Linus@EEUCA 2026: Multimodal and Text-Only Approaches to Vaccine-Critical Meme Detection.

Darwin Acharya
Kathmandu University
acharyadarwin5@gmail.com

Shiv Ram Saud
Kathmandu University
saudshivram373@gmail.com

Sunil Regmi
Kathmandu University
sunil.regmi@ku.edu.np

Abstract

In this paper, we describe our participation in the Shared Task on Multimodal Identification of Vaccine Critical Content on Social Media (VaxMeme) of EEUCA 2026, a satellite of ACL 2026. We tackle the classification of Twitter-based vaccine memes into anti-vaccine, neutral, and pro-vaccine categories using the VaxMeme dataset with 8,195 train, 1,024 val, and 1,025 test samples. We experiment with two different architecture families: (i) Multimodal hybrids: CLIP ViT-B/32 for images + BERT-based models for texts (BERT-base-uncased, ModernBERT) with late fusion strategy based on concatenation of L2-normalized feature vectors and (ii) Text-only: pre-trained models for texts (BERT-base-uncased, RoBERTa-base, ModernBERT-base, DistilBERT-base, Deberta-v3-base) for post_text. In both cases, we use a three-layer feed-forward network with GELU activation for classification. We use class-weighted cross-entropy loss, differential learning rates, AdamW optimizer, gradient accumulation, OneCycleLR scheduler, and early stopping on the val set for optimization. Data augmentation is applied for the multimodal CLIP-based approach only. The winning approach among those tested is the text-only BERT-base-uncased with a macro-F1 of 0.8102 which is ahead of the performance of the CLIP + BERT-base hybrid model, which achieves a test macro-F1 of 0.7603.

1 Introduction

The rapid spread of health misinformation online poses significant challenges to public health, potentially leading to confusion, undermining trust in health authorities, and hindering effective health interventions (Thapa et al., 2024). The internet meme, defined by its concise and visually salient nature and its reliance on a combination of image and text for information and meaning, has risen to become a powerful and spreading vehicle for vaccine-critical information. Unlike plain text-based misinformation, which can often be addressed using conventional natural language processing techniques, the use of humor and irony in such memes makes them highly engaging and significantly more challenging to detect and mitigate automatically. The EEUCA 2026 shared task (Thapa et al., 2026b; Hürriyetoğlu et al., 2026), which is a satellite event of ACL 2026, involves the classification of vaccination-related memes into three types: anti-vaccine, neutral, and pro-vaccine. The evaluation is done based on macro-averaged F1 score. The task is based on the VaxMeme benchmark dataset (Naseem et al., 2023). Our task is part of the main track.

Our main strategy is based on the development of a classification framework that is modular and enables a comparison of unimodal and multimodal settings in a controlled and identical manner. Two different approaches are implemented: the first approach is a multimodal hybrid pipeline which is based on a late fusion multimodal approach, where visual and textual information are first encoded with separate transformer-based models, which are then concatenated for a classification task. We are using the ViT-B/32 encoder from clip (Radford et al., 2021) for the visual modality, which was trained with contrastive learning on image-text pairs and has shown excellent performance in multimodal reasoning tasks. For the textual modality, we experimented with two different encoders: BERT-base-uncased (Devlin et al., 2019), and ModernBERT (Warner et al., 2025) fusing the L2-normalized representations through late concatenation. The second part of the pipeline is a text-only pipeline, which consists of the fine-tuning of five pre-trained language models, namely BERT-base-uncased (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), ModernBERT-base (Warner et al., 2025), DistilBERT-base (Sanh et al., 2020), and DeBERTa-v3-base (He et al., 2023)—using only the post_text metadata field. Both pipelines utilize

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 223–232

the identical feedforward classification head, along with the identical training approach, thus making it possible to compare them directly.

Carrying out this task has led to some interesting findings. Firstly, the best-performing model, BERT-base-uncased with text-only input, achieves a test macro-F1 of 0.8102, outperforming the CLIP + BERT-base hybrid model, achieving 0.7603. This is unexpected, suggesting that the post_text tweet feature is a highly discriminative feature for vaccine stance classification, and CLIP image features, although individually useful (CLIP-only: 0.7189), do not seem to bring consistent improvements over powerful text encoders when used with late concatenation. Secondly, the RoBERTa-base model, when used with text-only input, achieves a competitive test macro-F1 of 0.8091, showing how effective the post_text feature is across models. Finally, from error analysis, it is clear that, across all models, sarcasm and implicit culture remain the main failure cases, especially when the intended stance depends on the nuanced interaction between the images and the text. Our top-performing model ranked 14th on the official CodaBench leaderboard with a macro-F1 score of 0.81. The code is available at: `https://github.com/sunilRegmi-ai/VCC-Social-Media`.

2 Background

The VaxMeme Shared Task (Thapa et al., 2026b; Hürriyetoğlu et al., 2026) is a three-class classification task that attempts to predict the stance expressed in a meme. Each data point is composed of a meme image in PNG format and post_text metadata corresponding to the tweet. The task is to predict one of three classes: 0 for anti-vaccine, 1 for neutral, and 2 for pro-vaccine. For example, a meme showing a syringe with the caption "CCP Virus Variant Affects Vaccinated People More Than Unvaccinated People: Study" is classified as anti-vaccine (class 0), while a meme showing a healthcare worker holding a vaccination record is classified as pro-vaccine (class 2). Memes that depict vaccination without expressing any opinion are classified as neutral. Evaluation is done using macro F1 score which assigns equal importance to each class.

The dataset used from VaxMeme consists of 10,244 English-language memes which are labeled as 0, 1, 2. The dataset has been split into official training, evaluation, and test sets. The training set comprises 8,195 samples, whereas the evaluation set has been used to provide a held-out validation partition.

This work was inspired by the extensive amount of prior work in the area of multimodal harmful content detection. For instance, the Hateful Memes Challenge (Kiela et al., 2021) highlighted the importance of multimodal reasoning in meme classification tasks and showed that it was possible to achieve state-of-the-art results by utilizing a model that excelled in each modality individually, albeit at a much lower level of performance than humans. Most recently, (Bhandari et al., 2023) proposed a multimodal dataset, CrisisHateMM, for detecting hate speech from text-embedded conflict images. This demonstrates the generalization of vision-linguistic approaches for detecting harmful content. In the context of vaccine misinformation, (Pramanick et al., 2021) presented MOMENTA, a framework that makes use of both global image and text features, as well as local entity-level representations and object attributes, in order to effectively capture the fine-grained semantic information that is necessary in meme analysis tasks. (Hayawi et al., 2022) also presented ANTi-Vax, a text-based dataset that targets vaccine misinformation related to COVID-19. In a further extension of the above studies, (Naseem et al., 2023) presented the VaxMeme dataset, which targets the multimodal setting of vaccine misinformation analysis. At the same time, (Thapa et al., 2026a) also proposed concept-grounded vision-language models for interpretable vaccine misinformation detection, offering an alternative approach to the classification-centric models used in this shared task. Unlike the aforementioned works, our work differs in the following aspects: (i) comparative evaluation of five encoders in a unified text-only setting, (ii) direct comparison of the proposed approaches with the multimodal CLIP fusion variants in the same setting, and (iii) the evaluation of the proposed approaches on recently proposed state-of-the-art models such as ModernBERT (Warner et al., 2025) and DeBERTa-v3 (He et al., 2023), which have not been previously evaluated on this benchmark.

3 System Overview

3.1 Fusion and Classification Head

All systems both unimodal and multimodal, use a classification head that is the same and acts on the

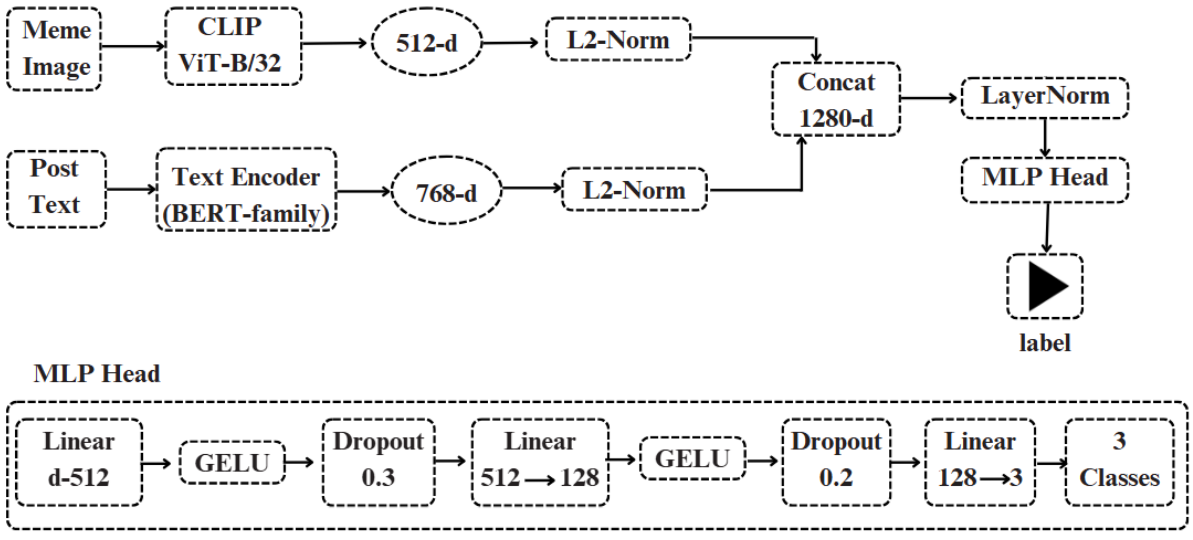

Figure 1: Architecture diagram showing hybrid system with MLP Head

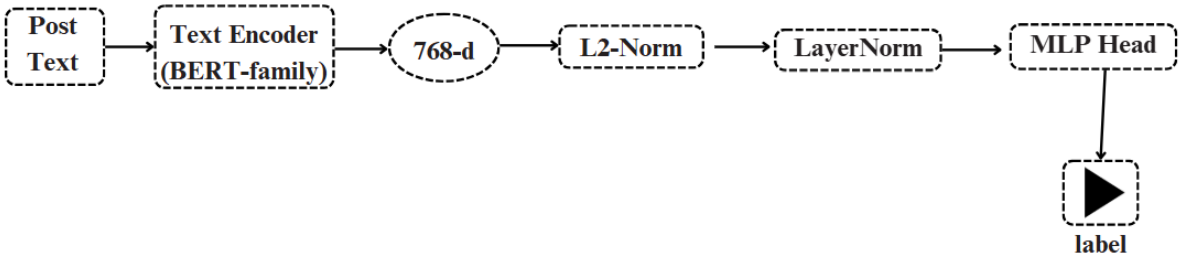

Figure 2: Architecture diagram showing text-only systems.

last feature representation. Let $z \in \mathbb{R}^d$ denote the input to the classification head.

In the multimodal (hybrid) setting, the L2-normalized visual feature vector $v \in \mathbb{R}^{512}$ and textual feature vector $t \in \mathbb{R}^{768}$ are concatenated to form a joint representation:

$$z = [v; t] \in \mathbb{R}^{1280}. \quad (1)$$

In the unimodal settings, z corresponds directly to the modality-specific representation:

$$z = \begin{cases} v \in \mathbb{R}^{512}, & \text{(image-only)} \\ t \in \mathbb{R}^{768}, & \text{(text-only)} \end{cases} \quad (2)$$

Thus, the input dimension d varies depending on the modality: $d = 512$ (image-only), $d = 768$ (text-only), and $d = 1280$ (hybrid).

Before classification, Layer Normalization is applied to z. The classification head is implemented as a three-layer feedforward network with GELU activations and dropout regularization:

$$\text{Linear}(d \to 512) \to \text{GELU} \to \text{Dropout}(0.3)$$
$$\text{Linear}(512 \to 128) \to \text{GELU} \to \text{Dropout}(0.2)$$
$$\text{Linear}(128 \to 3)$$

The last layer of the network generates logits for the three classes of stances. The classification head is randomly initialized and has a learning rate higher than the one used for training the pre-trained backbone encoders.

3.2 Multimodal Hybrid Systems (CLIP + Text Encoder)

The same late fusion multimodal framework is used by our hybrid systems. A visual encoder and a BERT-family textual encoder (discussed in Section 3.3) are used to get fixed-length feature vectors for the memes. The modality flag is used to specify the active encoders (image only, text only, or hybrid) and to specify if the normalized feature vectors need to be concatenated.

3.2.1 Visual Encoder: CLIP ViT-B/32

For all hybrid experiments, we use the visual subnetwork of CLIP ViT-B/32 (Radford et al., 2021) as the image encoder. CLIP is pre-trained via contrastive learning on 400 million web-scraped image-text pairs. This model is particularly beneficial for meme classification because the semantic meaning

of visual elements is often defined in relation to text.

In this work, we used the model via the OpenAI CLIP library, loading the pre-trained weights for ViT-B/32 with clip.load("ViT-B/32", jit=False) .

Each image is preprocessed through the default CLIP pipeline (resizing to 224x224 and normalizing through CLIP channel statistics). The visual encoder then produces a 512-dimensional vector. During training time, the images also go through a stochastic augmentation pipeline (as described in Section 3.5) before CLIP preprocessing. This model is cast to float32. During training, the whole visual encoder is fine-tuned end-to-end with a backbone learning rate of 2×10^{-6}. In the case of ModernBERT systems, a learning rate of 3×10^{-6} is used.

3.3 Text Encoders

We experiment with five text encoders from BERT-family to assess the effect of text model choice on multimodal meme and text-only systems classification performance.

3.3.1 BERT-base-uncased (Devlin et al., 2019)

The most commonly used transformer-based English language pre-trained model is trained using BooksCorpus and English Wikipedia datasets using a masked language and next sentence prediction task. The input is passed through a tokenizer and a maximum length of 128 is considered. The [CLS] token representation from the last hidden state is used for text representation, with a dimensionality of 768.

3.3.2 ModernBERT (Warner et al., 2025)

ModernBERT is a recent advancement in the BERT family of transformer-based language encoders that leverage FlashAttention, an alternating attention mechanism, and extensive pre-training on 2 trillion tokens using a context window of up to 8,192 tokens. In our experiment, we are using a variant of ModernBERT developed by answerdotai/ModernBERT-base variant. The hidden state of the last layer of the [CLS] symbol is used for text representation and has a dimensionality of 768. The ModernBERT model, along with the BERT-base model, has been utilized as a performance benchmark in all the experiments for both families of systems.

3.3.3 RoBERTa-base (Liu et al., 2019)

The RoBERTa-base is a 125 million parameter encoder that is based on the BERT-base architecture but employs a significantly enhanced pre-training procedure. This includes dynamic masking, where the masking is not static as in the original BERT, the removal of the next sentence prediction task, the use of ten times more data in the pre-training procedure (approximately 160 GB of data, as opposed to the 16 GB of the original BERT), and the employment of bigger batches and longer training steps. In fact, the RoBERTa-base outperforms the original BERT-base in all the standard natural language processing tasks. In our work, the RoBERTa-base is our primary strong text-only baseline, as our hypothesis is that the more and varied pre-training data, especially the web crawled material which is closer in style to the Twitter dataset due to its informal nature, might transfer particularly well to our meme post task.

3.3.4 DistilBERT-base-uncased (Sanh et al., 2020)

DistilBERT-base-uncased is a knowledge-distilled model with 66 million parameters that is a variant of BERT-base. It has been designed to reproduce the output distributions of BERT-base while reducing the number of parameters by 40%. It also reduces inference time by 60%. The knowledge distillation process uses a compound loss function that includes a loss from a masked language model (MLM), cosine embedding alignment from a BERT teacher model, and a soft cross-entropy loss over teacher logits. Nevertheless, DistilBERT has around 97% of the performance of BERT-base on the GLUE test set. It has been designed to be an efficiency baseline to calculate the performance cost of model compression for meme stance classification tasks and to check whether the reduced model performance has a significant effect on task performance.

3.3.5 Deberta-v3-base (He et al., 2023)

DeBERTa-v3-base is an encoder model with 86M parameters. It has two major improvements over BERT: "disentangled attention" unlike traditional BERT, disentangled attention calculates weight using distinct content and position vectors, which enables the model to more effectively understand token relationships at different positions, and "Enhanced Mask Decoding" (EMD), which is similar to ELECTRA and includes "replaced token detec-

tion." It is used in text-only systems as a text encoder.

For all encoders, the `post_text` field is tokenized with padding and truncation to a maximum of 128 tokens. The representation at the `[CLS]` position from the last hidden state (768-d) is extracted and ℓ_2-normalized as

$$\hat{\mathbf{f}} = \frac{\mathbf{f}}{\|\mathbf{f}\|_2 + \epsilon}, \quad \epsilon = 10^{-8}, \tag{3}$$

before being passed to the classification head.

3.4 Loss Function

We employ a weighted cross-entropy loss to address class imbalance in the training data. Let $y \in \{0, 1, 2\}$ denote the ground-truth label and $\hat{p} \in \mathbb{R}^3$ the predicted class probabilities. The loss is defined as:

$$\mathcal{L} = -\sum_{c=1}^{3} w_c \, y_c \log(\hat{p}_c), \tag{4}$$

where w_c denotes the class-specific weight, and $y_c \in \{0, 1\}$ is the one-hot indicator for class c.

Accordingly, the class weights $w = [1.2, 1.2, 1.0]$ associated with [anti-vaccine, neutral, pro-vaccine] respectively were defined based on the distribution of training set (pro-vaccine: 3,199; anti-vaccine: 2,535; neutral: 2,461). This weighting strategy boosts the influence of relatively infrequent classes in both training and evaluation, which helps address bias towards larger classes.

When training with gradient accumulation, the per-batch loss will be divided by the number of accumulation steps before backpropagation is executed to ensure correct gradient scaling.

3.5 Data Augmentation

To improve visual generalization, we apply a stochastic data augmentation pipeline to all training images prior to CLIP preprocessing. It is not applied to the text-only systems, for which no image data is processed. The augmentation pipeline consists of: (i) random resized cropping to 224×224 with a scale range of $(0.8, 1.0)$, ensuring that at least 80% of the original image content is retained; (ii) random horizontal flipping with a probability of 0.5; (iii) random rotation within $\pm 15°$; and (iv) color jittering with brightness, contrast, and saturation factors of 0.2.

No data augmentation is applied to validation or test images. The same augmentation pipeline is used across all hybrid model and image only model (CLIP Vit-B/32) configurations.

4 Experimental Setup

4.1 Data Splits

For model training, we use the official VaxMeme training partition, and for hyperparameter tuning and early stopping, we use the official evaluation partition as the development (validation) set. The test labels are not available, and evaluation is conducted via the official CodaBench submission system. No external data is used for training.

4.2 Hyperparameters

All models are trained using a batch size of 32 with gradient accumulation over 4 steps, yielding an effective weight update batch size of 128. The CLIP visual encoder and text backbone use AdamW for optimization with a learning rate of 2×10^{-6} (for BERT-base systems) or 3×10^{-6} (for ModernBERT systems) and a weight decay of 0.01.

The classification head is trained with a learning rate of 1×10^{-4} and weight decay of 0.05 for BERT-base systems, and a learning rate of 1.25×10^{-3} for ModernBERT systems.

We use a OneCycleLR scheduler with pct_start $= 0.2$, which performs linear warm-up over the first 20% of training steps and then cosine anneals to zero. Prior to each optimizer step we perform gradient clipping with maximum norm of 0.5.

We train for up to 50 epochs with early stopping based on validation macro-F1 and a patience of 5 epochs. The checkpoint with the best validation macro-F1 is selected for evaluation on the final test set.

4.3 Preprocessing

We load all images with the Pillow library, in RGB format. In training, each image is processed through data augmentation as described in Section 3.5, followed by CLIP-specific preprocessing, consisting of resizing images to 224x224 and normalization using CLIP channel statistics, whereas validation and test use CLIP preprocessing only. In case an image is either missing from the filesystem or corrupted, a zero tensor of shape $(3, 224, 224)$ is used as a filler for those images.

We tokenize the textual input (*post_text*) separately for all models using its respective tokenizer with padding and truncation to up to a maximum sequence length of 128 tokens. In particular, we use `padding="max_length"`, `truncation=True`, `max_length=128` and `return_tensors="pt"`. All inputs are explicitly cast to string type for robustness against missing or null data.

4.4 Evaluation Measures

The evaluation measure is based on the macro-averaged F1 score. This is calculated using the unweighted mean of F1 scores for individual classes. This ensures that equal weight is given to every class irrespective of their frequency. In addition to this, we also calculate macro-averaged precision and recall.

The evaluation is done using the `scikit-learn` library. Specifically, we have used `f1_score`, `precision_score`, and `recall_score` with `average="macro"` arguments. In addition, we have used `zero_division=0` to handle undefined values.

4.5 Tools and Libraries

The experiments use Python 3.10, PyTorch 2.x, and Hugging Face Transformers v4.40+. Other libraries include OpenAI CLIP for visual encoding, scikit-learn for evaluation, Pillow for images, and pandas for data handling.

For training, experiments use free Google Colab GPUs with CUDA. If multiple GPUs are present, data parallel training is performed via PyTorch `DataParallel`.

5 Results

Table 1 displays the entire results of all systems over the official validation and test sets. The top-performing systems over all metrics are the BERT-base-uncased text-only model with a test macro-F1 of 0.8102, which slightly outperforms the second-best-performing model, RoBERTa-base text-only model with 0.8091. Other top-performing systems are the Deberta-v3-base text-only model with 0.7950 and the DistilBERT-base text-only model with 0.7938. The top-performing hybrid systems are the CLIP + ModernBERT model with 0.7783, followed by the CLIP + BERT-base model with 0.7603. The performance of the CLIP-only baseline model is 0.7189. This shows that the textual post_text features are much more discriminative than image features alone for this dataset.

5.1 Text-Only vs. Multimodal Hybrid

One of the interesting observations from the results in Table 1 is the consistent outperformance of text-only systems over their multimodal hybrid counterparts, even when the same text encoder is used. For example, the text-only system using the BERT-base-uncased text encoder achieves 0.8102, which is 4.99 percentage points ahead of the CLIP + BERT-base hybrid system, which achieves 0.7603. In another case, the ModernBERT text-only system achieves 0.7797, which is almost comparable to the performance of the CLIP + ModernBERT hybrid system, which achieves 0.7783. The metadata field "post_text" describing the text of the tweet in which the meme was originally shared is seen to perform exceptionally well as a feature for vaccine stance classification. The addition of the CLIP features does not seem to provide any advantage over the text-only system, which is seen to be the best-performing configuration in our experiments. The reason for this might be the fact that the text of the tweet originally shared by the author of the meme actually contains the stance, which might not require the use of the image. Future work might involve the use of cross-attention fusion or the use of the text read directly from the embedded text in the images, i.e., OCR.

5.2 Comparison Across Text Encoders

Among the text-only models, BERT-base-uncased has the highest test macro-F1 of 0.8102. This is followed closely by RoBERTa-base with a test macro-F1 of 0.8091. These two models have high performance compared to ModernBERT-base with a test macro-F1 of 0.7797, DistilBERT-base with a test macro-F1 of 0.7938 and Deberta-v3-base with a test macro-F1 of 0.7950. The underperformance of ModernBERT-base compared to BERT-base is also noteworthy, especially considering that ModernBERT-base was designed with a more recent architecture and a pretraining corpus several orders of magnitude larger. This may be due to the nature of the post_text field, which consists of short-form social media text with fewer than 128 tokens. In this case, the large context window of up to 8,192 tokens and the complex attention mechanisms of ModernBERT-base do not provide any additional value and may pose optimization difficulties. The test macro-F1 score for DeBERTa-v3-base is 0.7950, ranking it in the third position among text-only models despite its superior bench-

Table 1: Results on the official validation and test sets. P = macro precision; R = macro recall. Val = validation set

System	Pipeline	Val F1	Val P	Val R	Test F1	Test P	Test R
BERT-base-uncased	Text-only	0.8166	0.8182	0.8162	0.8102	0.8126	0.8113
RoBERTa-base	Text-only	0.8140	0.8141	0.8158	0.8091	0.8121	0.8117
DistilBERT-base	Text-only	0.8140	0.8135	0.8149	0.7938	0.7941	0.7949
ModernBERT-base	Text-only	0.7923	0.7945	0.7955	0.7797	0.7819	0.7819
Deberta-v3-base	Text-only	0.8024	0.8020	0.8030	0.7950	0.7957	0.7964
CLIP + BERT-base	Hybrid	0.7604	0.7676	0.7587	0.7603	0.7682	0.7603
CLIP + ModernBERT	Hybrid	0.7791	0.7874	0.7785	0.7783	0.7882	0.7789
CLIP ViT-B/32	Image-only	0.7217	0.7223	0.7220	0.7189	0.7192	0.7190

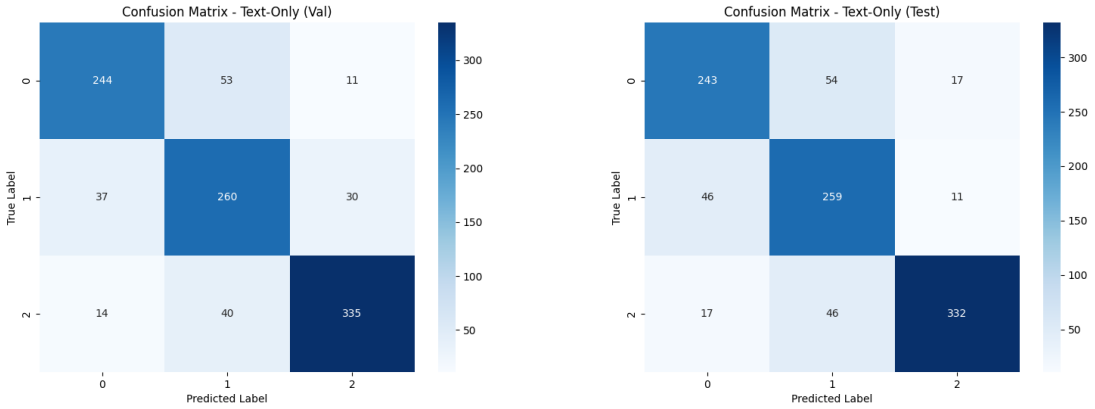

Figure 3: Confusion Matrices for text-only model BERT-base-uncased on validation and test sets.

mark performance on traditional NLP tasks. This shows that its disentangled attention mechanism provides little advantage in handling short informal text in tweets. The high performance of the RoBERTa-base model is also expected due to the pretraining of this model over a larger and more diverse corpus compared to BERT-base. The corpus used by RoBERTa-base also includes web-crawled text, which may be closer to Twitter text. The high validation F1 of 0.8140 and relatively lower test F1 (0.7938) by DistilBERT-base suggests some overfitting or train-test distribution mismatch, consistent with its reduced model capacity.

5.3 Error Analysis

The confusion matrices of the best-performing text-only system are presented in Figure 3 and its corresponding hybrid model's confusion matrices are presented in Figure 4, showing their performance on the validation and test sets. For both models, the neutral class, which corresponds to label 1, is the key challenge. For the test set, the hybrid system misclassifies 84 vaccine-critical and 57 pro-vaccine memes as neutral, whereas the text-only system misclassifies 54 and 46, respectively. Therefore, the text-only system is better for the neutral class. The pro-vaccine class, which corresponds to label 2, is the best for both models. For the hybrid system, the F1 score is 0.8548, whereas for the text-only system, the score is 0.8795, likely because pro-vaccine memes tend to have strong sentiment.

We manually analyzed the 100 cases of misclassification in the validation set of our best-performing model (BERT base uncased, text only). Three types of misclassifications were found.

i. **Sarcasm and ironic framing (42%):** Firstly, there are 42% of misclassified memes that express the vaccine-critical message with the use of sarcasm or irony. In these memes, words such as "Trust the science!" and "Safe and effective!" are used in an ironic way. However, the text encoder does not recognize the irony.

ii. **Implicit cultural references (31%):** Secondly, 31% of misclassifications are due to the meme implicitly referring to something, such as knowledge of a political figure, an in-

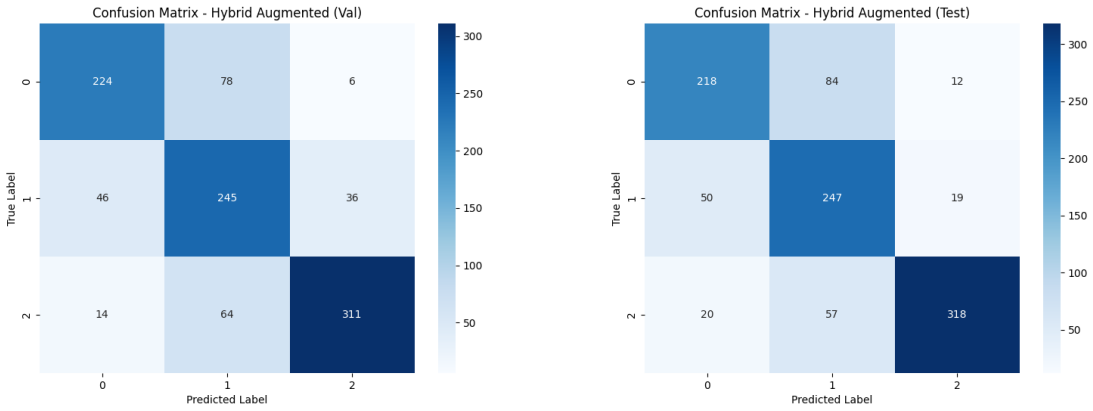

Figure 4: Confusion Matrices for hybrid model CLIP + BERT-base on validation and test sets.

ternet meme template, or a cultural event that cannot be inferred from the text alone. The model does not have enough world knowledge to comprehend the reference.

iii. **Neutral/vaccine-critical ambiguity (27%):** Lastly, 27% of misclassifications are due to the ambiguity of the meme, which represents a nuanced view that acknowledges both the benefits of vaccines as well as the concerns, thus placing it near the boundary of neutral and vaccine-critical in the original dataset.

6 Conclusion

We present our comprehensive system for the VaxMeme Shared Task at EEUCA 2026, comparing the performance of multimodal hybrid CLIP-fusion models with text-only fine-tuned language models for vaccine-critical meme stance classification. Our main discovery is that text-only models, especially the BERT-base-uncased model (0.8102 test macro F1), even outperform their multimodal counterparts when using the post_text tweet metadata signal. This suggests that the tweet text signal is highly discriminative for vaccine stance classification in this dataset. The CLIP model, using only image data, performs worse (0.7189) than text-based models, validating that vaccine stance cannot be reliably inferred from visual content alone. Common problems among configurations include sarcasm, culture, and ambiguity at the neutral/vaccine critical boundary.

Several avenues for enhancement have been identified. Firstly, the text extracted from the meme image using OCR can be leveraged as an additional visual-textual feature, different from post_text. Secondly, using cross-attention for fusion can improve the modeling of inter-modal relationships, thereby improving the understanding of sarcasm and implicit cultural cues. Thirdly, larger vision-language models, namely LLaVA and InstructBLIP, can improve the results for sarcasm detection and implicit cultural understanding. Lastly, using a retrieval model with an external knowledge base can improve the results by addressing the principal failure modes identified.

7 Limitations

- **Data dependency and overfitting:** The model heavily relies on *post_text*, which caused overfitting and weak generalization to visual inputs, as seen in the poor performance of image-only CLIP.
- **Simplistic multimodal fusion:** The fusion method is relatively simple and fails to capture complex relationships between text and images.
- **No image-text extraction:** The system does not extract text from images, missing important semantic information in memes.
- **Class ambiguity:** The model struggles with sarcasm, cultural context, and the distinction between neutral and vaccine-critical classes.
- **No external knowledge:** The approach does not use external knowledge or retrieval methods to resolve ambiguity.

8 Acknowledgments

We thank the organizers of the VaxMeme Shared Task at EEUCA 2026 (Thapa et al., 2026b; Hür-

riyetoğlu et al., 2026) for their efforts in creating the VaxMeme dataset and for running the shared task, along with maintaining the CodaBench platform. We also thank the anonymous reviewers for their helpful comments. This work was done without any external funding.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Kadhim Hayawi, Sakib Shahriar, Mohamed A. Serhani, Issam Taleb, and Sujith S. Mathew. 2022. Anti-vax: a novel twitter dataset for covid-19 vaccine misinformation detection. *Public Health*, 203:23–30.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

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)*.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

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. *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.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Preprint*, arXiv:2103.00020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

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, 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)*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

A Full Hyperparameter Table

Table 2 provides a complete listing of all hyperparameter settings used in each experimental pipeline to allow full reproduction of reported results.

Table 2: Hyperparameter settings for text-only and hybrid pipelines.

Hyperparameter	**Text-Only**	**Hybrid**
Batch size	32	32
Gradient accumulation	4	4
Effective batch size	128	128
Max epochs	50	50
Early stopping (patience)	5	5
LR — backbone	3×10^{-6}	2×10^{-6} (BERT), 3×10^{-6} (ModernBERT)
LR — classifier head	1×10^{-4}	1×10^{-4}
Weight decay (backbone / head)	0.01 / 0.05	0.01 / 0.05
Optimizer	AdamW	AdamW
Scheduler	OneCycleLR (pct_start = 0.2)	OneCycleLR (pct_start = 0.2)
Gradient clipping	0.5	0.5
Dropout (512 / 128 layer)	0.3 / 0.2	0.3 / 0.2
Max text tokens	128	128
Image size	N/A	224×224
Image augmentation	N/A	RandomResizedCrop, HFlip, Rotation ($\pm 15^{\circ}$), ColorJitter
Class weights [0,1,2]	[1.2, 1.2, 1.0]	[1.2, 1.2, 1.0]