COSE474-2024F: Final Project
Sarcasm-Aware CLIP and RoBERTa based Multimodal Hateful Meme Classification

Mirza Syahmi Bin Afandi

1. Introduction

The classification of hateful memes using deep learning techniques has emerged as a critical area of research due to the increasing prevalence of such content on social media platforms. Hateful memes, which combine visual and textual elements to convey harmful messages, present unique challenges for detection systems. This complexity arises from the need to analyze both modalities simultaneously, as the meaning often derives from their interaction rather than from either modality in isolation (Pan et al., 2022; Zhou et al., 2021). The Hateful Memes Challenge, initiated by Facebook, has significantly contributed to this field by providing a large dataset of over 10,000 memes, which includes both hateful and benign examples, thereby facilitating the development and evaluation of multimodal classification models (Kiela et al., 2020).

The primary problem addressed in this research is the difficulty in accurately classifying hateful memes, particularly those that employ sarcasm as a rhetorical device. Sarcasm can conceal the intended meaning of a meme, making it challenging for existing models to distinguish whether the content is genuinely hateful or benign. This issue is compounded by the multimodal nature of memes, which requires an understanding of both visual and textual contexts. Previous studies have demonstrated that unimodal approaches often fall short in detecting hate speech within memes, necessitating the development of more robust multimodal frameworks that can leverage the strengths of both image and text analysis (Kirk et al., 2021).

Therefore, in response to these challenges, we present a novel multimodal approach for hateful meme classification that integrates visual and textual features with sarcasm detection. Our key **contributions** are:

- We proposed SARCMeme (**S**arcasm-**A**ware **C**LIP and **R**oBERTa based Multimodal Hateful **Meme** Classification), a novel architecture combining Contrastive Language–Image Pretraining (CLIP) and Robustly Optimized BERT Pre-training Approach (RoBERTa) models, incorporating a sarcasm detection module to better interpret nuanced and sarcastic text often used to mask hateful intent.
- Our model outperforms baseline methods on a benchmark dataset for hateful meme classification, demonstrating the effectiveness of incorporating sarcasm detection and multimodal fusion.

2. Method

The approach combines multimodal representation learning with a specialized linguistic feature—sarcasm detection—to improve the classification of hateful memes. Traditional hateful meme detection methods often focus on standard visual-linguistic fusion without incorporating deeper pragmatic cues like sarcasm. By integrating a pre-trained sarcasm detection module (RoBERTaSarcasmDetector) with a CLIP-based vision-language encoder, the method enhances understanding of the text's intent and tone. This complementary interaction allows the model to better distinguish between genuinely hateful content and content that may appear hateful on the surface but is intended ironically or sarcastically. Such an approach is particularly novel in that it leverages sarcasm awareness to inform a complex multimodal decision boundary, improving robustness and interpretability in hateful content detection tasks.

2.1. SARCMeme

We proposed a framework that integrates vision language embeddings from a pre-trained CLIP model with a linguistic sarcasm signal extracted via a RoBERTa-based sarcasm detector, resulting in a robust multimodal representation for hateful meme classification. Specifically, we begin with an image-text pair (I, T) from a meme. CLIP produces text embeddings $\mathbf{t} \in \mathbb{R}^{d_t}$ and image embeddings $\mathbf{i} \in \mathbb{R}^{d_i}$, while a frozen RoBERTa-based sarcasm detector outputs a scalar sarcasm score $s \in \mathbb{R}$. To unify these representations, we learn projection parameters W_t, W_i, and W_s that map each modality into a common latent space of dimension d:

$$\mathbf{t}' = W_t\mathbf{t}, \quad \mathbf{i}' = W_i\mathbf{i}, \quad s' = W_s[s]. \tag{1}$$

After projection, we concatenate the transformed features into a single vector $\mathbf{z} = [\mathbf{t}' \| \mathbf{i}' \| s'] \in \mathbb{R}^{3d}$.

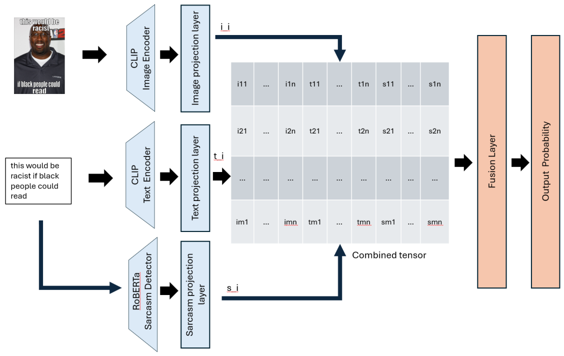

Figure 1. SARCMeme architecture. The input meme's text and image are first processed by a CLIP model to produce text and image embeddings. In parallel, the text is analyzed by a frozen RoBERTa-based sarcasm detector, yielding a scalar sarcasm score. Each representation—text, image, and sarcasm—is projected into a common embedding dimension and then concatenated into a single combined vector. This multimodal vector is passed through a fusion network that integrates the features and outputs a final probability indicating whether the meme is hateful.

This combined vector $\mathbf{z}$ is then processed by a fusion network consisting of fully connected layers, nonlinear activations (e.g., ReLU), and dropout layers to mitigate overfitting. The fusion network outputs a logit $\ell \in \mathbb{R}$, which is mapped to a probability $p \in [0, 1]$ via a sigmoid function. The training objective for the binary classification task (hateful vs. not hateful) is the binary cross-entropy (BCE) loss:

$$\mathcal{L}(\Theta; D^{tr}) = -\sum_{(\mathbf{x},y)\in D^{tr}} \big[y \log p(\mathbf{x}) + (1-y)\log(1-p(\mathbf{x}))\big], \tag{2}$$

where Θ denotes all learnable parameters and D^{tr} is the training set.

To reproduce our results, one must consistently follow the described data preprocessing steps, use the same pretrained model checkpoints for both CLIP and RoBERTa, and apply identical hyperparameters and random seeds. By maintaining the same dimensionalities (d_t, d_i, and d), the same projection matrices W_t, W_i, W_s, and the same fusion network architecture, as well as the training configurations (e.g., learning rate, batch size, number of epochs), other researchers can replicate our exact experimental conditions. The described equations and architecture details provide a clear and verifiable blueprint for implementing and confirming the performance of this hateful meme classification framework.

3. Experiments

To evaluate the efficacy of our proposed framework for hateful meme classification, we conducted a series of experiments leveraging two publicly available datasets and state-of-the-art computational resources. The experiments were designed to assess the individual components of the model, including the sarcasm detection module and the multimodal fusion network, as well as their collective performance in accurately classifying hateful memes. This section outlines the datasets used, the computational setup, and the experimental pipeline. Additionally, we detail how the framework was trained and evaluated using real-world data and modern machine learning tools.

3.1. Datasets

To evaluate the performance of our proposed hateful meme classification framework, we utilized two distinct datasets:

- **Hateful Memes Dataset**: Developed by Kiela et al., this dataset comprises memes labeled as hateful or non-hateful, specifically designed for the task of hateful meme detection (Kiela et al., 2020). It was employed for both training and testing our classification model.
- **Memotion Analysis Dataset**: Sourced from SemEval, this dataset contains memes annotated for various attributes, including sarcasm (Sharma et al., 2020). We leveraged this dataset to train our sarcasm detection module, which is crucial for understanding the nuanced meanings behind meme texts.

3.2. Computational Resources

Our experiments were conducted on Google Colab Pro, which provides access to high-performance computing resources. Specifically, we utilized:

GPU: NVIDIA A100 Tensor Core GPU, designed for high-performance computing tasks, enabling efficient training of deep learning models.
CPU: Intel Xeon processors, offering robust performance for data preprocessing and model training.
OS: Linux, which is commonly used in machine learning environments for its compatibility with various frameworks.
Software: The models were implemented using PyTorch, a flexible deep learning framework that supports dynamic computation graphs and extensive libraries for model development.

3.3. Experimental Design

Our experimental pipeline is structured as follows:

Data Preprocessing: Textual components of memes were tokenized using the RoBERTa tokenizer, while images were

processed with the CLIP processor to ensure compatibility with their respective models.
Sarcasm Detection Module Training: The RoBERTa-based sarcasm detector was trained on the Memotion Analysis dataset to predict sarcasm scores from meme texts, which enhances the model's ability to discern underlying meanings.
Feature Extraction with CLIP: Both textual and visual features were extracted from memes using a pretrained CLIP model, yielding embeddings for each modality.
Feature Projection: The extracted embeddings and sarcasm scores were projected into a unified latent space through learned linear transformations, facilitating effective multimodal integration.
Feature Fusion: The projected features were concatenated and passed through a fusion network comprising fully connected layers with ReLU activations and dropout for regularization, which is essential for preventing overfitting.
Classification: The output of the fusion network was fed into a sigmoid-activated layer to produce a probability score indicating the likelihood of the meme being hateful.
Training and Evaluation: The entire model was trained end-to-end on the Hateful Memes dataset, with performance evaluated using standard metrics such as accuracy, precision, recall, and F1-score.

This comprehensive experimental design ensures a robust evaluation of our framework, addressing the complexities inherent in hateful meme classification through a multimodal approach.

4. Results

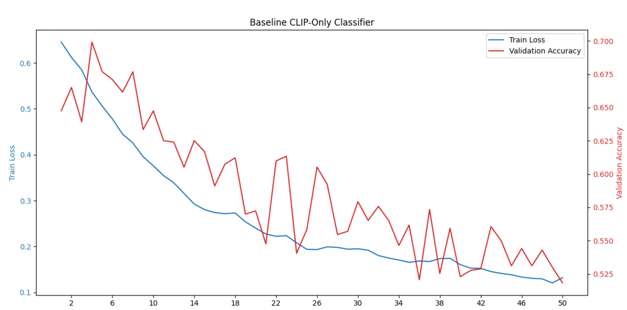

Figure 2. Training Loss and Validation Accuracy for Baseline CLIP-Only Classifier Across Epochs.

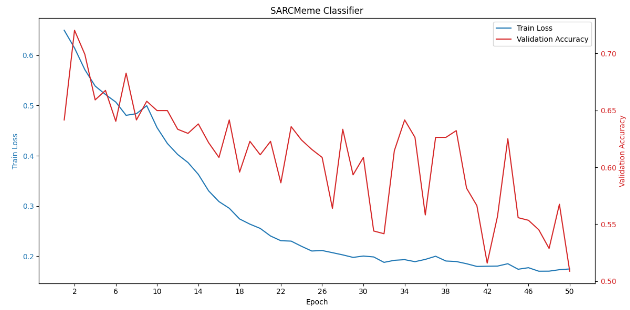

Figure 3. Training Loss and Validation Accuracy for SARCMeme Model Across Epochs.

Metric	Baseline CLIP-Only	SARCMeme
Accuracy	55.53%	57.29%
Precision	0.4037	0.3978
Recall	0.5016	0.5869
F1-Score	0.4474	0.4742

Table 1. Test Evaluation Metrics for Baseline CLIP-Only Classifier and SARCMeme Model.

4.1. Quantitative Results

The Baseline CLIP-Only Classifier exhibited a steady decline in training loss over 50 epochs, as shown in Figure 2. Validation accuracy for the baseline model peaked at 69.92% in epoch 4 but generally fluctuated between 50% and 70% across subsequent epochs. In comparison, the SARCMeme model also demonstrated a consistent reduction in training loss (Figure 3), with validation accuracy reaching a high of 72.03% in epoch 2 before stabilizing within a similar range. Table 1 summarizes the test evaluation metrics, revealing that SARCMeme achieved an accuracy of 57.29% and an F1-Score of 0.4742, outperforming the Baseline CLIP-Only Classifier, which attained 55.53% accuracy and a 0.4474 F1-Score. These quantitative results indicate that the SARCMeme model provides a marginal yet meaningful improvement in classification performance over the baseline.

4.2. Qualitative Results

Both the Baseline CLIP-Only Classifier and the SARCMeme model display signs of overfitting, as evidenced by decreasing training loss alongside fluctuating validation accuracy. The baseline model struggled with inconsistent precision and recall, suggesting challenges in reliably identifying hateful memes. On the other hand, SARCMeme improved recall significantly, enhancing its ability to detect positive instances of hateful content, albeit with a slight decrease in precision. This balance is reflected in the higher

F1-Score of the SARCMeme model, indicating a better trade-off between precision and recall. The integration of sarcasm detection through the RoBERTa-based component in SARCMeme appears to enhance the model's sensitivity to nuanced textual cues, thereby improving overall classification performance. However, the variability in validation metrics highlights the necessity for further optimization in model architecture and training strategies to achieve more stable and robust results.

4.3. Discussion

The proposed SARCMeme model demonstrates a modest improvement over the Baseline CLIP-Only Classifier, achieving higher accuracy and F1-Score on the test set (57.29% vs. 55.53% accuracy and 0.4742 vs. 0.4474 F1-Score). Notably, SARCMeme exhibits a significant increase in recall (0.5869 compared to 0.5016), indicating an enhanced capability to identify hateful memes. This improvement suggests that incorporating sarcasm detection via the RoBERTa-based component effectively aids the model in capturing nuanced textual cues associated with hatefulness. However, the slight decrease in precision (0.3978 vs. 0.4037) implies a marginal rise in false positives, which may necessitate further refinement. Additionally, both models exhibit signs of overfitting, as evidenced by the divergence between training and validation performance metrics. Overall, while SARCMeme achieves incremental gains, it underscores the potential benefits of integrating multimodal features, yet also highlights the need for continued optimization to fully harness the advantages of sarcasm detection in hateful meme classification.

5. Conclusion

In this study, we introduced the SARCMeme model, an enhancement over the Baseline CLIP-Only Classifier, aimed at improving the detection of hateful memes by incorporating sarcasm detection through a RoBERTa-based component. Our experimental results demonstrate that SARCMeme achieves improvement compared to the baseline. These findings suggest that integrating sarcasm detection effectively enhances the model's ability to identify nuanced hateful content. However, the slight decline in precision indicates a trade-off that necessitates further refinement to minimize false positives. Both models exhibited signs of overfitting, highlighting the need for advanced regularization techniques and optimization strategies.

Future work will focus on addressing overfitting by implementing methods such as early stopping, increased dropout rates, and data augmentation to enhance generalization. Additionally, exploring more sophisticated multimodal fusion techniques, such as attention mechanisms, could further leverage the interplay between textual and visual features. Expanding the dataset and incorporating more diverse examples of hateful and non-hateful memes can provide the models with a richer learning environment. Lastly, fine-tuning the integration of the sarcasm detection component, potentially allowing for joint training rather than freezing, may yield better synergistic performance between the textual and visual modalities. These advancements aim to create a more robust and accurate classifier for identifying hateful content in memes.

References

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33: 2611–2624, 2020.

Kirk, H. R., Jun, Y., Rauba, P., Wachtel, G., Li, R., Bai, X., Broestl, N., Doff-Sotta, M., Shtedritski, A., and Asano, Y. M. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. *arXiv preprint arXiv:2107.04313*, 2021.

Pan, X., Chen, P., Gong, Y., Zhou, H., Wang, X., and Lin, Z. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. *arXiv preprint arXiv:2203.07996*, 2022.

Sharma, C., Bhageria, D., Scott, W., Pykl, S., Das, A., Chakraborty, T., Pulabaigari, V., and Gamback, B. Semeval-2020 task 8: Memotion analysis–the visuo-lingual metaphor! *arXiv preprint arXiv:2008.03781*, 2020.

Zhou, Y., Chen, Z., and Yang, H. Multimodal learning for hateful memes detection. In *2021 IEEE International conference on multimedia & expo workshops (ICMEW)*, pp. 1–6. IEEE, 2021.