Multi modalities fusion for hateful meme classification

Mohamed Yassine Aouame [1] **Mattéo Berthet** [1] **Oussama Jaffal** [1]

Abstract

Every day, social media platforms witness an increasing influx of content, including a significant amount of hateful material. Addressing this issue is crucial, as humans often employ subtle and implicit language and memes that challenge the capabilities of automated models to classify them as hate speech. This paper proposes the development of a multimodal classifier that leverages highly descriptive feature extraction from pre-trained models, enhancing the classifier's ability to accurately identify implicit hate speech in memes. Our approach utilizes Vision Transformer (ViT) and HateBERT, fine-tuned on the TOXIGEN dataset, which specifically focuses on implicit hateful speech. By employing early fusion techniques (such as contrastive learning, concatenation, and attention mechanisms), we effectively combine the strengths of both modalities, making our model more robust in detecting implicit hateful memes.

Keywords: Hate speech, multimodal, memes, early fusion, deep learning, CLIP, BERT, ViT, Hateful Memes Challenge, HarMeme

1. Introduction

In this paper, we focus on the classification of implicit memes. In 2021, Meta published the Hateful Memes dataset [1], a particularly challenging dataset due to its inclusion of benign confounders, which test a model's ability to understand both text and images for accurate classification. The winner of the Hateful Memes Challenge achieved an accuracy of 73.2%, while human performance stood at 84%, indicating significant room for improvement. Subsequently, Microsoft released the TOXIGEN dataset, which comprises highly implicit content designed to deceive existing hate-speech classifiers. TOXIGEN demonstrated that current models could be easily fooled, leading to the development of fine-tuned versions of HateBERT and RoBERTa, which improved accuracy from 57% to 96% [2].

Our approach combines Vision Transformers (ViT), introduced by Dosovitskiy et al. [3], with fine-tuned HateBERT for feature extraction from text and ViT for feature extraction from images. By leveraging these advanced models, we aim to achieve superior performance in identifying and moderating implicit hateful content. We utilize state-of-the-art methods for extracting each modality before fusing them with attention-based techniques for classification.

2. Related Work

2.1. Datasets

Kiela et al. [1] introduced the Hateful Memes Challenge, a benchmark designed to test the capabilities of multimodal models in identifying hate speech within memes. A thorough work has been done to ensure the quality of this dataset, they reconstructed the memes by changing the images while keeping the same meaning, reducing bias in the images and avoiding any potential noise from optical character recognition (OCR). In addition, it includes benign confounders to mitigate biases. The latter also induces classes imbalance in the dataset. Nonetheless, its complexity and thorough annotation process make it ideal for testing implicit hate speech detection systems. The images used in this dataset are under a license negotiated to allow redistribution for research purposes.

Pramanick et al. [4] introduced the HarMeme dataset, focusing on harmful memes related to COVID-19. The dataset contains 3,544 memes, annotated for harm intensity and target entities. This dataset has also gone through a rigorous two-stage annotation process. Notably, the HarMeme dataset consists of actual memes from real users, providing authentic data. However, collecting data during this period introduces limitations due to the specific trends and topics prevalent at that time.

2.2. MultiModal Models

To capture the complex nature of memes, unimodal models are not suited for such tasks as they can not leverage the often nuanced combination of text and image. This fact motivates us to investigate some of the most commonly used approaches to combine different modalities.

Concat Model: This model concatenates output features

[1] Group 13.

EE-559: *Deep Learning*, 2024.

for both image and text before final classifier layer. This approach allows the model to consider both textual and visual information simultaneously. However, it struggles to capture intricate interactions between the modalities, potentially limiting its effectiveness in detecting subtle or context-dependent harmful content [1].

Attention Model: Attention mechanisms improve model performance by allowing the network to focus on specific parts of the input, such as particular words in text or regions in images and relate them. Different approaches are valid in this case: self attention [5], and cross attention [6].

Contrastive Learning: Developed by OpenAI, the CLIP model learns visual concepts from natural language descriptions. It is trained to predict the correct pairings of a batch of training (image, text) examples to enforce similarity. From our findings, the best performing model across evaluated metrics is ISSUES, which is built on the CLIP architecture with 77.7% Acc. and 85.51% AUROC [7].

The initial results for both datasets reveal that while models like ViLBERT and VisualBERT outperform unimodal baselines, they still fall short of human performance, underlying the complexity of this task, and the need for more complex fusion techniques that capture hidden patterns.

2.3. Features Extraction Models

As our focus is on image and text modalities, some unimodal modals have been effectively pre-trained in the last years.

HateBERT: HateBERT is an English pre-trained BERT model obtained by further training the English BERT base uncased model with more than 1 million posts from banned communites from Reddit. It was introduced in 2021 by Caselli et al. and was trained on 3 datasets (OffensEval 2019, AbusEval, HatEval) [8]. When fine tuned on TOXIGEN, it achieved exeptional accuracy of 96%.

ViT_224: Vision Transformer (ViT) model pre-trained on ImageNet-21k (14 million images, 21,843 classes) at a resolution of 224x224, and fine-tuned on ImageNet 2012 (1 million images, 1,000 classes) at a resolution of 224x224. This particular model reaches an accuracy of 88.55% on ImageNet, 90.72% on ImageNet-ReaL, 94.55% on CIFAR-100, and 77.63% on the VTAB suite of 19 tasks [3].

3. Method

Our approach involves utilizing state-of-the-art pre-trained unimodal models acting as encoders to extract features from both text and images to feed into different multimodal models.

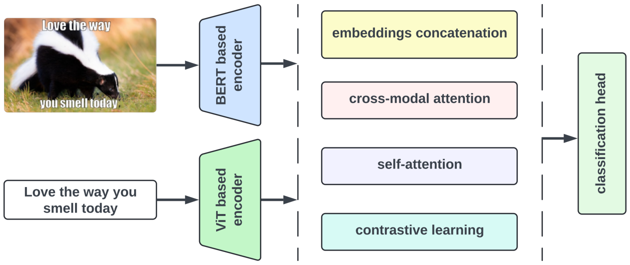

Figure 1. High level diagram of the pipeline where feature extraction and classification head stays the same for all our methods but the fusion is declined in 4 different implementations.

3.1. Datasets

We train, test and validate on HarMeme and Hateful Meme Challenge datasets separately. The HarMeme dataset has labels from three classes: harmless, partially harmful, and very harmful. We combine the last two classes in a single hateful class to match the other dataset. As a preprocessing step, we use basic transforms on images and tokenize our text inputs.

3.2. Encoder modules

As encoder modules, we use TOXIGEN roberta and TOXIGEN HateBERT for text modalities. Image embeddings are generated by resnet152, ViT_224 Efficientnet_b7. Best results were held by ViT_224 and TOXIGEN HateBERT, which we use for our evaluation benchmark. [5]

3.3. Combiner modules

3.3.1. Self-Attention

This attention model computes attention matrice for each embedding by multiplying weight for each model by its embedding. It then applies softmax to get attention weights for both of them. The scores for both models are concatenated and the weights are combined and multiplied by the total embedding for each item in the batch. This fused tensor integrates the important information from both text and image modalities based on the attention weights. The limitations for this model are the non exploration of relations between each models.

3.3.2. Cross-Attention

In this model, we are implementing a cross-attention mechanism. Text and image features are transformed into a common space using linear transformations. After normalization, cross-attention is applied, where the text features attend to the image features and vice versa, using additional linear transformations and the tanh activation. Attention scores are computed and used to weight the combined features. The weighted features are then summed to create a fused representation, which is passed through a dropout layer and a fully connected layer for classification. This model performs the best in our setup, it captures better the hidden relationship existing in memes.

3.3.3. Contrastive learning

The module optimizes two types of losses: the contrastive loss, which encourages similar text and image pairs to be close in the projection latent space, and the classification loss. The product of both losses encourages the model to penalize dissimilar memes wrongly classified.

4. Validation

Method	Acc.	AUROC
Concatenation	61.5	61.0
Cross-modal attention	**63.9**	**63.0**
Self-attention	60.9	60.6
Contrastive learning	49.4	50.6
Human	84.7	-
VisualBERT COCO	69.5	75.4
ViLBERT CC	65.9	74.5
HMC 1st prize [9]	73.2	84.5
Hate-CLIPper [10]	77.2	84.3
ISSUES [7]	**77.7**	**85.5**

Table 1. Hateful Meme Dataset comparision of Acc and AUROC for the best baseline method (VisualBERT and ViLBERT) [1], the winner of the competition and the state of the art methods with our 4 implementations on the test seen data. TOXIGEN HateBert and ViT_224 were used for embeddings

Method	Acc.	F1	AUROC
Concatenation	78.9	58.2	70.8
Cross-modal attention	**80.7**	57.5	**71.1**
Self-attention	76.8	53.1	68.2
Contrastive learning	55.6	40.2	54.0
Human	84.7	-	-
VisualBERT COCO	75.8	65.8	-
ViLBERT CC	72.6	57.2	-
Hate-CLIPper [10]	83.9	-	91.9
ISSUES [7]	**81.6**	-	**92.83**

Table 2. HarMeme comparison of Acc, F1 and AUROC for the best baseline method (VisualBERT and ViLBERT) [4] and the state of the art methods with our 4 implementations on the test data. We used TOXIGEN HateBERT and ViT_224 for embeddings

The accuracy obtained by our best model is 63.9% (AUC: 63.0), which is lower than the baseline performance of pre-trained multimodal models such as VisualBERT COCO, which achieves an accuracy of 69.5% (AUC: 74.4%). The challenging nature of the Hateful Memes dataset, due to benign image confounding, tests the model's ability to understand the relationship between text and image. Our model demonstrates limitations in grasping the intricate interactions between these modalities, even when employing cross-modal attention mechanisms.

To further contextualize these results, we compared our model's performance on the Hateful Memes dataset with its performance on the HarMeme dataset, where it achieved an accuracy of 80.7% (AUC: 70.8%). This comparison indicates that while our model can learn and classify effectively in less challenging contexts, it struggles significantly with the more complex Hateful Memes dataset. We attribute these results to the dataset's inherent difficulty in requiring nuanced understanding of the interplay between text and images.

A model built and pre-trained using multimodal dataset such as COCO manage to better grasp the context and meaning of the meme. Hate-CLIPper and ISSUES are based on CLIP architecture from pretrained model and fined tuned on both dataset achieved much better accuracies for the Hatefeul Memes Dataset.

When comparing the results for the HarMeme dataset, we managed to beat the baseline with 80.7% accuracy and 57.5% F1 score but we are still far away from ISSUES and Hate-CLIPper.

We can see that using state of the art feature extraction with most accurate textual and visual models such as TOXIGEN HateBERT and ViT_224 is not enough when context in image and text is different. To alleviate this we need to focus on CLIP based architecture of train a model with multi-modal dataset so that the model learns the intricacies between text and image better and then fine-tune it on our dataset.

5. Conclusion

In this paper we used different approaches for a multimodal classification of implicit hateful memes. Our best model was the cross-attention model, which compared to baselines scores better in the HarMeme dataset, but falls short in the HMC dataset [1] evaluated against other models. One approach to improve the way we capture the complexity of the dataset, is by improving our contrastive learning approach, this is inspired by the Hate-CLIPer model. In the context of memes, we do not want the images and text to capture similar things because memes often use implicit information that usual multimodal models for image and text wouldn't capture. To address this, we need to retrain the encoder projections of each uni-modal model to better capture this implicit nature in memes.

References

[1] A. M. Douwe Kiela, Hamed Firooz, "The hateful memes challenge: Detecting hate speech in multimodal memes."

[2] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, and E. Kamar, "Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[3] A. K. D. W. X. Z. T. U. M. D. M. M. G. H. S. G. J. U. N. H. Alexey Dosovitskiy, Lucas Beyer, "An image is worth 16x16 words: Transformers for image recognition at scale."

[4] R. M. S. S. M. S. A. P. N. T. C. Shraman Pramanick, Dimitar Dimitrov, "Detecting harmful memes and their targets."

[5] G. S. Q. H. Yiling Wu, Shuhui Wang, "Learning fragment self-attention embeddings for image-text matching."

[6] A. C. Vandana Rajan, Alessio Brutti, "Is cross-attention preferable to self-attention for multi-modal emotion recognition?."

[7] L. A. M. B. A. D. B. Giovanni Burbi1, Alberto Baldrati1, "Mapping memes to words for multimodal hateful meme classification."

[8] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for abusive language detection in English," in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, eds.), (Online), pp. 17–25, Association for Computational Linguistics, Aug. 2021.

[9] R. Zhu, "Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution," 2020.

[10] G. K. Kumar and K. Nandakumar, "Hate-clipper: Multimodal hateful meme classification based on cross-modal interaction of clip features," 2022.