# _alexcristea@EEUCA 2026: A Robust Early-Fusion ERNIE Pipeline for Multimodal COVID-19 Vaccine Meme Classification

**Cristea Alexandru-Marian**
**University of Bucharest**
**alexandru-marian.cristea@s.unibuc.ro**

**Ionescu Costin Ioan**
**University of Bucharest**
**costin-ioan.ionescu@s.unibuc.ro**

## Abstract

This paper presents our team's submission for the EEUCA 2026 shared task on Multimodal Vaccine Critical Meme Detection. To tackle the inherent challenges of internet sarcasm and high label noise, we propose a robust, heavily regularized early-fusion text pipeline. Bypassing computationally heavy visual encoders, we extract text directly from meme images via OCR, concatenate it with the user's social media post, and process the unified context through an ERNIE-2.0-Large encoder. To combat severe overfitting, we replace the standard classification head with a Multi-Sample Dropout architecture ($p = 0.3$). Our optimized, lightweight text-only pipeline achieved a peak Macro F1 score of 0.834, securing 4th place on the official leaderboard. Furthermore, an ablation study utilizing Focal Loss demonstrates that our primary solution using standard Cross-Entropy provides superior robustness against the inherent label noise found in internet meme datasets.

## 1 Introduction

In recent years, social media memes have become a primary vehicle for both public health communication and the spread of medical misinformation (Thapa et al., 2024). While memes can be used to promote awareness, they are also frequently used to spread vaccine skepticism and vaccine-critical narratives, often employing heavy sarcasm, irony, and culturally specific visual puns to subvert their literal text. This ambiguity makes automated stance detection highly susceptible to overfitting, as models tend to memorize surface-level lexical cues rather than learning generalized semantic representations.

The EEUCA 2026 shared task (Thapa et al., 2026b), held in conjunction with the EEUCA Workshop (Hürriyetoğlu et al., 2026), provides a targeted benchmark for this challenge. The overarching goal of the competition is to advance reliable systems for monitoring vaccine-related discourse and supporting myth-debunking efforts on social media platforms.

While previous state-of-the-art approaches to this task have relied on massive, multi-stream architectures combining Graph Neural Networks and deep image encoders, we propose a highly efficient early-fusion text architecture relying on the Enhanced Representation through Knowledge Integration (ERNIE) framework. By explicitly demarcating the original social media post from the embedded image text, we allow the transformer's self-attention mechanism to cross-reference contextual cues effectively without the latency of visual feature extraction. Aggressively regularized via Multi-Sample Dropout and Cross-Entropy loss, our pipeline successfully secured 4th place on the final leaderboard with an F1 score of 0.834.

## 2 Background

**Task Setup and Dataset:** The EEUCA 2026 shared task requires systems to process a multimodal input (a social media text post paired with an image) and predict a single output stance representing the post's attitude toward COVID-19 vaccines. For example, an input might consist of a user's post saying "They want us to take it" paired with an image of a grim reaper. The target output is a three-class classification: *Vaccine Critical*, *Neutral*, or *Pro-Vaccine*.

To train and evaluate systems, the shared task utilizes a curated dataset of over 10,000 multimodal social media posts (Thapa et al., 2026a; Naseem et al., 2023). The dataset consists entirely of English-language content originating from social media platforms. The genre is highly informal, characterized by internet slang, grammatical irregularities, and image-macro memes. The annotation schema for determining the stance relies on complex multimodal interactions, mirroring methodologies established in prior multimodal hate



*Proceedings of the 9th Workshop on Event Extraction and Understanding: Challenges and Applications*, pages 185–191

speech and crisis informatics research (Bhandari et al., 2023).

**Related Work:** Early work in automated stance detection primarily focused on text-only social media posts, consistently struggling with implicit sentiment, sarcasm, and irony. With the emergence of memes as a dominant form of internet communication, the field shifted toward multimodal approaches. Datasets mapping COVID-19 discourse have shown that vaccine-critical memes often rely on deep cultural context and dog-whistles rather than explicit anti-vaccine terminology (Naseem et al., 2023). Furthermore, while Large Language Models (LLMs) have shown immense promise in computational social science for capturing these nuanced social dynamics (Thapa et al., 2025), their application to non-compositional memes (where a benign image paired with a benign text creates a highly sarcastic message) remains challenging. Consequently, researchers frequently utilize massive Large Vision-Language Models (LVLMs) to capture cross-modal interactions. However, these dual-encoder systems often suffer from high computational costs and struggle to align disparate visual and textual feature spaces.

## 3 System Overview

We propose a highly regularized, early-fusion text pipeline that bypasses the computational overhead of dual-encoder LVLMs. The core algorithm relies on reducing the multimodal task into an advanced text-pair classification problem.

**Concrete Algorithmic Flow Example:** 1. *Input:* A user posts an image of a grim reaper holding a sign, accompanied by the social media caption: *"They want us all to take it."* 2. *Extraction:* We extract the text embedded within the image via OCR: *"COVID VACCINE WILL KILL YOU!"* 3. *Early Fusion:* The algorithm concatenates these texts into a single string separated by tokens: *"[CLS] They want us all to take it. [SEP] COVID VACCINE WILL KILL YOU! [SEP]"* 4. *Encoding:* ERNIE 2.0 processes this unified string, allowing its self-attention mechanism to instantly recognize the contextual contrast between the vague post and the aggressive image text. 5. *Prediction:* The pooled representation passes through our Multi-Sample Dropout head, predicting *"Vaccine Critical"*.

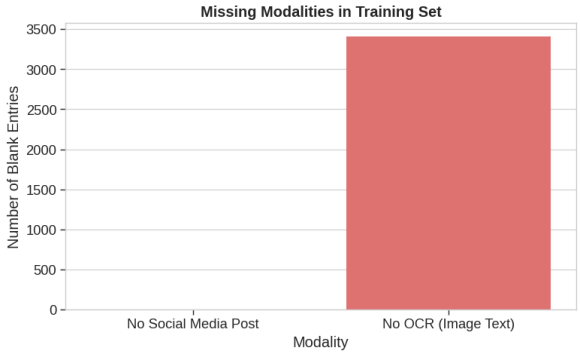



Figure 1: Missing modalities in the training set. A significant portion of the dataset lacks OCR text from the image, necessitating an early-fusion approach to avoid dropping critical contextual data.

### 3.1 Data Processing and Early Fusion

Because memes frequently rely on the juxtaposition between an image's embedded text and a user's accompanying caption, processing these modalities independently limits the model's ability to capture irony. Furthermore, as illustrated in Figure 1, a massive chunk of the dataset is missing either the social media post or the OCR image text entirely. To address this, we employ our early-fusion strategy.

For a given data instance, we define the social media caption as $T_{post}$ and the OCR-extracted text from the image as $T_{image}$. We concatenate these into a single sequence, separated by the special '[SEP]' token, allowing the Transformer's self-attention mechanism to cross-reference contextual cues between the two sources:

$$S = [\text{CLS}] \oplus T_{post} \oplus [\text{SEP}] \oplus T_{image} \oplus [\text{SEP}]$$

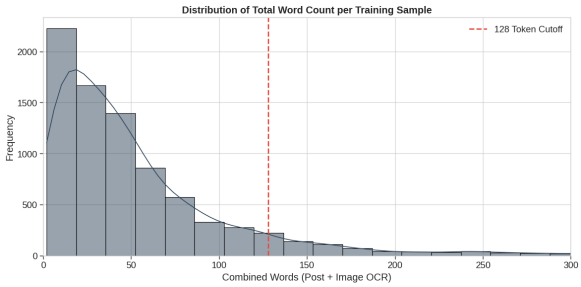



Figure 2: Distribution of the combined word count (post text + OCR image text). The 128-token cutoff captures the vast majority of instances without introducing unnecessary padding latency.

The combined sequence is tokenized using the ERNIE subword tokenizer. As shown in Figure 2, a maximum sequence length of 128 tokens perfectly

encapsulates the vast majority of the combined text sequences. Sequence lengths are strictly padded or truncated to this limit to accommodate the relatively short nature of meme text while maintaining optimal computational efficiency.

### 3.2 The ERNIE 2.0 Encoder

We utilize the pre-trained ERNIE 2.0-Large (Enhanced Representation through Knowledge Integration) model as our core representation learner (Sun et al., 2020). Unlike standard BERT, ERNIE is pre-trained via continual multi-task learning, explicitly capturing lexical, syntactic, and semantic information through entity-level masking. This makes it particularly adept at recognizing the named entities (e.g., vaccine names, political figures) heavily prevalent in COVID-19 discourse. We extract the hidden state of the '[CLS]' token from the final layer, $h \in \mathbb{R}^d$, where $d = 1024$, to serve as the aggregate representation of the meme.

### 3.3 Multi-Sample Dropout Classification Head

Standard Transformer classifiers utilize a single dropout layer preceding a linear transformation. However, due to the high variance and ambiguity in subjective meme datasets, a single dropout mask can inadvertently zero-out the specific neurons holding crucial "sarcasm" features for a given batch, leading to unstable gradient updates and poor generalization.

To combat this, we replace the standard head with a Multi-Sample Dropout architecture (Inoue, 2019). Instead of a single pass, the pooled representation $h$ is passed through $N$ parallel, independent dropout layers. The surviving vectors are all processed by the exact same fully connected linear classifier, and the resulting predictions are averaged to produce the final output logits:

$$\text{Logits} = \frac{1}{N} \sum_{i=1}^{N} W(Dropout_i(h)) + b$$

In our optimal configuration, we set $N = 5$ and apply a heavy dropout rate of $p = 0.3$. This architecture acts as an implicit ensemble within a single forward pass, aggressively regularizing the network and smoothing the loss landscape without introducing additional trainable parameters.

### 3.4 Loss Function and Optimization

As illustrated in Figure 3, the dataset exhibits a class imbalance, which naturally biases standard neural networks toward the majority "Pro-Vaccine" class. To mitigate this, we apply inverse class weighting to our loss functions. The weight for class $c$ is calculated as:

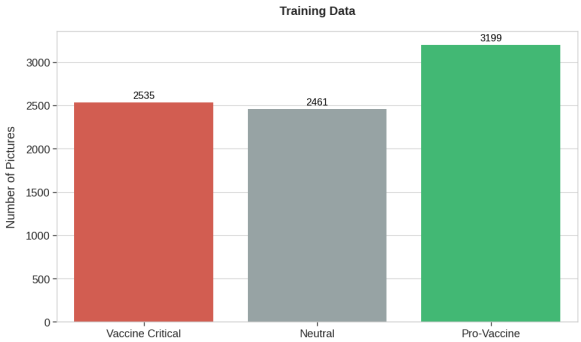



Figure 3: Class distribution of the training dataset, highlighting the natural class imbalance that necessitates our inverse-frequency weighted loss function.

$$w_c = \frac{N_{total}}{C \times N_c}$$

where $N_{total}$ is the total number of training samples, $C = 3$ is the number of classes, and $N_c$ is the number of samples in class $c$.

For our final solution, the model is optimized using the standard PyTorch Cross-Entropy Loss function, modulated by these class weights:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i)$$

To validate this choice, we also formulated an alternate configuration utilizing Weighted Focal Loss (Lin et al., 2017) to test if aggressive hard-example mining would yield better results (detailed in Section 4.2). We optimize the network using AdamW with a learning rate of $2 \times 10^{-5}$ and an increased weight decay of 0.05 to further penalize large weights. We employ a linear learning rate scheduler with a 10% warmup phase over 10 training epochs, utilizing mixed-precision training (FP16) with a gradient accumulation factor of 8 to achieve an effective batch size of 256.

## 4 Experimental Setup

### 4.1 Dataset Splits and Preprocessing

We evaluate our proposed pipeline on the official benchmark dataset provided for the EEUCA 2026 shared task (Naseem et al., 2023; Thapa et al., 2026a). The dataset consists of multimodal social media posts annotated for one of three stances:

Vaccine Critical, Neutral, and Pro-Vaccine. We strictly utilized the official shared task data splits: the training set was used for model optimization, the development (dev) set was used for hyperparameter tuning and ablation validation, and the hidden test set was exclusively used for the final leaderboard submission.

During preprocessing, we extracted the OCR text and social media caption, concatenating them using the standard '[SEP]' token. As qualitative analysis of the filtered vocabulary demonstrates, there are distinct lexical signatures for each class, confirming strong textual signals exist for our early-fusion pipeline to exploit. Based on text-length distributions, all sequences were truncated or padded to a maximum length of 128 tokens.

### 4.2 Implementation Details and Tools

Our pipeline was implemented using the PyTorch framework[1] and the HuggingFace Transformers library[2]. We initialized our text encoder using the pre-trained `ernie-2.0-large-en` weights.

To ensure the statistical validity of our findings and account for the high variance inherent in fine-tuning large language models on noisy data, every architectural variation in our ablation study was trained across three distinct random seeds (42, 22, and 100). The network was optimized using AdamW with a learning rate of $2 \times 10^{-5}$ and mixed-precision training (FP16). Due to space constraints, the complete list of hyperparameter configurations required to replicate our experiments is provided in Appendix A.

### 4.3 Evaluation Measures

Following the official EEUCA 2026 Codabench evaluation protocol, system performance is primarily measured using the Macro F1-score. The Macro F1 calculates the F1-score independently for each of the three classes and then computes their unweighted average. This ensures that performance on the minority classes contributes equally to the final score, preventing systems from artificially inflating their metrics by over-indexing on the majority class.

[1] https://pytorch.org, version 2.0.1

[2] https://huggingface.co/docs/transformers, version 4.30.0

## 5 Results

### 5.1 Main Quantitative Findings

Our final optimized pipeline (ERNIE 2.0 + Multi-Sample Dropout + Cross-Entropy) achieved a peak Macro F1 score of **0.8340** on the official hidden test set. This performance secured the **4th place** ranking on the final EEUCA 2026 competition leaderboard, demonstrating that a highly regularized, early-fusion text-only pipeline is highly competitive against complex, multi-stream vision-language baseline models.

### 5.2 Quantitative Analysis: Ablation Studies

To better understand our design decisions, we conducted several ablations evaluated on the development split (summarized in Table 1).

**The Impact of Multi-Sample Dropout:** Our first experiment evaluated the impact of the classification head. The baseline, utilizing a single 10% dropout layer, achieved a 3-seed average of 0.8136 on the dev split. By replacing this with our Multi-Sample Dropout architecture (five parallel masks at $p = 0.3$), the average Macro F1-score significantly increased. This confirms that creating an implicit ensemble of dropout masks prevents over-indexing on noisy, batch-specific features.

**Cross-Entropy vs. Focal Loss:** To validate our choice of the primary loss function, we conducted an ablation utilizing Weighted Focal Loss. Contrary to our initial hypothesis, our final solution using standard Cross-Entropy Loss slightly outperformed the Focal Loss ablation. Focal Loss aggressively penalizes misclassifications on "hard" examples. However, in the context of subjective internet sarcasm, these "hard" examples are frequently mislabeled outliers. By forcing the optimizer to over-index on these noisy data points, the Focal Loss model suffered a degradation in generalization. Therefore, standard Cross-Entropy provides the highest robustness against dataset noise.

**Cross-Lingual Knowledge Transfer:** To test the boundaries of text-only meme classification, we conducted an ablation utilizing ERNIE 3.0 Base, an architecture trained natively on Chinese corpora, by translating the English memes into Chinese. As expected, this cross-lingual pipeline suffered a catastrophic performance drop to 0.7240 Macro F1. This confirms that meme comprehension relies heavily on culturally specific slang and regional context, which machine translation actively strips away.

| Model Architecture | Macro F1-Score |
|---|---|
| Standard Head (Baseline) | $0.8136 \pm 0.0042$ |
| ERNIE 2.0 + MSD + Focal Loss (Ablation) | $0.8221 \pm 0.0028$ |
| **ERNIE 2.0 + MSD + Cross-Entropy (Final)** | $\mathbf{0.8238 \pm 0.0102}$ |
| ERNIE 3.0 (Translated Data Ablation) | $0.7240 \pm 0.0013$ |

Table 1: Macro F1-scores for our experimental models and ablation studies. Mean and standard deviation are calculated on the development split across three random seeds.

### 5.3 Error Analysis

To better understand the limitations of our text-only approach, we generated a confusion matrix based on the development split predictions (Figure 4). While the strong diagonal confirms the efficacy of our inverse class-weighting, the off-diagonal clusters reveal that our pipeline primarily struggles with two highly specific multimodal phenomena, illustrated in Figure 5.

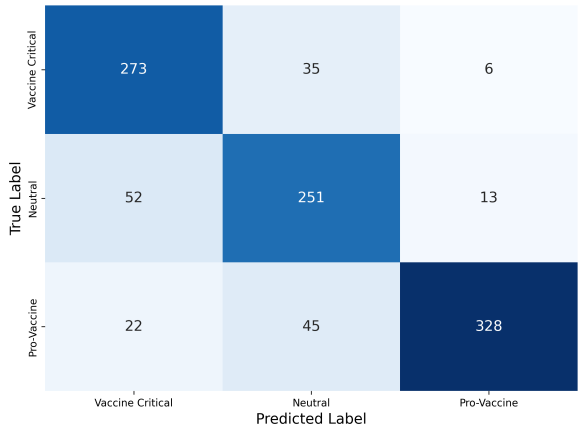



Figure 4: Confusion matrix of our final ERNIE 2.0 early-fusion pipeline on the development split. The diagonal demonstrates balanced learning, while off-diagonal clusters highlight specific vulnerabilities to satirical overlap and visual punchlines.

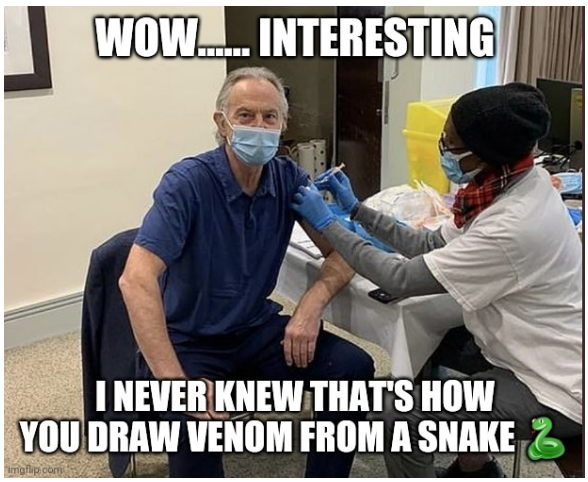
First, the model frequently fails due to the inability to parse complex pragmatic intent in "debunking" or satirical posts. As seen in Figure 5a, the model incorrectly predicts 'Vaccine Critical' due to the presence of highly critical hashtags, such as #vaccineinjury and #tonyblairisawarcriminal. Even though the user included the #vaxxed hashtag, indicating their actual Pro-Vaccine stance, the pooling mechanism fails to recognize that the user is presenting the other information as an object of ridicule. The model essentially lacks the human world knowledge required to distinguish between endorsement and ironic sharing.

Second, the text-only pipeline inherently fails on memes where the stance relies exclusively on visual sarcasm. When a user posts a benign, neutral, or even positive caption paired with a sarcastic "reaction face" (Figure 5b), the true stance is completely subverted by the visual modality. Because our ERNIE encoder is blind to facial expressions and visual tropes, it evaluates the positive text at face value and misclassifies the post. This confirms that while text-only pipelines are highly efficient and robust against general noise, achieving true human-level comprehension on internet sarcasm will ultimately require lightweight visual-feature integration.

## 6 Conclusion

In this paper, we presented a highly regularized, early-fusion text pipeline for the EEUCA 2026 COVID-19 multimodal meme classification task. Rather than relying on computationally expensive Large Vision-Language Models, we demonstrated that extracting image text via OCR and utilizing a specialized ERNIE 2.0 encoder provides a competitive and efficient alternative.

Crucially, our experiments highlight the dangers of applying hard-example mining techniques, such as Focal Loss, to inherently noisy internet datasets. We found that utilizing standard Cross-Entropy loss combined with a Multi-Sample Dropout architecture yields superior generalization by smoothing the loss landscape and ignoring mislabeled outliers. Without requiring task-specific generative vision fine-tuning, our robust text-only pipeline achieved a peak Macro F1-score of 0.834. Future work will explore applying this lightweight, noise-resistant classification head to open-source multimodal encoders to further capture nuanced visual irony without sacrificing training stability.




(a) Ablating Pragmatic Intent in Debunking Post

(b) Visual Sarcasm (Reaction Face)

Figure 5: Examples of systematic pipeline failures. **(a)** The model predicts 'Vaccine Critical' because it heavily over-indexes on critical hashtags like #vaccineinjury, completely missing the ironic, 'Pro-Vaccine' intent implied by the #vaxxed declaration. **(b)** The model fails to recognize the 'Vaccine Critical' stance because the sarcasm relies entirely on the visual reaction face, which our text-only pipeline cannot process.

## 7 Limitations

While our early-fusion text pipeline is highly efficient, it inherently fails on memes where the intended stance relies exclusively on visual sarcasm. When a user posts a benign, neutral, or even positive caption paired with a sarcastic "reaction face," the true stance is completely subverted by the visual modality. Because our ERNIE encoder is blind to facial expressions and visual tropes, it evaluates the text at face value and misclassifies the post. Furthermore, it struggles to parse complex pragmatic intents, such as users sharing critical misinformation specifically to debunk it. This confirms that while text-only pipelines are highly robust against general noise, achieving true human-level comprehension on multimodal internet sarcasm will ultimately require lightweight visual-feature integration.

## 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.

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)*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *Neural Networks*, 111:66–73.

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*, pages 2980–2988.

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.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8963–8970.

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 Hyperparameters and Reproducibility

To ensure full reproducibility of our ERNIE 2.0-Large fine-tuning experiments, we detail the complete set of hyperparameters in Table 2. All models were trained using PyTorch with Mixed Precision (FP16) on a single NVIDIA T4 GPU.

| Hyperparameter | Value |
|---|---|
| Encoder Architecture | ERNIE 2.0 Large (English) |
| Max Sequence Length | 128 tokens |
| Batch Size | 32 |
| Gradient Accumulation | 8 steps |
| Effective Batch Size | 256 |
| Epochs | 10 |
| Optimizer | AdamW |
| Learning Rate | $2 \times 10^{-5}$ |
| Weight Decay | 0.05 |
| Learning Rate Scheduler | Linear with 10% Warmup |
| Multi-Sample Dropout | 5 layers, $p = 0.3$ |

Table 2: Detailed training hyperparameters for the optimal Cross-Entropy + MSD configuration.

## B Dataset Distribution and Class Weights

The EEUCA 2026 dataset exhibits a natural class imbalance reflecting real-world social media distributions. To prevent the model from collapsing into predicting only the majority class, we applied inverse frequency weighting to the loss functions. The class weights were calculated as $w_c = N_{total}/(C \times N_c)$.

## C Detailed Per-Class Performance

While the primary evaluation metric is the unweighted Macro F1-score, analyzing the per-class F1-scores provides deeper insight into the model's behavior. As discussed in the main text, the baseline models consistently struggled with the "Neutral" class due to its highly ambiguous and sarcastic nature. The combination of early text fusion, Multi-Sample Dropout, and weighted Cross-Entropy yielded the most balanced performance across all three classes, preventing catastrophic failure on the minority "Vaccine Critical" and "Vaccine Neutral" instances while maintaining high accuracy on the clear-cut "Pro-Vaccine" memes.

To quantitatively evaluate these dynamics, we report the per-class Precision, Recall, and F1-scores on the development split in Table 3.

| Class Stance | Precision | Recall | F1-Score |
|---|---|---|---|
| Vaccine Critical | 0.7867 | 0.8694 | 0.8260 |
| Neutral | 0.7583 | 0.7943 | 0.7759 |
| Pro-Vaccine | 0.9452 | 0.8304 | 0.8841 |
| **Macro Average** | 0.8301 | 0.8314 | **0.8238** |

Table 3: Detailed performance breakdown per stance class evaluated on the development split for the final ERNIE 2.0 + MSD configuration.