CSECU-Learners@EEUCA 2026: Vaccine Critical Memes Identification using Two-Stage Early Fusion of Transformers

Monir Ahmad
Department of Computer Science and Engineering
University of Chittagong,
Chattogram-4331, Bangladesh
ahmad.csecu@gmail.com

Md. Saif Uddin
Department of Computer Science and Mathematics
Bangladesh Agricultural University,
Mymensingh-2202, Bangladesh
saifuddin.csm@bau.edu.bd

Abstract

Memes have emerged as a fast and influential way to share information online, particularly during major public health events like COVID-19 vaccination. While they can support awareness and encourage positive behavior, they are also widely used to spread misinformation and vaccine-critical views. These messages are often expressed through sarcasm and implicit meaning, which makes automatic detection difficult. To tackle this problem, EEUCA 2026 introduces a shared task based on the VaxMeme dataset for multimodal vaccine critical meme detection. The task encourages us to design models that can jointly understand both image and text, capturing the underlying context more effectively. In this work, we present our approach to this task by proposing a two-stage early fusion framework that integrates multiple transformer-based encoders. We train our model using focal loss to give more attention to difficult samples. Our experimental results show that our method performs competitively in the shared task, demonstrating its effectiveness for this problem.

1 Introduction

The rapid growth of social media has transformed how information is created, shared, and consumed, with memes emerging as one of the most influential forms of communication. Memes, which typically combine images and short textual elements, are highly engaging due to their humorous, sarcastic, and easily shareable nature (Pramanick et al., 2021a).

However, this same virality makes them a powerful vehicle for spreading misleading or harmful narratives (Wang et al., 2020). In the context of public health, vaccine-critical memes have become particularly concerning, as they can promote misinformation, reinforce vaccine hesitancy, and negatively influence public perception toward immunization efforts. Given the demonstrated relationship between online exposure and real-world attitudes, the automatic detection of such content is crucial for enabling timely interventions and supporting public awareness campaigns (Wang et al., 2020). The organizer of EEUCA 2026 (Hürriyetoğlu et al., 2026) proposes a shared task to support multimodal vaccine-critical meme detection (Thapa et al., 2026b). The task allows participants to develop models that jointly leverage both visual and textual representations to capture the global and local contextual cues embedded in memes.

Early research in this domain has predominantly focused on text-based analysis of social media content (Zhang et al., 2020; Naseem et al., 2021). These approaches leverage traditional machine learning models and, more recently, transformer-based architectures such as BERT (Devlin et al., 2019) to classify vaccine-related opinions and sentiments. While these methods have shown promising results, they are inherently limited in their ability to capture the full meaning of memes. To address these shortcomings, recent studies have explored multimodal approaches that jointly analyze textual and visual information (Pramanick et al., 2021b; Volkova et al., 2019). These methods have demonstrated improved performance across various tasks, including fake news detection, hateful meme identification, and misinformation analysis. Many models focus primarily on either global or local feature representations, without effectively combining both. Another critical limitation is the scarcity of publicly available, well-annotated multimodal datasets for vaccine critical content detection.

Our submitted system, CSECU-Learners addresses this challenge through a multi-encoder two-stage early fusion architecture. Specifically, we employ a Twitter-domain RoBERTa (Barbieri et al., 2020) to encode the textual content of each post, a Vision Transformer (ViT) (Dosovitskiy et al., 2020) to capture visual features from the

Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications, pages 169–176

meme image, and a Vision-and-Language Transformer (ViLT) (Kim et al., 2021) to obtain joint cross-modal representations. In Stage 1, the pooler outputs of RoBERTa and ViT are combined via performance-weighted summation. In Stage 2, this visual contextualized representation is concatenated with the ViLT pooler output that is passed to a linear classification layer. To mitigate the effect of class imbalance in the training corpus, we adopt Focal Loss (Lin et al., 2017).

We structure the remainder of this paper as follows. Section 2 introduces the proposed approach for the EEUCA-2026 multimodal vaccine-critical meme detection task. Section 3 outlines the experimental design, including implementation details and parameter settings. The results and their analysis are presented in Section 4. Finally, we conclude the paper by highlighting future research directions in Section 5, followed by a discussion of the limitations of the proposed method.

2 System Overview

This section provides an overview of our proposed system for the shared task on multimodal identification of vaccine critical content on social media at EEUCA 2026. Figure 1 presents a high-level illustration of our proposed system.

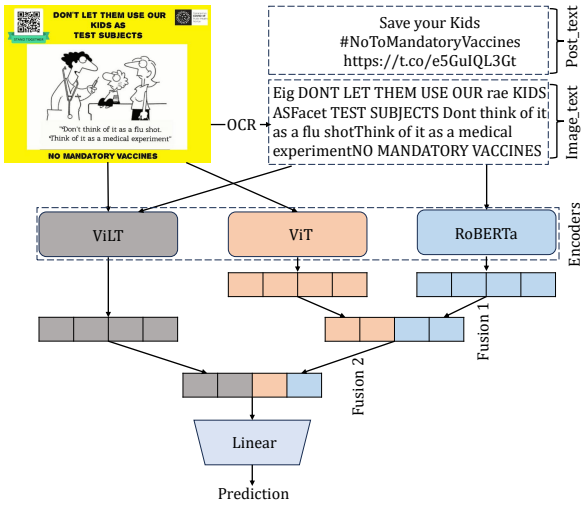

Figure 1: Overview diagram of our proposed system for EEUCA 2026: the shared task on multimodal identification of vaccine critical content on social media.

A social media post typically consists of two distinct information sources: the accompanying caption (post text) and the text embedded within the image itself (image text), the latter of which is recovered via Optical Character Recognition (OCR). We design a multi-encoder fusion architecture that jointly processes the visual and textual signals extracted from each post.

At the encoder stage, three pre-trained transformer-based models operate in parallel. Vision-and-Language Transformer (ViLT) (Kim et al., 2021) processes the raw image alongside its associated textual content, leveraging its native cross-modal attention to learn joint image–text representations. Vision Transformer (ViT) (Dosovitskiy et al., 2020), by contrast, focuses exclusively on the visual content of the post, while the Robustly Optimized BERT Pretraining Approach (RoBERTa) (Liu et al., 2019) handles the purely linguistic dimensions of the task.

We subsequently consolidate the representations produced by these three encoders through a two-stage fusion strategy. In the first fusion stage, the contextualized pooler output from ViT and the contextualized pooler output from RoBERTa are fused to form a combined multimodal representation. In the second fusion stage, this combined representation is merged with the holistic vision–language embedding produced by ViLT, yielding a unified feature vector that integrates all three perspectives. This final fused representation is passed through a linear classification layer, which produces logits (unnormalized scores) from which our model derives its final prediction.

2.1 Encoder Models

We fine-tune ViLT to obtain contextualized multimodal representations, employ ViT to capture visual information from the given image, and utilize RoBERTa to extract contextualized textual feature representations.

2.1.1 RoBERTa

For text encoding, we employ a RoBERTa-base model (Liu et al., 2019) fine-tuned on Twitter data[1]. It was originally introduced by Barbieri et al. (Barbieri et al., 2020) as part of the TweetEval benchmark. Unlike general-domain language models, this checkpoint was trained on roughly 58 million tweets, making it particularly sensitive to the informal grammar, hashtags, URLs, and emotionally charged phrasing that characterize vaccine-related discourse on social media. Given an input sequence, the model produces a contextualized

[1] https://huggingface.co/cardiffnlp/twitter-roberta-base

[CLS] pooler output, which serves as the textual representation fed into the fusion stage.

2.1.2 ViT

For visual encoding, we adopt the Vision Transformer (ViT) (Dosovitskiy et al., 2020) in its base configuration with 32×32 patch size, pre-trained on ImageNet-21k[2]. ViT treats an input image as a sequence of fixed-size non-overlapping patches, projects each patch into a linear embedding, and processes the resulting sequence through a standard transformer encoder. The larger patch size reduces sequence length and computational cost while retaining sufficient visual detail for meme-style social media images. It typically carries meaning through coarse layout and salient objects rather than fine-grained texture. The [CLS] pooler output of ViT is used as the visual representation at the first fusion stage.

2.1.3 ViLT

To capture cross-modal interactions directly at the encoder level, we incorporate the Vision-and-Language Transformer (ViLT) (Kim et al., 2021). We specifically utilize the base model pre-trained with masked language modeling on image–text pairs[3]. Unlike pipeline approaches that extract visual features with a separate object detector before cross-modal fusion, ViLT encodes image patches and text tokens within a single unified transformer. This approach enables direct attention between visual and linguistic elements at every layer. The model receives the raw post image together with the text as inputs, and its pooler output provides a holistic vision–language representation that complements the independently encoded visual and textual streams at the second fusion stage.

2.2 Two-Stage Early Fusion

Rather than relying on a single encoder to capture all modality-specific signals, we propose a two-stage early fusion strategy that progressively consolidates the representations from RoBERTa, ViT, and ViLT into a unified feature vector for downstream classification.

2.2.1 Stage 1: Weighted Fusion of Visual and Textual Representations

In the first stage, we combine the pooler outputs of RoBERTa and ViT through a performance-aware weighted summation followed by a $\tanh$ activation. Let $\mathbf{p}_R \in \mathbb{R}^{768}$ and $\mathbf{p}_V \in \mathbb{R}^{768}$ denote the pooler outputs of RoBERTa and ViT, respectively.

To determine the fusion weights, we rank the two encoders by their individual performance on the validation set, assigning an order number k to each model such that the better-performing model receives $k = 1$ (Du et al., 2022). Since RoBERTa outperforms ViT on the validation set, RoBERTa is assigned $k_R = 1$ and ViT is assigned $k_V = 2$. The weight for each encoder is then computed as:

$$w_k = \frac{1}{\sqrt{k}}, \quad k \in \{1, 2\} \tag{1}$$

This yields $w_R = 1/\sqrt{1} = 1.000$ for RoBERTa and $w_V = 1/\sqrt{2} \approx 0.707$ for ViT. The weighted sum is then computed as:

$$\mathbf{f}_1 = w_R \, \mathbf{p}_R + w_V \, \mathbf{p}_V \tag{2}$$

where $\mathbf{f}_1 \in \mathbb{R}^{768}$ is the resulting visual-contextualized representation that jointly encodes textual semantics and visual content while preserving the relative contribution of each encoder according to its predictive capability.

2.2.2 Stage 2: Fusion with Cross-Modal ViLT Representation

In the second stage, we incorporate the joint image–text representation produced by ViLT. Let $\mathbf{p}_L \in \mathbb{R}^{768}$ denote the pooler output of ViLT. A key limitation of ViLT is that it accepts a maximum input sequence length of 40 tokens. However, vaccine-critical content on social media is often considerably longer. Specifically, in the VaxMeme training set, approximately 63% of non-empty image texts and 72% of non-empty post texts exceed this 40-token limit when tokenized using the `cardiffnlp/twitter-roberta-base` tokenizer. Consequently, ViLT alone cannot fully encode the linguistic content of such posts, whereas the Stage 1 fusion — which employs RoBERTa — does not suffer from this constraint.

To complement the full-sequence textual encoding from Stage 1 with the cross-modal attention capability of ViLT, we concatenate $\mathbf{f}_1$ and $\mathbf{p}_L$ as follows:

$$\mathbf{f}_2 = \text{Concat}(\mathbf{f}_1, \, \mathbf{p}_L) \tag{3}$$

where $\mathbf{f}_2 \in \mathbb{R}^{1536}$ is the final fused representation obtained by concatenating the 768-dimensional visual-contextualized output from Stage 1 with the

[2] https://huggingface.co/google/vit-base-patch32-224-in21k

[3] https://huggingface.co/dandelin/vilt-b32-mlm

768-dimensional ViLT pooler output. This unified representation is subsequently passed to the linear classification layer for prediction.

2.3 Classification

The unified representation $\mathbf{f}_2 \in \mathbb{R}^{1536}$ obtained from Stage 2 fusion is fed into a single linear feed-forward layer that maps the fused embedding to the output space. Formally, the unnormalized class scores (logits) are computed as:

$$\hat{\mathbf{y}} = \mathbf{f}_2 \, \mathbf{W}^\top + \mathbf{b} \tag{4}$$

where $\mathbf{W} \in \mathbb{R}^{n \times d}$ is the weight matrix and $\mathbf{b} \in \mathbb{R}^n$ is the bias vector of the linear layer, with $d = 1536$ being the dimensionality of the fused representation from Stage 2 and n denoting the number of target classes. The final predicted class label $\hat{y}$ is determined by selecting the class corresponding to the maximum logit.

2.4 Focal Loss

Training on real-world social media datasets often involves skewed class distributions, where certain categories are substantially under-represented relative to others. Standard cross-entropy loss (Zhang and Sabuncu, 2018) tends to be dominated by the more frequent, easily classified examples, which can impede the model from learning discriminative patterns for minority or harder instances. To address this, we adopt Focal Loss (Lin et al., 2017).

Let t denote the index of the ground-truth class for a given input sample, and let $\hat{\mathbf{y}} \in \mathbb{R}^n$ be the logit vector. The predicted probability for the true class p_t is obtained via the softmax function. Focal loss addresses limitation of cross-entropy loss by augmenting the cross-entropy term with a modulating factor $(1 - p_t)^\gamma$:

$$\mathcal{L}_{\text{FL}}(p_t) = -(1 - p_t)^\gamma \log(p_t) \tag{5}$$

where $\gamma \geq 0$ is the focusing parameter that governs the rate at which well-classified examples are down-weighted. When $p_t \to 1$, the modulating factor $(1 - p_t)^\gamma \to 0$, effectively reducing the loss contribution of confidently correct predictions. Conversely, when the model misclassifies a sample and p_t remains small, the factor approaches unity, preserving the full loss signal for that instance.

3 Experimental setup

3.1 Dataset Overview

To evaluate our proposed framework on the shared task on multimodal identification of vaccine critical content on social media at EEUCA 2026, we use the annotated benchmark dataset provided by the task organizers (Thapa et al., 2026a). The dataset builds upon the VaxMeme corpus (Naseem et al., 2023), with an annotation schema shared with CrisisHateMM (Bhandari et al., 2023). Each instance in the dataset includes three components: the image, the text extracted from the image using OCR, and the accompanying post caption. In our approach, we utilize the image and the post text, while excluding the OCR-extracted image text, as it is often empty for most samples and its incorporation with the post text leads to performance degradation on the validation set. Each sample is assigned one of three class labels: *Vaccine Critical*, *Neutral*, or *Pro-vaccine*. The distribution of samples across the training, validation, and test splits is summarized in Table 1.

Class Label	Train	Validation	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: Distribution of data samples across splits in the shared task dataset.

The training set comprises 8,195 samples in total, with *Pro-vaccine* being the most frequent class at 3,199 instances (39.03%), followed by *Vaccine Critical* at 2,535 instances (30.93%) and *Neutral* at 2,461 instances (30.03%). A similar trend is observed in the validation set of 1,024 samples, where *Pro-vaccine* again constitutes the largest proportion at 389 instances (37.99%), with *Neutral* and *Vaccine Critical* accounting for 31.93% (327 instances) and 30.08% (308 instances), respectively. The test set contains 1,025 samples.

Across both the training and validation splits, the dataset exhibits a moderate degree of class imbalance. The *Pro-vaccine* class consistently outnumbers the other two categories, with a ratio of approximately 1.30:1 and 1.26:1 relative to *Vaccine Critical* and *Neutral* in the training set, respectively. While this imbalance is not extreme, it is sufficient to bias a naively trained classifier towards the majority class, potentially at the expense of cor-

rectly identifying *Vaccine Critical* content. This motivates our adoption of Focal Loss (Section 2.4), which mitigates the adverse effect of class imbalance by down-weighting the loss contribution of over-represented, easily classified samples during training.

3.2 Parameter Settings

This section describes the experimental setup for our submission to the shared task at EEUCA 2026. We fine-tune the Twitter-RoBERTa, ViT, and ViLT models available through the Hugging Face Transformers library (Wolf et al., 2020) using a Kaggle notebook[4] equipped with an NVIDIA Tesla T4 GPU. To ensure reproducibility across experimental runs, the random seed is fixed at 66 throughout all experiments. We employ the AdamW optimizer (Loshchilov and Hutter, 2017) for parameter updates, and model selection is performed by saving the checkpoint that achieves the best macro-averaged F1 score on the validation set. The optimal hyperparameter values identified through our experiments are summarized in Table 2. The maximum input sequence length is set to 256 tokens for RoBERTa. The focusing parameter of Focal Loss is set to $\gamma = 1$, which provides a mild emphasis on harder, misclassified samples.

Hyperparameter	Optimal Value
Train batch size	8
Test batch size	8
Learning rate	3e-5
Random seed	66
Max sequence length	256
Dropout probability	0.3
Number of train epochs	3
Focusing parameter (γ)	1

Table 2: Optimal hyperparameter configuration used in our experiments.

3.3 Evaluation Measures

To assess the effectiveness of the systems proposed by participants, the organizers use the macro-averaged F1 score (Sokolova and Lapalme, 2009) as the primary evaluation metric. This evaluation metric is well-suited for datasets with long-tail class distributions, as it treats all classes equally. By taking the harmonic mean of precision and recall for each class and then averaging the results, it offers a balanced view of overall model performance.

[4] https://www.kaggle.com

This formulation ensures that minority classes such as *Vaccine Critical* and *Neutral* contribute equally to the overall score as the majority *Pro-vaccine* class, penalizing systems that achieve high accuracy by predominantly predicting the dominant category.

4 Results and Analysis

In this section, we present and analyze the performance of our proposed CSECU-Learners system on the EEUCA 2026 shared task on multimodal identification of vaccine critical content on social media. The full training set is used to train our proposed model, while the validation set is reserved exclusively for hyperparameter tuning. The official evaluation metric is the macro-averaged F1 score, as described in the previous section.

4.1 Performance Comparison with Participating Systems

Table 3 summarizes the performance of our system alongside a selection of participating teams. Our CSECU-Learners system (Codabench username: anchy) achieved a macro-averaged F1 score of 0.8308 and an accuracy of 0.8341, securing 6th place among all participating teams. These results demonstrate that our proposed two-stage early fusion architecture, which consolidates complementary signals from RoBERTa, ViT, and ViLT, yields competitive performance on this multimodal classification task.

Among the top-ranked systems, *lili12* attains the highest macro-F1 of 0.8494, outperforming our system by a margin of 0.0186 points. The 2nd and 3rd ranked systems, *TIU-MI* and *CUET_Synthetica*, achieve macro-F1 scores of 0.8389 and 0.8357, respectively, placing them 0.0081 and 0.0049 points ahead of our submission. Notably, the performance gap between the 1st and 6th ranked systems is relatively narrow at approximately 1.86 percentage points, indicating that our framework operates within a highly competitive performance band at the upper end of the leaderboard. In contrast, the lower-ranked systems exhibit considerably weaker results. The 23rd, 24th, and 25th ranked teams *abs123*, *thatgrass*, and *kannanrrk* record macro-F1 scores of 0.7846, 0.7754, and 0.7436, respectively. Our system surpasses these by margins of 0.0462, 0.0554, and 0.0872 points. The ranking is not unique for each team. On the Codabench test phase leaderboard, we observe multiple entries un-

Team	Macro-F1	Accuracy	Precision	Recall	Rank
lili12	0.8494	0.8517	0.8494	0.8517	1st
TIU-MI	0.8389	0.8420	0.8386	0.8409	2nd
CUET_Synthetica	0.8357	0.8390	0.8383	0.8359	3rd
alexcristea72	0.8340	0.8380	0.8338	0.8351	4th
CUET_Synthetica	0.8332	0.8361	0.8345	0.8340	5th
CSECU-Learners (Ours)	**0.8308**	**0.8341**	**0.8309**	**0.8309**	**6th**
abs123	0.7846	0.7912	0.7868	0.7864	23rd
thatgrass	0.7754	0.7844	0.7858	0.7802	24th
kannanrrk	0.7436	0.7502	0.7435	0.7437	25th

Table 3: Comparative performance of selected systems on the EEUCA 2026 shared task. Our system is highlighted in bold.

der the same team. For example, CUET_Synthetica appears in both 3rd and 5th positions on the leaderboard.

4.2 Analysis of Different Modality Baseline Models

To motivate the design of our proposed multi-encoder fusion framework, we conduct a systematic baseline analysis across three modality categories: textual, visual, and multimodal. For the textual baseline, we adopt the Twitter RoBERTa model, which is particularly well-suited to this task, given that the underlying data originates from social media platforms with informal and hashtag-rich linguistic characteristics. As the visual baseline, we employ ViT, which has demonstrated strong performance across a broad range of image classification benchmarks. For the multimodal baseline, we utilize ViLT, which processes both image and text modalities within a single unified transformer, affording equal priority to visual and linguistic signals through its linear modality interaction mechanism. Each baseline model is evaluated independently on the validation set, and the results are reported in Table 4.

Method	Modality	Prec.	Rec.	Macro-F1
Twitter RoBERTa	Text	0.8084	0.8100	0.8082
ViT	Image	0.6991	0.6993	0.6973
ViLT	Multimodal	0.7718	0.7723	0.7716

Table 4: Performance of individual modality baseline models on the validation set. Here Prec. and Rec. indicate Precision and Recall metrics respectively.

Among the three baselines, Twitter RoBERTa attains the highest macro-F1 score of 0.8082, demonstrating that the textual content carries the most discriminative signal for identifying vaccine-critical content. This is consistent with the nature of the task, where vaccine critical stance is frequently expressed through explicit linguistic cues such as hashtags, emotionally charged phrasing, and misinformation-laden statements. ViT, operating solely on the visual modality, records the weakest performance with a macro-F1 of 0.6973, falling 11.09 percentage points below RoBERTa. This substantial gap suggests that visual features alone are insufficient for reliable vaccine-critical content identification.

ViLT, which jointly encodes image and text within a single transformer through cross-modal attention, achieves a macro-F1 of 0.7716 — surpassing ViT by 7.43 percentage points but falling 3.66 percentage points short of RoBERTa. This intermediate performance highlights both the benefit and the limitation of ViLT in this setting. While its multimodal design allows it to leverage visual-textual interactions, its restricted maximum sequence length of 40 tokens prevents it from fully encoding the often lengthy post captions this dataset, as discussed in Section 2.2.

4.3 Ablation Study

To quantify the individual contribution of each component in our proposed framework, we conduct an ablation study on the validation set by selectively disabling one component at a time while keeping the remaining components. The results are presented in Table 5.

Our proposed CSECU-Learners system achieves the highest macro-F1 of 0.8251, confirming that every component contributes positively to the overall performance. Replacing Focal Loss with standard cross-entropy loss reduces the macro-F1 from 0.8251 to 0.8216, a drop of 0.35 percentage points. It indicates that the class imbalance present in the training data, where *Pro-vaccine* samples constitute approximately 39% of the corpus. Focal Loss effectively mitigates this by suppressing the gradi-

Method	Macro-F1
CSECU-Learners	**0.8251**
−Focal Loss	0.8216
−Fusion 2	0.8180
−Fusion 1	0.7716

Table 5: Ablation study results on the validation set. Each row removes one component from the full system. −Focal Loss denotes training with standard cross-entropy loss; −Fusion 2 removes ViLT and retains only the weighted RoBERTa–ViT fusion; −Fusion 1 removes Stage 1 and relies solely on ViLT.

ent contribution of easily classified majority-class samples.

Disabling Stage 2 fusion, that is, discarding the ViLT pooler output and relying solely on the Stage 1 weighted combination of RoBERTa and ViT, results in a macro-F1 of 0.8180, a decline of 0.71 percentage points relative to the full system. This indicates that the cross-modal attention mechanism of ViLT contributes complementary vision–language interaction signals that the independently encoded RoBERTa and ViT representations alone cannot fully replicate.

The most pronounced degradation occurs when Stage 1 fusion is removed entirely, reducing the system to ViLT alone and yielding a macro-F1 of 0.7716, a drop of 5.35 percentage points relative to the full model. The substantial performance recovery achieved by incorporating Stage 1, which pairs RoBERTa's full-sequence textual encoding with ViT's visual representation. It addresses the ViLT's architectural sequence length limitation and captures the richer linguistic content present in vaccine-critical social media posts.

5 Conclusion and Future Direction

In this study, we tackle the problem of detecting vaccine-critical memes in a multimodal setting as part of the EEUCA 2026 shared task. We introduce a two-stage early fusion framework built on multiple transformer-based encoders. In the first stage, representations from RoBERTa and ViT are merged using a weighted summation guided by their relative performance. In the second stage, this fused representation is further integrated with the joint image–text embedding generated by ViLT. To better handle difficult samples and class imbalance during training, we adopt focal loss as the optimization objective. Experimental findings indicate that the proposed method achieves strong performance, demonstrating its capability in identifying vaccine-critical memes.

For future work, we plan to investigate advanced transformer models that are pre-trained on biomedical corpora, as they may provide more domain-relevant representations. Additionally, we aim to replace the fixed fusion formulation in Stage 1 with a learnable scalar or a lightweight attention-based gating mechanism, allowing the model to adaptively weight feature contributions and improve generalization across diverse datasets.

Limitations

Despite its effectiveness, our approach has several limitations. The use of multiple transformer-based encoders and their fusion increases computational cost, making the model relatively slow and resource-intensive. In this work, we rely on base-sized transformers; however, larger variants are known to achieve better performance in many tasks, which we did not investigate here. Moreover, the performance of the model is influenced by manual hyperparameter tuning, which may vary across different datasets and may not always generalize well to real-world applications.

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.

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 the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Xiyang Du, Dou Hu, Jin Zhi, Lianxin Jiang, and Xiaofeng Shi. 2022. Pali-nlp at semeval-2022 task 6: isarcasmeval-fine-tuning the pre-trained model for detecting intended sarcasm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 815–819.

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)*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Usman Naseem, Matloob Khushi, Jinman Kim, and Adam Dunn. 2021. Classifying vaccine sentiment tweets by modelling domain-specific representation and commonsense knowledge into context-aware attentive gru. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

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, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021a. Detecting harmful memes and their targets. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2783–2796.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021b. Momenta: A multimodal framework for detecting harmful memes and their targets. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 4439–4455.

Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437.

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, 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)*.

Svitlana Volkova, Ellyn Ayton, Dustin L Arendt, Zhuanyi Huang, and Brian Hutchinson. 2019. Explaining multimodal deceptive news prediction models. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 659–662.

Zuhui Wang, Zhaozheng Yin, and Young Anna Argyris. 2020. Detecting medical misinformation on social media using multimodal deep learning. *IEEE journal of biomedical and health informatics*, 25(6):2193–2203.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and 1 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Li Zhang, Haimeng Fan, Chengxia Peng, Guozheng Rao, and Qing Cong. 2020. Sentiment analysis methods for hpv vaccines related tweets based on transfer learning. In *Healthcare*, volume 8, page 307. MDPI.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.