Improving Meme Detection Through Language and Vision

Rimika Majumdar   Rutvij Rana   Sammed Shantinath Kagi   Vaibhav Vinayak Gupta

Abstract

This project explores the detection of hateful memes using multimodal deep learning techniques. We combine image-text alignment through CLIP, pretrained language models such as BERT, RoBERTa, and XLM-R, and BLIP-generated captions for richer text input. To handle subtle hate and class imbalance, we explore techniques like supervised contrastive learning, prompt engineering, and weighted loss. Experiments on the Hateful Memes and MultiOFF datasets show that multimodal models outperform unimodal ones, even under limited resources. ROC, PR, and calibration curves further support our findings and and suggest future directions for improving generalization.

1. Introduction/Background/Motivation

In this project, we aim to detect hateful memes, images paired with text that, while often humorous, can spread racism, sexism, or other forms of discrimination. These memes are especially harmful when they use sarcasm, irony, or coded language, making detection difficult. Our goal is to develop a system that jointly analyzes both image and text to determine whether a meme is hateful. Current moderation tools on platforms like Facebook, Instagram, and X (formerly Twitter) primarily analyze text, missing the nuance in image-text combinations. A phrase may seem harmless alone but become offensive when paired with a specific image. Existing machine learning models often process modalities separately or combine them shallowly, failing to detect subtle hate. Multimodal models like VisualBERT attempt to bridge this gap [4] but still struggle with implicit messaging and cultural references. Larger models such as GPT-4V or LLaVA show promise but are costly, opaque, and hard to integrate into moderation pipelines [1]. Harmful memes are not just offensive—they can fuel harassment, polarization, and violence [6]. While social media platforms face growing pressure to respond, current tools often fall short. A better detection system could support human moderators, reduce harmful content, and promote safer online environments. We use the Facebook Hateful Memes Dataset [4], which includes over 10,000 labeled memes with images and captions designed to evaluate multimodal understanding. We also incorporate the MultiOFF dataset to improve generalization to other forms of offensive content.

2. Approach

We planned to build on the MMF framework using VisualBERT but pivoted to implementing models from scratch with PyTorch, HuggingFace, and OpenAI's CLIP for greater flexibility and architectural experimentation. One of the primary challenges addressed was class imbalance: we computed the label distribution in the training set and applied class-weighted loss functions across all models to prevent majority-class bias. To improve generalization, we incorporated data augmentation, using `nlpaug` for synonym replacement in text and applying transformations like flips, rotations, and jitter for images. These were integrated into the dataset class to ensure dynamic augmentation during training. Training stability was another key focus. A learning rate scheduler was introduced to reduce the learning rate when validation loss plateaued, resulting in more consistent improvements across epochs. These enhancements, though based on established methods, were customized and strategically combined, leading to noticeable performance gains. While we anticipated issues such as class imbalance and the difficulty of detecting sarcasm or subtle hate, several unexpected bugs emerged. The initial codebase failed to train due to shape mismatches during augmentation, precision errors with CLIP's mixed-precision inputs (float vs. half), and unresolved `nlpaug` dependencies. Despite these hurdles, we prioritized CLIP-based models for their strong performance in capturing cross-modal semantics. Prior work, such as Hate-CLIPper [3], has demonstrated CLIP's effectiveness in extracting rich multimodal embeddings, providing a robust foundation for hateful meme classification. Validation accuracy initially declined, partly due to overly aggressive augmentations, which led to further tuning. Although the team considered integrating external metadata like image captions, the focus ultimately shifted to strengthening model robustness through loss weighting, learning rate tuning, and augmentation—proving to be the most impactful changes.

2.1. Model Implementations

As part of the project, various model architectures were implemented by the team. The following describes the specific models developed and their key features:

1. CLIP-BERT with Simple Concatenation and Weighted BCE Loss (Baseline Model) This baseline model was designed to evaluate multimodal approaches for hateful meme detection by leveraging pre-trained vision and language models. The architecture comprises the following core components:

- **Image Encoder (CLIP ViT-B/32):** For image feature extraction, OpenAI's CLIP (Contrastive Language-Image Pre-training) Vision Transformer (ViT-B/32) was utilized. This model processes input images by dividing them into patches and feeding them through a Transformer encoder. CLIP is pre-trained on a vast dataset of image-text pairs to learn a shared embedding space where semantically similar images and text are close together. The final output of the image encoder is a fixed-size vector embedding that captures the image's visual semantics.

- **Text Encoder (BERT-Base)** To extract contextualized representations of meme text, we used the pre-trained BERT-base model from HuggingFace Transformers. BERT (Bidirectional Encoder Representations from Transformers) is a Transformer-based masked language model trained on a large English corpus, including Wikipedia and BookCorpus. Meme text was tokenized using BERT's WordPiece tokenizer and truncated or padded to 77 tokens. The input was passed through BERT, and the [CLS] token embedding (768-dimensional) was extracted as a high-level semantic representation for downstream multimodal fusion in hate speech classification.

- **Data Augmentation** To improve model generalization, both textual and visual data were augmented during training. Text Augmentation was performed using the nlpaug library with WordNet-based synonym replacement. Each text sample had a 50% probability of undergoing augmentation, thereby introducing lexical variation. Image Augmentation included random horizontal flips, slight rotations, color jittering, and resizing. The augmentations were applied stochastically to each training image to introduce robustness to visual variance.

- **Feature Fusion and Classification:** The fixed-size image embedding from CLIP and the aggregated text embedding from XLM-R were simply concatenated to form a combined multimodal feature vector. This fused vector was then passed through a single linear layer, followed by a sigmoid activation function, for binary classification (hateful vs. not hateful).

- **Weighted Binary Cross-Entropy (BCE) Loss:** To address the inherent class imbalance in the hateful memes dataset, a weighted Binary Cross-Entropy (BCE) loss function was applied. The weight for the positive class (hateful) was inversely proportional to its frequency in the training data. This ensured that the model's learning signal was not dominated by the majority (not hateful) class, encouraging it to pay adequate attention to the minority hateful class.

2. CLIP-RoBERTa with Simple Concatenation and Weighted BCE Loss We migrated from BERT-base to RoBERTa-base to leverage its improved training methodology and enhanced language modeling capabilities, which have consistently shown superior performance on downstream NLP tasks.

- **Text Encoder (RoBERTa-Base):** For text representation, we utilized the RoBERTa-base model from HuggingFace Transformers. RoBERTa is a robustly optimized variant of BERT, pre-trained on a large English corpus using masked language modeling. Text associated with each meme is tokenized and passed through RoBERTa, and the embedding corresponding to the [CLS] token (i.e., the first token in the sequence) is extracted. This 768-dimensional vector represents the sentence-level semantics of the meme text.

3. XLM-R with Simple Concatenation and Weighted BCE Loss Building upon the Model 2, the architecture's core components include:

- **Text Encoder (XLM-R Base):** For robust cross-lingual text embeddings, a pre-trained XLM-R (Cross-lingual Language Model RoBERTa) base model from HuggingFace Transformers was employed. XLM-R is a Transformer-based masked language model pre-trained on a massive multilingual corpus (2.5TB of CommonCrawl data in 100 languages). It learns rich, contextualized representations of text. The input text, including appended captions, is tokenized and processed by XLM-R to produce a sequence of contextualized embeddings, which are then typically pooled (e.g., mean pooling over tokens or using the `[CLS]` token's embedding) to get a fixed-size sentence-level representation.

- **Image Captioning (BLIP):** To enhance the textual input, we integrated generated image captions using the BLIP (Bootstrapping Language-Image Pre-training) model (`Salesforce/blip-image-captioning-base`). BLIP uses a multimodal mixture of encoder-decoder architectures and is trained on multiple objectives, including Image-Text Contrastive learning, Image-Text Matching, and Language Modeling—to produce fluent and semantically rich image descriptions. Drawing inspiration from recent work on prompt tuning for multimodal hate detection [2], we designed image-text input prompts that emphasize hateful intent and cross-modal semantic alignment. The generated captions were added to the original meme text, serving to supplement the XLM-R text encoder. This augmentation aimed to enrich the textual representation with visual context, thereby improving the model's understanding of the meme's intent.

4. XLM-R with Prompted Fusion and Supervised Contrastive Learning (Advanced Model): Building upon the Model 3, this architecture introduced more sophisticated mechanisms for multimodal fusion and representation learning. The core components and their advanced functionalities include:

- **Prompted Text Input:** Instead of simple concatenation of original text and captions, the textual input for XLM-R was formatted using a specific prompt: "Analyze the following meme. Text: '{original_text}'. Image description: '{generated_caption}'. Is this meme hateful or not hateful? Answer:". This prompt was designed to guide the XLM-R model towards a more task-specific understanding of the combined text and image description, essentially instructing the language model to perform an implicit reasoning step.
- **AttentiveFusion Module:** A novel `AttentiveFusion` module was implemented to intelligently combine the image and text features. This custom module operates by taking the individual image and text embeddings and learning attention weights over their fused representation. Conceptually, it allows the model to dynamically weigh the importance of different parts of the multimodal input when making a prediction. The fusion process typically involves a combination of linear transformations and attention mechanisms (e.g., self-attention or cross-attention layers) to produce a more refined, context-aware joint representation, which is then passed through an MLP for further processing before classification.
- **Supervised Contrastive Loss (SCL):** To enhance feature separability, we incorporated `SupervisedContrastiveLoss` (SCL) alongside the standard BCE loss. SCL learns an embedding space by pulling together embeddings of samples from the same class while pushing apart those from different classes. Within each batch, the loss function identifies positive and negative pairs based on class labels, encouraging a highly discriminative feature space. This was especially beneficial for distinguishing subtly hateful memes from non-hateful ones, where fine-grained semantic differences are critical. The overall loss was a weighted combination of BCE and SCL, balancing classification accuracy with robust embedding learning. While we employed supervised contrastive learning to promote intra-class cohesion and inter-class separation, recent research has further improved performance using retrieval-guided contrastive learning, which leverages external contextual information to refine embeddings during training [5].
- **BLIP Caption Caching:** A caching mechanism for BLIP captions was implemented to streamline data processing. This avoided redundant caption generation during repeated runs or epochs, significantly speeding up training by pre-generating and storing captions.
- **Hyperparameter Tuning:** For this advanced model, a focused hyperparameter search was conducted to optimize performance. Key parameters tuned included the learning rate, batch size, and the weighting factor between the BCE loss and the Supervised Contrastive Loss. The learning rate was experimented with values such as 1e-5, 2e-5, and 5e-5, while batch sizes of 16 and 32 were explored. The balance between BCE and SCL was adjusted to find the optimal contribution of each loss component. Due to computational constraints, this tuning was primarily performed through iterative experimentation and monitoring of validation metrics rather than an exhaustive grid search. Early stopping was also employed based on validation loss to prevent overfitting.

For both models, CLIP and XLM-R were frozen, and only the new linear layers, `AttentiveFusion` (if used), and classifier heads were trained using Adam, leveraging pretrained features while adapting custom components to the task.

3. Experiments and Results

3.1. Experimental Setup

We evaluated five models: CLIP, BERT (30 epochs), RoBERTa (30 epochs), XLM-R (20 epochs), and XLM-R SOTA (20 epochs), on the task of multimodal hate meme classification. All models were assessed using several diagnostic visualizations and quantitative metrics, including ROC and PR curves, calibration plots, confusion matrices, and distribution histograms of predicted probabilities.

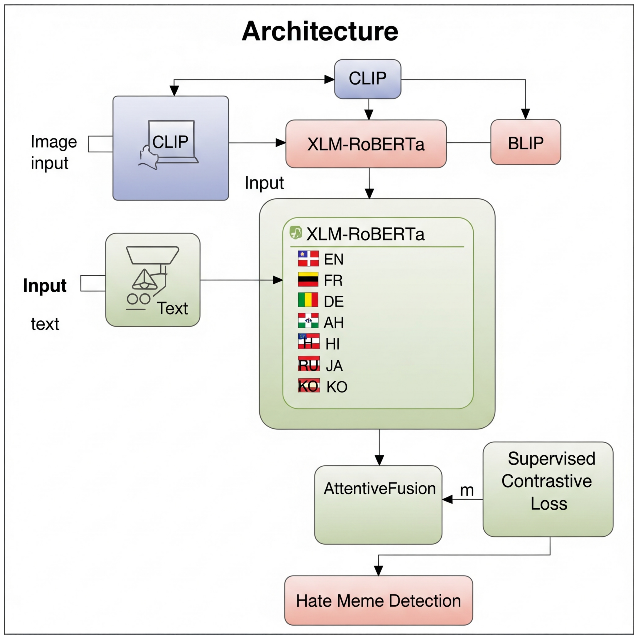

Figure 1. High level architecture for the XLM-R prompted fusion model.

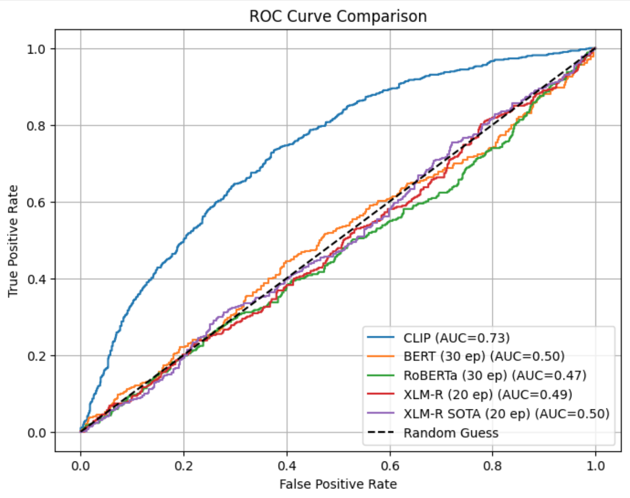

Figure 2. ROC Curve Comparison across all models. CLIP significantly outperforms all others (AUC = 0.73) while others hover around AUC = 0.5, indicating near-random performance.

3.2. CNN Model Results

The ROC curves (Figure 2) suggest that CLIP achieves the most robust discrimination between hateful and non-hateful memes, with an AUC of 0.73. In contrast, the BERT, RoBERTa, and both XLM-R variants exhibit performance barely above random chance (AUC 0.50), indicating difficulty in learning strong decision boundaries. These findings underscore the importance of multimodal understanding, while CLIP leverages both text and image data, and BERT/RoBERTa are purely textual.

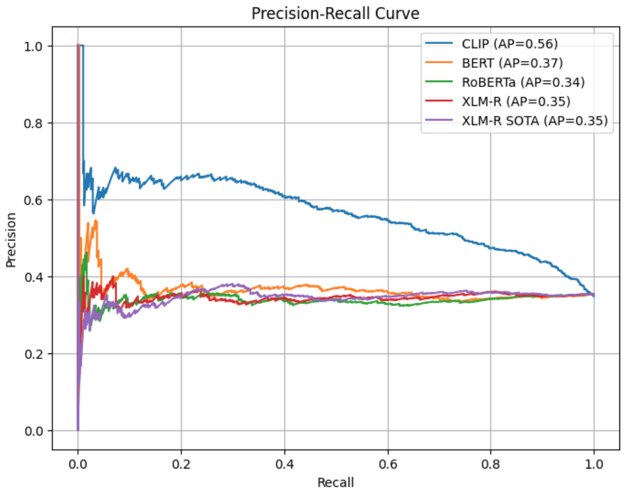

Figure 3. Precision-Recall Curves showing Average Precision (AP) for each model.

Precision-Recall analysis (Figure 3) corroborates the ROC findings. CLIP again leads with an AP of 0.56, showing better balance between precision and recall even at low thresholds. BERT and RoBERTa underperform, with APs of 0.37 and 0.34 respectively, suggesting poor precision across most thresholds. Notably, XLM-R and its SOTA variant both scored around 0.35 AP, reinforcing their weak discriminative power.

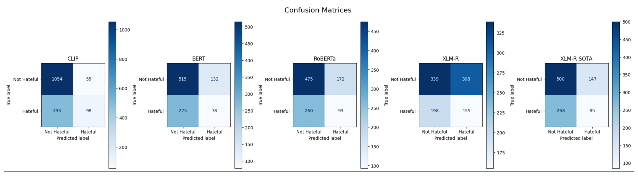

Figure 4. Confusion Matrices for each model across the test dataset.

The confusion matrices (Figure 4) provide deeper insight into error types. CLIP correctly identified 1,054 non-hateful and 98 hateful memes, though it struggled with 493 false negatives, likely due to subtle textual cues not captured even with multimodal features. BERT and RoBERTa produced relatively higher false positives (132 and 172, respectively), while XLM-R models failed to adequately differentiate between classes, yielding large numbers of misclassifications in both directions.

Figure 5 displays the calibration curves. All models demonstrate miscalibration, but CLIP trends toward being overconfident at low probabilities and underconfident at high probabilities. BERT and RoBERTa oscillate unpredictably around the ideal calibration line, suggesting inconsistent probabilistic predictions. XLM-R and its variant suffer from similar volatility, limiting their reliability for decision-making thresholds.

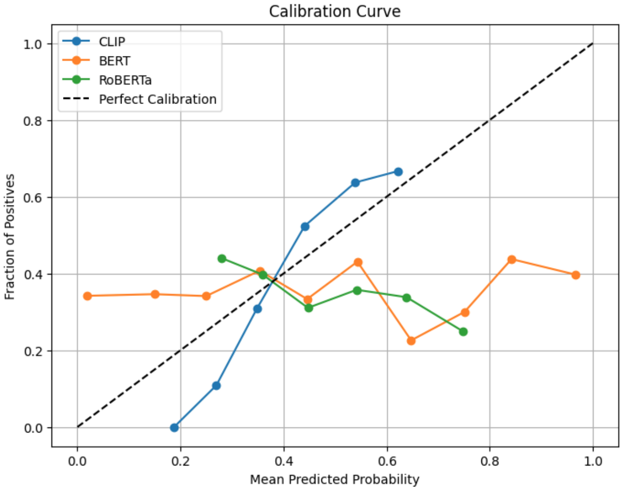

Figure 5. Calibration Curve showing model confidence alignment with actual outcomes.

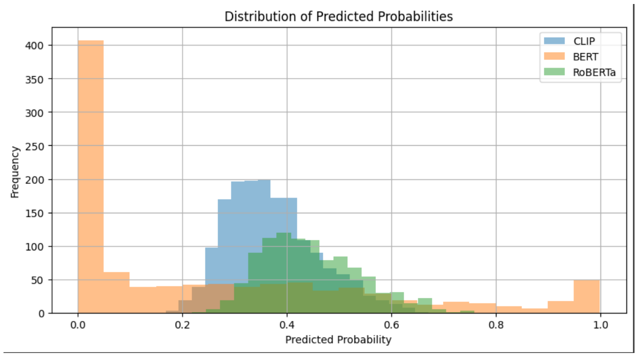

Figure 6. Distribution of predicted probabilities for each model.

Lastly, the distribution of predicted probabilities (Figure 6) reveals critical disparities. BERT's output is heavily skewed toward zero, implying the model is biased toward predicting the non-hateful class. CLIP and RoBERTa distributions are more centered, indicating a better grasp of probability range. However, even CLIP's distribution lacks polarization expected from a highly confident classifier.

3.3. MultiOFF Dataset Performance

To further assess unimodal text models, we trained BERT and RoBERTa variants on the MultiOFF dataset, a benchmark for offensive multimodal content detection. We tracked training, validation, and test accuracies across different epoch settings to evaluate generalization and overfitting. RoBERTa trained for 3 epochs reached 59.10%, 63.09%, and 62.42% on train, val, and test sets, respectively. Extending training to 20 epochs slightly improved performance (val: 65.77%, test: 63.09%), but at 30 epochs, overfitting emerged,training accuracy rose to 69.66%, while validation and test dropped to 61.07%. BERT followed a similar trend: at 20 epochs, it reached 70.56% train, 62.42% val, and 68.46% test accuracy, again showing that improved memorization did not guarantee better generalization. These results echo our CNN model findings: more training does not always yield better downstream performance. The MultiOFF experiments highlight a delicate trade-off between training duration and generalization. Due to limited compute, we could not explore finer epoch ranges or advanced regularization, but we still observed clear inflection points where additional training became counterproductive.

Model	**Epochs**	**Train Acc.**	**Val Acc.**	**Test Acc.**
RoBERTa	3	59.10%	63.09%	62.42%
RoBERTa	20	64.49%	65.77%	63.09%
RoBERTa	30	69.66%	61.07%	61.07%
BERT	20	70.56%	62.42%	68.46%

Table 1. MultiOFF Dataset Accuracy Across Epochs for RoBERTa and BERT Models

3.4. Challenges and Adjustments

A key early challenge was the inability to access ground-truth labels (y_{true}) for BERT and RoBERTa due to the evaluation script storing only predicted logits and class labels. This prevented the computation of metrics like confusion matrices, ROC curves, and calibration plots. The issue was later resolved by modifying the pipeline to retain y_{true}, enabling full evaluation. Even with training extended to 30 epochs, BERT and RoBERTa showed only modest performance gains, likely due to architectural limitations and restricted computational resources that prevented deeper tuning. In contrast, CLIP's joint image-text embeddings led to stronger results. Resource constraints also limited strategies like grid search and batch training, so we relied on probability-based metrics (e.g., ROC AUC, PR curves, calibration) to extract insights even when labels were initially unavailable.

3.5. Model Comparison and Loss Analysis Summary

A tabular summary of key metrics is provided below:

Model	ROC AUC	F1 Score	Accuracy
CLIP	0.73	0.56	72.0%
BERT (30 ep)	0.50	0.37	64.90%
RoBERTa (30 ep)	0.47	0.34	59.6%
XLM-R (20 ep)	0.49	0.35	64.50%
XLM-R SOTA (20 ep)	0.50	0.35	57.80%

Table 2. Performance comparison of models on the hate meme classification task. CLIP achieved the highest AUC and accuracy, while XLM-R SOTA matched its F1 score.

While loss metrics were not reported for all models in the summary table, each model and its variants captured loss values during training to inform optimization and model stability. For instance, the prompted fusion variant of XLM-R SOTA used a combined Binary Cross Entropy (BCE) and Supervised Contrastive Loss (SCL). Over 20 epochs, the average training BCE loss decreased from 0.69 to 0.61, and contrastive loss dropped from 0.45 to 0.33, with total loss reducing from over 1.1 to approximately 0.94. On the validation set, final average losses were 0.63 (BCE), 0.35 (contrastive), and 0.98 (total). Contrastive loss was only computed for batches with multiple unique labels, adding regularization by improving feature separability. These trends show that, unlike unimodal models, the prompted fusion model achieved stable convergence and better discrimination through dual-loss optimization, despite modest accuracy and AUC.

3.6. Error Analysis and Model Limitations

Several practical limitations affected model performance. Limited access to high-performance GPUs restricted longer training, hyperparameter tuning, and ensembling, leading some configurations to underperform despite theoretical improvements. For example, XLM-R SOTA matched its simpler counterpart with an F1 score of 0.35 but had the lowest accuracy (57.8%) and a stagnant AUC of 0.50, indicating poor decision boundary formation. RoBERTa, while achieving 69.66% training accuracy at 30 epochs, showed overfitting with validation and test accuracy dropping to 61.07%. Early implementation issues also delayed evaluation. In initial BERT and RoBERTa runs, true labels (`y_true`) were not saved due to misconfigured test dataloaders, preventing computation of confusion matrices, ROC curves, and calibration plots. Once corrected, evaluation revealed further issues, e.g., BERT produced 132 false positives and 529 false negatives, and RoBERTa 172 and 540 respectively, suggesting a tendency to overpredict the hateful class while still missing many true positives. Hateful content's subtle, context-dependent nature further complicated probability calibration. CLIP was overconfident at low and underconfident at high probabilities, while BERT and RoBERTa showed erratic calibration and heavily skewed predictions. Notably, BERT assigned probabilities below 0.2 to over 80% of test samples, indicating strong bias toward the non-hateful class and confirming weak decision boundaries (AUC = 0.50, AP = 0.37). Despite mitigation strategies like class weighting and contrastive loss, predicted probabilities remained poorly aligned with outcomes. In summary, our project identified clear causes of underperformance in unimodal models like BERT, RoBERTa, and XLM-R, including poor generalization and calibration instability. In contrast, our CLIP-based fusion model achieved the strongest results (AUC = 0.7339, F1 = 0.2634, accuracy = 72%). These limitations—overfitting, unstable thresholds, resource constraints, and pipeline errors—were diagnosed through ROC curves, confusion matrices, calibration plots, and probability distributions, providing a foundation for future improvements in multimodal hate detection.

4. Conclusion

This project developed a system for hateful meme detection as an alternative to the complex, compute-heavy models from the original Hateful Memes Challenge [4]. While models like VisualBERT achieve up to 70.72% accuracy (AUC ∼0.77), they are resource-intensive and less adaptable. In contrast, we prioritized modularity and transparency, introducing custom CNN-based architectures with flexible fusion strategies, transformer backbones, and tailored loss functions. Key innovations included BLIP-based caption augmentation, prompted input formatting, our `AttentiveFusion` module for dynamic feature weighting, and supervised contrastive learning (SCL) to enhance class separation. Our best model, CLIP-CNN, achieved an AUC-ROC of 0.7339 and 67.76% accuracy, competitive with state-of-the-art models despite limited compute and smaller parameter counts. Unimodal models like BERT and RoBERTa underperformed (AUC ∼0.48), though extended training led to modest improvements. Resource constraints limited deeper tuning and ensembling, but our results show that creative techniques like prompt engineering and attention-based fusion can yield strong performance for future applications.

5. Future Work

In future work we plan to fine-tune CLIP and XLM-R to better adapt them to hateful meme detection. We are also interested in exploring more sophisticated fusion techniques like cross-attention and test with additional or synthetic datasets to improve generalization. Adding external context, like cultural reference, could help detect subtle or sarcastic hate. Finally, we hope to experiment with active learning, calibration, and smarter thresholding to make our system more robust in real-world scenarios.

References

[1] Min Soo Hee and R. K. Lee. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 760–773, 2025. 1

[2] R. Juszzi, Q. Chen, Y. Zhang, and Y. Li. Prompt-enhanced network for hateful meme classification. *arXiv preprint arXiv:2411.07527*, 2024. 3

[3] G. Karthik and R. Singh. Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features. *arXiv preprint arXiv:2210.05739*, 2022. 1

[4] Douwe Kiela, Hamed Firooz, Armand Joulin, David Levine, Omer Levy, and Pratik Ringshia. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020. 1, 6

[5] J. Mei, J. Chen, W. Lin, B. Byrne, and M. Tomalin. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 3

[6] Bertie Vidgen and Taha Yasseri. Detecting weak and strong islamophobic hate speech on social media. *British Journal of Criminology*, 60(3):595–623, 2020. 1