Quasar@EEUCA 2026: Multimodal Deep Learning for Vaccine Stance Detection in Memes

Adiba Fairooz Chowdhury, MD. Sagor Chowdhury
Department of Computer Science and Engineering
Chittagong University of Engineering and Technology, Bangladesh
{u2004014, u2004010}@student.cuet.ac.bd

Abstract

Vaccine stance detection in multimodal memes has emerged as an important yet challenging task, requiring models to interpret both textual and visual cues that jointly convey opinions. The difficulty lies in capturing subtle semantic interactions and handling class imbalance across stance categories. In this paper, we present our system developed for the VaxMeme 2026 Shared Task at EEUCA 2026. Our approach leverages a soft-voting ensemble of complementary models, combining DeBERTa-v3-large and RoBERTa-large for rich textual representation with CLIP (ViT-B/32) for joint vision-language understanding. We incorporate domain-specific preprocessing, techniques such as random token deletion, image enhancement, and balanced class oversampling to address dataset limitations. Through extensive ablation studies, we identify balanced class oversampling as the most effective component, significantly improving performance across models. Our final system achieves a macro F1-score of 0.8306, securing 8th place among 25 teams, demonstrating the effectiveness of ensemble-based multimodal learning for stance detection.

1 Introduction

Vaccine hesitancy has emerged as a critical public health challenge, with social media playing a significant role in the spread of both pro-vaccine advocacy and vaccine misinformation (Sallam, 2021). Among the many forms of health-related content on social media, memes have proven particularly influential: they combine image and text into a single communicative act, exploiting sarcasm, irony, and cultural reference to encode stances that neither modality alone reveals (Kiela et al., 2020). The multimodal nature of memes makes automatic stance detection substantially harder than text-only misinformation detection, and has direct relevance to public health (Thapa et al., 2024).

The VaxMeme 2026 Shared Task (Thapa et al., 2026b), organised as part of the EEUCA 2026 workshop (Hürriyetoğlu et al., 2026), which focuses on event extraction and understanding challenges has introduced a benchmark for this problem. The task requires systems to classify English-language vaccine memes from the VaxMeme dataset (Naseem et al., 2023; Thapa et al., 2026a; Bhandari et al., 2023) into three stance categories: *vaccine-critical*, *neutral*, and *pro-vaccine*. With 8,195 memes and moderate class imbalance, the dataset presents challenges for both model selection and training strategy.

To address this task, we have developed a multimodal ensemble system. We have systematically evaluated text-only models (TF-IDF, BERT, RoBERTa variants, DeBERTa variants), image-only models (ResNet-50, ViT, Swin, ConvNeXt, EfficientNet, CLIP Vision), and multimodal models (CLIP, BLIP, LLaVA), applying domain-specific text preprocessing, random token deletion augmentation, image enhancement, and class balancing via data augmentation. Our experiments show that balancing the dataset through augmentation is the single most impactful intervention, boosting even simple TF-IDF models by approximately 4.3 macro F1 points, while the specific choice of augmentation strategy has minimal effect. Furthermore, CLIP multimodal with image enhancement and balanced training data outperforms all text-only models. Our final soft-voting ensemble of DeBERTa-v3-large, RoBERTa-large, and CLIP multimodal achieves a macro F1 of 0.8306, placing us 8th out of 25 participating teams.

The main contributions of this work are:

- We have conducted a comprehensive comparison of text-only, image-only, and multimodal architectures across multiple preprocessing and augmentation configurations, providing a systematic ablation of what helps for vaccine

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 122–132

stance detection in memes.

- We have shown that addressing class imbalance is the most impactful single intervention, improving all model families substantially and implicating class imbalance as a primary bottleneck.
- We have shown that random token deletion generally matches or slightly outperforms synonym replacement as a text augmentation strategy for the short, domain-specific vocabulary of vaccine memes.
- We have developed a domain-adapted preprocessing pipeline for meme text that deliberately preserves stance-relevant informal signals such as emoji, hashtags, and repeated characters.

Further implementation details will be available via our code repository.[1]

2 Background

The VaxMeme 2026 Shared Task (Thapa et al., 2026b) presents a three-class stance classification problem: given a meme consisting of an image and its OCR-extracted text overlay, a system must assign one of three labels—vaccine-critical, neutral, or pro-vaccine. For example, a meme showing a syringe with the caption "they want to inject you with poison" would be labelled vaccine-critical, while one showing a vaccination queue with "protecting our community" would be pro-vaccine; a neutral meme might present statistics without implicit endorsement or criticism. Table 1 provides representative examples of multimodal inputs and their corresponding stance labels. The dataset (Naseem et al., 2023; Thapa et al., 2026a; Bhandari et al., 2023) contains 8,195 English-language social media memes with moderate class imbalance: pro-vaccine (39.0%), vaccine-critical (30.9%), and neutral (30.0%). For the official evaluation, 8,195 training samples, 1,024 validation samples (labels released at the test phase), and 1,025 unlabeled test samples are provided. In the development phase we use a stratified split without access to official validation labels: 5,736 train / 820 val / 1,639 test. The primary evaluation metric is macro F1, which equally weights performance across all three classes regardless of their frequency. We participated in the single track of this shared task.

[1] GitHub Repository

The task is organised as part of the EEUCA 2026 workshop (Hürriyetoğlu et al., 2026), which focuses on event extraction and understanding challenges. Meme classification has been studied extensively in the context of hate speech (Kiela et al., 2020), where multimodal models such as VisualBERT and UNITER demonstrated that jointly encoding image and text substantially outperforms unimodal approaches. The CrisisHateMM dataset (Bhandari et al., 2023) provides an annotation schema for multimodal hate content in social media images, which informs the VaxMeme annotation design. Vaccine misinformation detection on text has received considerable attention (Jennings et al., 2021; Hayawi et al., 2022; Thapa et al., 2024), and the VaxMeme dataset (Naseem et al., 2023; Thapa et al., 2026a) extends this line of work to multimodal meme-level stance classification. Thapa et al. (2025) survey the use of large language models in computational social science, providing broader context for NLP-based misinformation detection.

CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) provide state-of-the-art joint vision-language pretraining suitable for multimodal meme understanding. DeBERTa (He et al., 2021) introduces disentangled attention that separately encodes content and positional representations, yielding strong results on social media text. Prior ensemble work on social media classification (Cai et al., 2019) confirms that combining diverse model families reliably outperforms any single model.

3 System Overview

Figure 1 illustrates the overall system pipeline. Each meme is processed through parallel text and image pipelines. These pipelines perform preprocessing and optional data augmentation before extracting features using their respective models. The extracted features are then fed into three different model families, and their output probability distributions are combined via soft voting to produce the final predictions. Figure 2 shows the detailed architecture of the final ensemble, highlighting how features from different models are integrated to improve classification performance.

3.1 Preprocessing

Meme text is often noisy and informal. We apply a lightweight normalization pipeline consisting of: (1) whitespace trimming and normaliza-

Input			Output
Image	**Post Text**	**Image Text**	**Label**
	Unvaccinated is Sexy AF. https://t.co/SJQ6DpBh17	Unvaccinated MATTER	0
	@pauraenisciun Unvaccinated https://t.co/KMDvQ...	Bullying in 2022	1
	IM VAXXED YALL https://t.co/xfixZZV9oI	D	2

Table 1: Example memes from the VaxMeme dataset. The first three columns represent the multimodal inputs (image, post text, and OCR-extracted image text), while the final column is the stance label. Labels: 0 = Vaccine-critical, 1 = Neutral, 2 = Pro-vaccine.

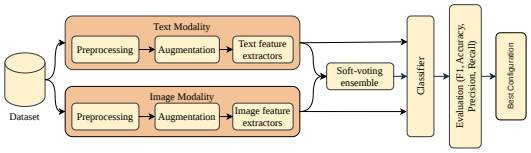

Figure 1: Overall system pipeline for processing memes through text and image models, with outputs combined via soft voting.

tion; (2) URL replacement with `[URL]`; (3) emoji demojization (e.g. `:syringe:`); (4) reduction of repeated characters; (5) hashtag normalization; (6) mention normalization; and (7) removal of non-alphanumeric characters except common punctuation. Although some steps (e.g., whitespace cleanup and emoji handling) have minimal effect on this dataset, they are retained for consistency and robustness. Importantly, normalization is conservative to preserve stance-indicative tokens such as hashtags and informal expressions. Table 2 shows representative examples of each step.

All images are uniformly resized to 224×224 pixels and enhanced using a three-step procedure: contrast adjustment ($\times 1.2$), brightness scaling ($\times 1.1$), and sharpness enhancement ($\times 1.1$). Figure 3 illustrates the enhancement effect on example memes.

3.2 Class Balancing and Augmentation

The original training set is moderately imbalanced: pro-vaccine 39.0%, vaccine-critical 30.9%, and neutral 30.0%. To address this, we perform class-balanced oversampling, increasing each class to a fixed target size (e.g., 3,200 or 4,000 samples per class depending on the experiment). Oversampling is implemented via sampling with replacement. For the duplicated samples only, we apply random token deletion as a text augmentation strategy. Specifically, each token is removed with probability $p = 0.1$, and augmentation is applied to a sample with probability 0.3. We intentionally avoid geometric augmentations (e.g., flipping, rotation, cropping), as preliminary experiments in our ablation study (Appendix A.4) showed a drop in validation F1 from 0.8140 to 0.8069 when these transformations were applied. This suggests that meme semantics depend heavily on layout and text placement, which such transformations may distort.

This approach preserves domain-specific vocabulary (e.g., vaccine names, slang) while introducing slight variability, making it more suitable than synonym replacement, which risks altering semantic

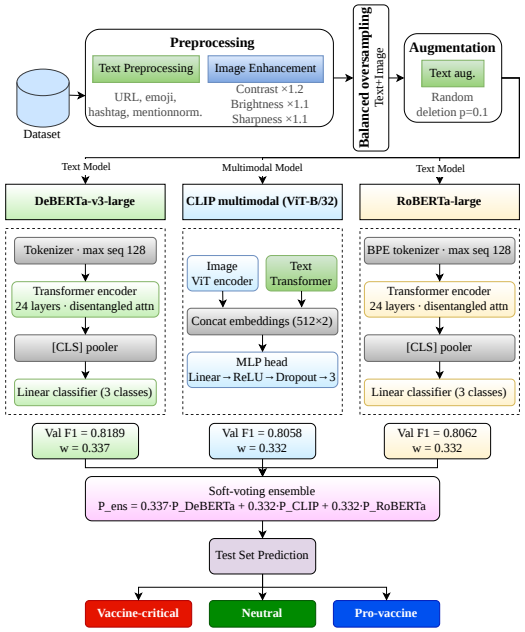

Figure 2: Final ensemble architecture: DeBERTa-v3-large, CLIP multimodal (ViT-B/32), and RoBERTa-large with soft voting. Weights are proportional to individual validation F1.

Before	Preprocessing	After
`Vaxxed!  Yass! https://t.co/xxx`	Normalise space	`Vaxxed! Yass! https://t.co/xxx`
`Vaxxed https://t.co/xxx`	Remove URL	`Vaxxed [URL]`
`vaxxed???????? [URL]`	Repeat char	`vaxxed?? [URL]`
`I got #vaxxed [URL]`	Hashtag	`I got hashtag_vaxxed [URL]`
`@UKvaxxed Vaxxed [URL]`	Mention	`mention_UKvaxxed Vaxxed [URL]`
`Vaxxed [URL]`	Clean	`Vaxxed URL`

Table 2: Preprocessing pipeline with before-and-after examples for each step.

Figure 3: Image enhancement examples: original (left) and enhanced (right).

meaning. Notably, augmentation is applied only to oversampled instances, while original samples remain unchanged.

Class balancing is the most impactful intervention in our pipeline. As shown in Table 10, even simple models such as TF-IDF classifiers benefit significantly (+4–5 macro F1 points), indicating that performance gains primarily stem from improved class distribution rather than augmentation alone. Figure 4 shows the class distribution before and after balancing.

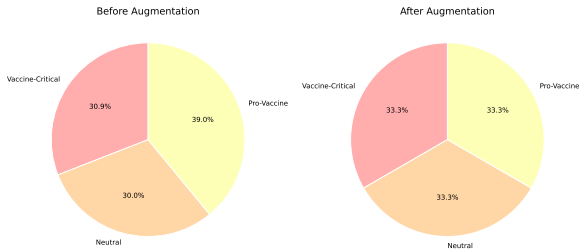

Figure 4: Class distribution before (left) and after (right) balancing to 4,000 samples per class.

3.3 Model Selection and Justification

We select three models for the final ensemble, each chosen based on ablation evidence:

DeBERTa-v3-large achieves the strongest text-only result in ablation (on the development split: 0.8107 without augmentation, 0.8110 with P+D). Its disentangled attention mechanism, which separately encodes content and positional information (He et al., 2021), is particularly suited to the short, stylised text of memes.

RoBERTa-large reaches 0.8130 with P+D augmentation on the development split, complementing DeBERTa through a different pretraining objective and corpus. Despite similar macro F1, the two models make different errors, making them effective ensemble partners.

CLIP multimodal (ViT-B/32) is the only model in our system that jointly encodes image and text. With image enhancement, balanced oversampling, text preprocessing and text deletion augmentation, it reaches macro F1=0.8502 on the development set—outperforming all text-only models—confirming that visual features carry stance signals not captured by text alone. BLIP was evaluated but plateaued at 0.7892 regardless of pipeline configuration, likely due to insufficient fine-tuning, and was excluded. We evaluated ViT-B/32 as it fits within the 16 GB memory budget of a single Kaggle P100; ViT-L/14 requires approximately 2× the GPU memory for the same batch size and was not feasible under our computational constraints.

3.4 Ensemble Strategy

We combine the three models via soft voting, averaging class probability vectors with weights proportional to validation macro F1:

$$P_{\text{ens}} = w_1 P_{\text{DeBERTa}} + w_2 P_{\text{CLIP}} + w_3 P_{\text{RoBERTa}} \quad (1)$$

where $w_1 = 0.337$, $w_2 = 0.332$, $w_3 = 0.332$. We explored majority voting and four-model configurations (adding TF-IDF and BLIP); none exceeded the three-model soft-voting result (Section 5).

3.5 End-to-End Inference Example

Figure 5 shows the inference pipeline for a representative sample (index 30): an image of a man holding a CDC COVID-19 Vaccination Record Card and the post text `Fully vaxxed w/Pfizer. https://t.co/OzvacyQTT9`. Text is preprocessed via MemeTextPreprocessor (URL replaced with

[URL], whitespace normalized, special characters removed), yielding Fully vaxxed wPfizer. URL". The 900×900 image is enhanced (contrast $\times 1.2$, brightness $\times 1.1$, sharpness $\times 1.1$) and resized to 224×224. DeBERTa-v3-large and RoBERTa-large encode the text, CLIP encodes image and text embeddings which are concatenated and passed through a 2-layer classification head. Weighted soft voting ($w = (0.337, 0.332, 0.332)$) combines outputs: DeBERTa predicts Neutral (0.638), CLIP and RoBERTa predict Pro-vaccine (0.968, 0.720), giving a final correct Pro-vaccine prediction (ensemble probability 0.611, computed from the full class probability vectors), showing how strong visual cues can override sparse textual input.

4 Experimental Setup

We use the official shared task splits described in Section 3. All models are trained on the full training set with balanced augmentation. Validation macro F1 guides model selection and weight calibration.

DeBERTa-v3-large and RoBERTa-large are fine-tuned with AdamW (lr=1e-5, weight decay=0.01), linear schedule with 10% warmup, batch size 8 (gradient accumulation steps 4 and 2 respectively), fp16, for 5 epochs with early stopping (patience=3 on val macro F1). CLIP uses AdamW (lr=2e-5), ReduceLROnPlateau scheduler, batch size 16, up to 10 epochs with early stopping (patience=3). Full hyperparameters are in Appendix B.

We use HuggingFace Transformers v4.40.0[2] for all transformer models, OpenAI CLIP[3] for the multimodal encoder, Salesforce BLIP[4] for the BLIP baseline, scikit-learn v1.4.2[5] for TF-IDF and logistic regression, and Pillow v10.3.0[6] for image enhancement.

All experiments use a single NVIDIA P100 (16 GB) via Kaggle.

Macro F1 is the primary metric, consistent with the shared task evaluation. We additionally report accuracy, per-class precision, recall, and F1, along with a confusion matrix for error analysis.

Formally, let $\mathbf{C} \in N^{K \times K}$ denote the confusion matrix, where C_{ij} is the number of samples with true class i predicted as class j. For class k, precision, recall, and F1 are defined as:

$$\text{Precision}_k = \frac{C_{kk}}{\sum_j C_{jk}}, \tag{2}$$

$$\text{Recall}_k = \frac{C_{kk}}{\sum_j C_{kj}}, \tag{3}$$

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}. \tag{4}$$

Macro F1 is computed as:

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} \text{F1}_k. \tag{5}$$

Overall accuracy is defined as:

$$\text{Accuracy} = \frac{\sum_k C_{kk}}{\sum_{i,j} C_{ij}}. \tag{6}$$

5 Results

5.1 Progressive System Development

Table 3 summarises our experiments as a progression from simple baselines to the final submission, grouped by model category. All results are on the official test set (1,025 samples) unless otherwise noted. Full per-model breakdowns are in Appendix D.

The results tell a clear story. Image-only models establish a ceiling of 0.7156, confirming text as the dominant modality. Plain text baselines without augmentation reach 0.8046 (TF-IDF+LogReg), which large pretrained transformers with augmentation push to 0.8252 (RoBERTa-large). Adding CLIP multimodal to the ensemble breaks the text-only ceiling, and increasing the augmentation target from 3,200 to 4,000 samples per class yields the final gain to 0.8306. Zero-shot LLaVA-1.5-13B (0.3447) confirms that general-purpose large vision-language models require task-specific fine-tuning for this domain.

5.2 Final Ensemble and Submission

Table 4 shows the individual validation scores of each component in the final ensemble. Weights are proportional to validation F1.

Adding TF-IDF or BLIP to the ensemble does not improve results: TF-IDF's lexical patterns are largely subsumed by the large transformers, and BLIP's contribution is already covered by CLIP.

[2] https://github.com/huggingface/transformers
[3] https://github.com/openai/CLIP
[4] https://github.com/salesforce/BLIP
[5] https://scikit-learn.org
[6] https://python-pillow.org

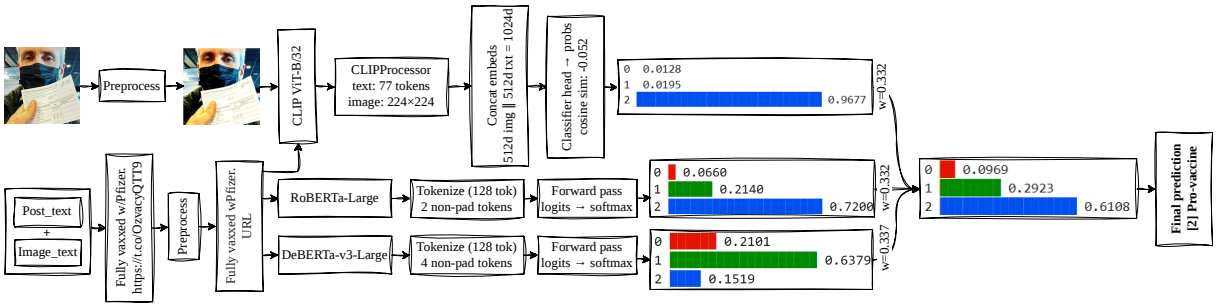

Figure 5: End-to-end inference for sample 30. The image and post text ("`Fully vaxxed w/Pfizer. [URL]`") are preprocessed separately and fed to DeBERTa, CLIP, and RoBERTa. Bars show class probabilities; weighted soft voting produces the final Pro-vaccine prediction, illustrating how visual signals can outweigh sparse textual cues.

Stage	Model / Configuration	F1
Image-only baselines		
	ResNet-50	0.6875
	ConvNeXt-base	0.6603
	CLIP Vision (image only)	0.7156
Text-only baselines (no aug)		
	TF-IDF + LogReg	0.8046
	BERT-base	0.7962
	RoBERTa-base	0.8028
Text models + preprocessing & aug		
	DeBERTa-v3-base	0.7967
	BERT-base	0.8088
	RoBERTa-base	0.8200
Large text models (+ preprocessing & aug)		
	XLM-RoBERTa-large	0.8026
	DeBERTa-v3-large	0.8146
	RoBERTa-large	**0.8252**
Multimodal		
	CLIP Multimodal (no aug)	0.7848
	CLIP Multimodal (aug=3200)	0.7957
	CLIP Multimodal (aug=4000)	0.8002
	LLaVA-1.5-13B (zero-shot)	0.3447
Ensemble configurations		
	RoBERTa + CLIP + TF-IDF	0.8252
	CLIP + RoBERTa + DeBERTa (3200)	0.8286
	RoBERTa + TF-IDF + CLIP	0.8299
	Stacking (+ BLIP)	0.8262
	CLIP + RoBERTa + DeBERTa (4000)	**0.8306**

Table 3: Progression of results from image-only baselines to the final submission. Bold = best in group.

5.3 Per-class Analysis and Error Analysis

Table 5 and Figure 6 show per-class results and the confusion matrix for the final ensemble on validation.

The neutral class is the hardest (F1=0.755), misclassified in both directions: 40 neutral memes are predicted vaccine-critical and 28 are predicted pro-vaccine. Neutral memes tend to present factual information without explicit stance markers, making them inherently ambiguous. Vaccine-critical memes have the highest recall (0.815), suggesting they carry strong distinctive lexical and visual signals. The test prediction distribution—vaccine-critical 30.6%, neutral 35.1%, pro-vaccine 34.2%—closely mirrors the training distribution, indicating no systematic class bias.

Model	Val F1	Acc.	Weight
DeBERTa-v3-large	0.8189	0.8213	0.337
CLIP Multimodal (ViT-B/32)	0.8058	0.8086	0.332
RoBERTa-large	0.8062	0.8076	0.332
Ensemble (submitted)	**0.8140**	**0.8154**	—

Table 4: Final ensemble components, validation scores, and soft-voting weights. Test set macro F1 = **0.8306**.

Class	Prec.	Recall	F1
Vaccine-critical	0.826	0.815	0.820
Neutral	0.721	0.792	0.755
Pro-vaccine	0.900	0.836	0.867
Macro avg	0.816	0.814	**0.814**

Table 5: Per-class validation results for the final ensemble.

6 Discussion

Balanced class oversampling emerges as the most impactful intervention in our study. While text-only models outperform CLIP individually (RoBERTa-large: 0.8062 val, DeBERTa-v3-large: 0.8189 val vs. CLIP: 0.8058 val), CLIP still improves ensemble performance. To understand this, we analysed pairwise prediction agreement between models on the test set: CLIP agrees with DeBERTa on only 86.3% of test samples and with RoBERTa on 85.8%, whereas the two text models agree with each other on 92.9%. Although DeBERTa and RoBERTa individually outperform CLIP on their respective disagreements, CLIP's lower correlation with both text models introduces sufficient diver-

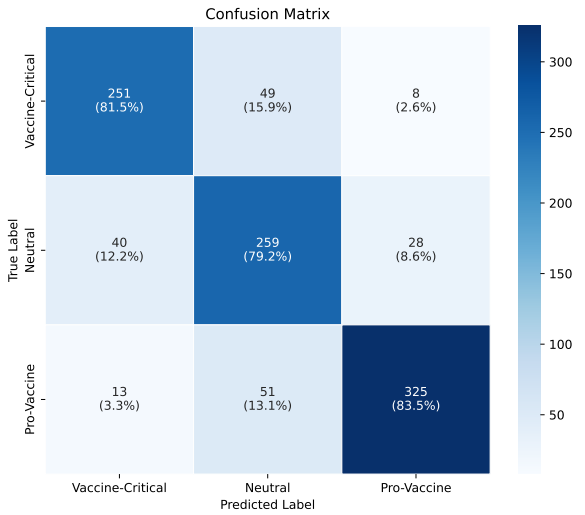

Figure 6: Confusion matrix for the final ensemble on the validation set (1,024 samples).

sity for soft voting to exploit—consistent with ensemble theory, where diversity among members is as important as individual accuracy. In terms of augmentation, random token deletion proves more effective than synonym replacement across all eight transformer models (Table 7), as synonym substitution risks altering domain-specific vocabulary such as vaccine names and slang, whereas deletion preserves the original distribution while introducing useful variation. Finally, large models such as DeBERTa-v3-large and RoBERTa-large exhibit early overfitting, peaking at epochs 2 and 3 respectively, with validation loss increasing thereafter, suggesting that the dataset size (8,195 samples) is limiting and that early stopping is crucial for stable performance. An experiment in which random crop, horizontal flip, colour jitter, and rotation were applied during training reduced ensemble validation F1 from 0.8140 to 0.8069 (Appendix A.4). We attribute this to the nature of meme images: meaning is encoded through composition and text placement, which geometric transforms disrupt. Image enhancement alone (contrast, brightness, sharpness) proved sufficient.

7 Conclusion

We presented a multimodal ensemble system for vaccine stance detection in memes, achieving 0.8306 macro F1 and ranking 8th out of 25 teams in the VaxMeme 2026 Shared Task (Thapa et al., 2026b). Our key finding is that balanced class oversampling is the most impactful intervention: it substantially boosts all model families and should be the first consideration on imbalanced meme datasets. A soft-voting ensemble of DeBERTa-v3-large, RoBERTa-large, and CLIP multimodal—trained with balanced oversampling and domain-adapted preprocessing—achieves competitive performance close to the top systems. Future work will explore cross-modal attention mechanisms and task-specific fine-tuning of larger vision-language models.

Limitations

Our system is trained on English-language COVID-19 vaccine memes; generalisation to other languages or vaccine contexts is unverified. Text deletion augmentation may introduce noise for very short meme texts. Computational constraints limited hyperparameter search to a single run per configuration.

Ethics Statement

All data consists of publicly available social media content. Stance detection tools carry a risk of misuse for content censorship; we advocate for their application in public health research and discourse analysis only.

Acknowledgments

We thank the VaxMeme 2026 shared task organisers for providing the dataset and evaluation infrastructure, and the EEUCA 2026 workshop chairs for hosting the competition. We also thank Kaggle for providing the computational resources used in this work.

References

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.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Kheir Hayawi, Sakib Shahriar, Mohamed Adel Serhani, Ikbal Taleb, and Sujith Samuel Mathew. 2022. ANTi-Vax: A novel twitter dataset for COVID-19 vaccine misinformation detection. *Public Health*, 203:23–30.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

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)*.

Will Jennings, Gerry Stoker, Hannah Bunting, Viktor Orri Valgarßson, Jennifer Gaskell, Daniel Devine, Lawrence McKay, and Melinda C. Mills. 2021. Lack of trust, conspiracy beliefs, and social media use predict COVID-19 vaccine hesitancy. *Vaccines*, 9(6):593.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Malik Sallam. 2021. COVID-19 vaccine hesitancy worldwide: a concise systematic review of vaccine acceptance rates. *Vaccines*, 9(2):160.

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)*.

A Ablation Study

All ablation results use a stratified development split: 5,736 train / 820 val / 1,639 test. Macro F1 is reported on the held-out test partition throughout.

A.1 Simple Baselines

Model	Macro F1
ResNet-50 (image only)	0.6518
Late Fusion (BERT+ResNet)	0.7041
TF-IDF + SVM	0.7919
BERT-base	0.7949
TF-IDF + LogReg	0.7995
RoBERTa-base	0.7959
CLIP Multimodal	0.7871
Enhanced TF-IDF	0.8017
Super Ensemble	**0.8075**

Table 6: Initial baselines without any preprocessing or augmentation.

A.2 Text Transformer Ablation

Model	None	Prep	Syn	Del	P+D
Base-sized models					
BERT-base	0.7912	0.8014	0.7938	0.7955	0.7998
RoBERTa-base	0.7990	0.7894	0.7888	0.7894	0.7795
Twitter-RoBERTa	0.7985	0.7988	0.7990	0.8012	0.7997
XLM-RoBERTa-base	0.7945	0.7912	0.7956	0.7889	0.7844
DeBERTa-v3-base	0.8058	0.8062	0.7993	0.8050	0.7966
Large models					
RoBERTa-large	0.7970	0.8026	0.8031	0.8050	0.8130
XLM-RoBERTa-large	0.7997	0.8014	0.7924	0.7975	0.8049
DeBERTa-v3-large	**0.8107**	0.8054	0.8001	0.8021	**0.8110**

Table 7: Text transformer macro F1 across all preprocessing and augmentation configurations (development split). Prep = text preprocessing; Syn = synonym augmentation; Del = random deletion; P+D = Prep+Del. Bold = best per group.

A.3 Image Enhancement and Oversampling

Table 8 reports image-only macro F1 before and after applying image enhancement and balanced oversampling.

Model	Baseline	w/ Enh+Oversample
ConvNeXt-base	0.6213	0.6964
EfficientNet-B3	0.6383	0.6753
ResNet-50	0.6518	0.7118
Swin-base	0.6670	0.7699
ViT-base	0.6672	**0.7834**
CLIP Vision	0.6915	0.7725

Table 8: Image-only macro F1 before and after image enhancement and balanced oversampling (development split).

A.4 Impact of Image Geometric Augmentation

We conducted a small ablation to evaluate the effect of geometric augmentation (flipping, cropping, rotation) on validation performance. Table 9 summarizes the results for our CLIP multimodal ensemble.

Pipeline	Val F1
Img Enh + Oversample only	0.8140
Img Enh + Oversample + Geometric Aug	0.8069

Table 9: Effect of geometric image augmentation on validation macro F1. Augmentation reduces performance, likely due to disruption of text layout and composition.

A.5 Effect of Balanced Class Distribution on TF-IDF

Table 10 reports macro F1 for TF-IDF models before and after balanced class distribution.

Model	Imbalanced	Balanced (Syn)	Balanced (Deletion)
TF-IDF + SVM	0.7919	0.8412	0.8423
TF-IDF + LogReg	0.7995	**0.8427**	0.8401

Table 10: TF-IDF macro F1 before and after balanced class distribution (3,200 per class) and with additional text deletion augmentation. Gains are due to class balancing alone.

A.6 Multimodal Development Results

Table 11 summarizes multimodal model results under different pipeline configurations.

Model	Pipeline	Macro F1
BLIP-base (5ep)	None	0.7892
CLIP Multimodal	None	0.7871
CLIP Multimodal (15ep)	Img Enh + Oversample	0.8415
BLIP-base (5ep)	Img Enh + Oversample	0.7892
CLIP Multimodal (15ep)	Img Enh + Oversample + Text Aug	**0.8502**

Table 11: Multimodal model results under different pipeline configurations (development split).

A.7 All Text Models (Main Evaluation)

Table 12 reports validation macro F1 for all text-only models with and without preprocessing and augmentation.

Model	w/o Pre+Aug	w/ Pre+Aug
Base-sized models		
BERT-base	0.7962	0.8088
DeBERTa-v3-base	0.8031	0.7967
RoBERTa-base	0.8028	0.8200
Twitter-RoBERTa	0.8092	0.8178
XLM-RoBERTa-base	0.8096	0.8136
4-model ensemble	0.8151	0.8165
Large models (w/ Pre+Aug only)		
XLM-RoBERTa-large	—	0.8026
DeBERTa-v3-large	—	0.8146
RoBERTa-large	—	**0.8252**
TF-IDF models		
TF-IDF + SVM	0.8003	0.7902
TF-IDF + LogReg	0.8046	0.7996

Table 12: All text-only model validation macro F1. Pre+Aug = text preprocessing + random deletion augmentation.

A.8 All Image and Multimodal Models (Main Evaluation)

Table 13 reports validation macro F1 for image-only and multimodal models.

Model	Aug Target	Val F1
ConvNeXt-base	—	0.6603
EfficientNet-B3	—	0.6622
CLIP Vision (image only)	—	0.7156
CLIP Multimodal	3200	0.7957
CLIP Multimodal	4000	**0.8002**
CLIP Multimodal (TTA)	4000	0.7856
LLaVA-1.5-13B (zero-shot)	—	0.3447

Table 13: Image-only and multimodal model validation macro F1. Aug Target = balanced samples per class.

B Training Hyperparameters

As shown in Table 14, CLIP multimodal (ViT-B/32) uses a 2-layer MLP classification head over concatenated image and text embeddings

Setting	DeBERTa-v3-large	RoBERTa-large
Learning rate	1e-5	1e-5
Batch size	8	8
Grad. accum. steps	4	2
Epochs	5	5
Best epoch (val F1)	2	3
Optimizer	AdamW, weight decay 0.01	
Scheduler	Linear decay, warmup ratio 0.1	
Max seq. len	128 tokens	
Precision	fp16	
Early stopping	Patience = 3 (val macro F1)	

Table 14: Transformer training hyperparameters for the final submission run.

(dim=512 $\times$ 2), trained with AdamW (lr=2e-5, weight decay=0.01) and a ReduceLROnPlateau scheduler (factor=0.5, patience=1). Batch size=16, up to 10 epochs with early stopping (patience=3 on val macro F1); best checkpoint at epoch 2 (val F1=0.8058). Images resized to 224 $\times$ 224 with enhancement (contrast $\times$1.2, brightness $\times$1.1, sharpness $\times$1.1). No geometric augmentation was applied. Cross-entropy loss with class-balanced weights. All models trained on a single NVIDIA P100 (16 GB) via Kaggle.

C Ensemble Weights

Final ensemble weights are proportional to validation macro F1: DeBERTa-v3-large: 0.337, CLIP Multimodal: 0.332, RoBERTa-large: 0.332 (normalised to sum to 1.0). Soft voting averages class probability vectors; majority voting was also evaluated and produced identical results on several configurations, suggesting the models largely agree.

D Full Test-Phase Results

All results below are on the official test set (1,025 samples) unless marked † (validation set). Models are grouped by family. This appendix covers all 65 experimental configurations run during the test phase.

D.1 Text-Only Models

Table 15 lists all text-only model results on the official test set.

D.2 Image-Only Models

Table 16 lists all image-only model results on the official test set.

D.3 Multimodal Models

Table 17 lists all multimodal model results on the official test set.

Model	Pre+Aug	F1	Acc.	Prec.
TF-IDF models				
TF-IDF + LogReg	No	0.8046	0.8068	0.8061
TF-IDF + SVM	No	0.8003	0.8020	0.8032
TF-IDF + LogReg	Yes	0.7996	0.8020	0.8016
TF-IDF + SVM	Yes	0.7902	0.7922	0.7943
Base-sized transformers (no Pre+Aug)				
BERT-base	No	0.7962	0.8000	0.7971
RoBERTa-base	No	0.8028	0.8059	0.8036
Twitter-RoBERTa	No	0.8092	0.8137	0.8088
XLM-RoBERTa-base	No	0.8096	0.8127	0.8120
DeBERTa-v3-base	No	0.8031	0.8059	0.8066
Base-sized transformers (with Pre+Aug)				
BERT-base	Yes	0.8088	0.8117	0.8101
RoBERTa-base	Yes	0.8200	0.8215	0.8232
Twitter-RoBERTa	Yes	0.8178	0.8205	0.8193
XLM-RoBERTa-base	Yes	0.8136	0.8156	0.8164
DeBERTa-v3-base	Yes	0.7967	0.8010	0.7978
Large models (with Pre+Aug)				
XLM-RoBERTa-large	Yes	0.8026	0.8049	0.8062
DeBERTa-v3-large	Yes	0.8146	0.8185	0.8147
BERT-base (large run)	Yes	0.8147	0.8156	0.8196
RoBERTa-large	Yes	0.8252	0.8293	0.8250

Table 15: All text-only model results on the test set.

Model	F1	Acc.	Prec.
ResNet-50	0.6875	0.6907	0.6904
ConvNeXt-base	0.6603	0.6663	0.6612
EfficientNet-B3	0.6622	0.6663	0.6671
CLIP Vision (image only)	0.7156	0.7220	0.7178
ResNet-50 (with enh+oversample)	0.6784	0.6810	0.6841

Table 16: All image-only model results on the test set.

D.4 Ensemble Configurations

Table 18 lists all ensemble configurations explored; the final submission is bolded.

Model / Configuration	F1	Acc.
CLIP Multimodal (baseline, no aug)	0.7848	0.7873
CLIP Multimodal (aug=3200)	0.7957	0.8000
CLIP Multimodal (aug=4200)	0.7849	0.7873
CLIP Multimodal (TTA)	0.7856	0.7912
CLIP Multimodal (aug=4000)	**0.8002**	**0.8020**
BLIP-base (5ep)†	0.7892	0.7906
LLaVA-1.5-13B (zero-shot)	0.3447	0.3971

Table 17: All multimodal model results. †Validation F1 (test set unlabeled).

Ensemble Configuration	F1	Acc.
RoBERTa + CLIP + TF-IDF (weighted)	0.8252	0.8293
RoBERTa + CLIP + TF-IDF (majority)	0.8252	0.8293
TF-IDF + CLIP + RoBERTa + DeBERTa	0.8221	0.8244
CLIP + RoBERTa + DeBERTa (aug=3200)	0.8286	0.8312
Stacking (CLIP+RoBERTa+DeBERTa+BLIP)	0.8262	0.8283
RoBERTa + TF-IDF + CLIP (majority)	0.8299	0.8312
RoBERTa + TF-IDF + CLIP (weighted)	0.8299	0.8312
CLIP + RoBERTa only	0.8132	0.8156
CLIP + RoBERTa + DeBERTa (aug=4000)	**0.8306**	**0.8322**

Table 18: All ensemble configurations explored. Final submission is bolded.