data has been collected or stored. Our work has aimed to support public health monitoring and has been developed as a research prototype rather than a deployment-ready moderation tool. Any real-world use of this system should involve human oversight and domain expert review before application.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

Aashish Bhandari, Siddhant B Shah, Surendrabikram Thapa, Usman Naseem, and Mehwish Nasim. 2023. Crisishatemm: Multimodal analysis of directed and undirected hate speech in text-embedded images from russia-ukraine conflict. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1994–2003.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Ali Hürriyetoğlu, Surendrabikram Thapa, Hristo Tanev, Laxmi Thapa, and Surabhi Adhikari. 2026. Overview of the workshop on event extraction and understanding: Challenges and applications. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.

Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. MemeFier: Dual-stage modality fusion for image meme classification. *arXiv preprint arXiv:2304.02906*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Usman Naseem, Jinman Kim, Matloob Khushi, and Adam G. Dunn. 2023. A multimodal framework for the identification of vaccine critical memes on Twitter. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 706–714.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

Laxmi Thapa, Aryaman Jain, Lakshmojee Koduru, Surabhi Adhikari, Junaid Rashid, Jungeun Kim, Surendrabikram Thapa, and Usman Naseem. 2026a. Concept-grounded detection of vaccine misinformation in multimodal content using interpretable vision-language models. In *Companion Proceedings of the ACM on Web Conference 2026*.

Surendrabikram Thapa, Kritesh Rauniyar, Hariram Veeramani, Aditya Shah, Imran Razzak, and Usman Naseem. 2024. Did you tell a deadly lie? evaluating large language models for health misinformation identification. In *International Conference on Web Information Systems Engineering*, pages 391–405. Springer.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large language models (llm) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining*, 15(1):1–30.

Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Kritesh Rauniyar, Laxmi Thapa, Surabhi Adhikari, Kristina T. Johnson, Ali Hürriyetoğlu, Hristo Tanev, and Usman Naseem. 2026b. Multimodal identification of vaccine content stance on social media. In *Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications (EEUCA)*.