"Female Astronaut: Because sandwiches won't make themselves up there!": Towards multi-modal misogyny detection in memes

Smriti Singh*
Dept. of Computer Science
UT Austin

Amritha Haridasan*
Dept. of Computer Science
UT Austin

Raymond Mooney
Dept. of Computer Science
UT Austin

Abstract

A rise in the circulation of memes has led to the spread of a new form of multimodal hateful content. Unfortunately, the degree of hate women receive on the internet is disproportionately skewed against them. This, combined with the fact that multimodal misogyny is more challenging to detect as opposed to traditional text-based misogyny, signifies that the task of identifying misogynistic memes online is one of utmost importance. To this end, the MAMI dataset was released, consisting of 12000 memes annotated for misogyny and four sub-classes of misogyny - shame, objectification, violence and stereotype. While this balanced dataset is widely cited, we find that the task itself remains largely unsolved. Thus, in our work, we[1] investigate the performance of multiple models in an effort to analyse whether domain specific pretraining helps model performance. We also investigate why even state of the art models find this task so challenging, and whether domain-specific pretraining can help. Our results show that pretraining BERT on hateful memes and leveraging an attention-based approach with ViT outperforms state of the art models by more than 10%. Further, we provide insight into why these models may be struggling with this task with an extensive qualitative analysis of random samples from the test set.

1 Introduction

With a rise in social media usage, memes have become an important part of expression, and communication today. Multiple research studies have found that memes play a role in shaping a wide range of beliefs, such as climate change, use as bonding icons, political discussion, and social development. This new form of media, however, is still host to old-school offensive content that was previously seen in non-multimodal settings. This includes hate speech of different forms, such as sexism and racism. The emergence of this popular media format has brought along the need to detect hateful content in multimodal formats, to ensure that the internet remains a safe space for all groups. Further, there has been evidence to show that women are disproportionately targeted on the internet. For example, 33% of women under 35 say they have been sexually harassed online, while 11% of men under 35 say the same[2]. It has also been shown through many psychological and social science-based studies that the effects of online hate speech are observed well beyond the boundaries of the cyber world (Pluta et al., 2023). Yet, traditional, language-based misogyny detection techniques are no longer fully effective when it comes to multimodal misogyny. This is because, unlike text-based misogyny, identifying multimodal misogyny involves picking up on visual cues combined with sarcasm and linguistic nuances.

To try and bridge this challenge, Fersini et al. (2022) developed, licensed, and released MAMI: Multimedia automatic misogyny detection, a dataset of 12000 memes, labeled for misogyny and four subclasses – shaming, objectification, violence, and stereotypes. The dataset is balanced across all classes and was released as a part of the SemEval Task in 2022. While this dataset is widely cited, and there have been multiple approaches developed to leverage this dataset for misogyny detection, we find that this challenging task remains unsolved to a large extent. To the best of our knowledge, no research aims to understand exactly why even the best models are unable to succeed at this task. Moreover, there is no research (to the best of our knowledge) that showcases the potential benefit (or lack thereof) of using models pre-trained on other hate-speech data. Therefore, the focus of our work is two-fold: Thus, in our work, instead of

[1]/* denotes equal contribution

[2]https://www.forbes.com/sites/ewelinaochab/2023/03/08/when-the-harassment-of-women-moves-online/?sh=3a9d64223f29

solely focusing on developing a model that outperforms the current state-of-the-art architectures, we focus on the following broader research questions:

- Do multimodal models understand misogyny in memes better than language-only or vision-only models?
- Do these models benefit from pre-training on text hate-speech datasets?
- What can't these models do? What mistakes do they make? Is there a pattern that can be observed in their mistakes?

Our contributions, per the aforementioned research questions, are as follows:

- We present a multimodal model, BERT*+VIT, that is pre-trained on hate-speech text data, finetuned on the MAMI dataset.
- An extensive quantitative analysis of the performance of various state-of-the-art models when fine-tuned on the MAMI dataset – text only, language only, and multimodal.
- A qualitative analysis of the mistakes made by different models.

The rest of this paper is organized as follows: Section 2 describes the related work, Section 3 elaborates on the experiments we conduct as a part of our methodology and Section 4 summarizes the results obtained.

2 Related Work

One of the first large-scale challenges that involved detecting hateful memes is the 'Hateful memes challenge' organized by Facebook AI (Kiela et al., 2020). To quote the authors, "Memes pose an interesting multimodal fusion problem: Consider a sentence like "love the way you smell today" or "look how many people love you". Unimodally, these sentences are harmless, but combine them with an equally harmless image of a skunk or a tumbleweed, and suddenly they become mean." They release the hateful memes dataset, consisting of 10,000 memes annotated for unimodal hate, multimodal hate, benign text, benign image, and random non-hateful examples.

This was followed by many research efforts to categorize memes beyond hateful, such as Zia et al. (2021), who looked at classifying memes as racist and/or sexist, Nafiah and Prasetyo (2021) who focused on analyzing and identifying sexist memes during the COVID-19 pandemic, and Suryawanshi et al. (2020) who used the presidential election to develop a dataset of memes consisting of racism, sexism and homophobia.

The MAMI dataset (Fersini et al., 2022) was the first of its kind to motivate the sub-classification of misogynistic memes. This task, as a part of the SemEval 2022 contest, showcased many noteworthy methodologies for the proposed problem. For example, Sharma et al. (2022b) proposed an R2D2 architecture that used pre-trained models as feature extractors for text and images. They used these features to learn multimodal representation using methods like concatenation and scaled dot product attention. This methodology achieved an F1 score of 0.757 and was ranked 3rd in Subtask, and 10th on Subtask B, with an F1 score of 0.690. In another study, Mahadevan et al. (2022) develop an ensemble model consisting of XLM-RoBERTa, DistilBERT, ResNext, and Data-efficient Image Transformer to achieve an average F1 of 0.71 on Task A and 0.69 on Task B. However, these authors established an SVM as their baseline. Our goal is to explore a wider range of similar models, using such models as a baseline, and hopefully develop a better one. For now, we plan to use precision, recall, and F1 score as our evaluation metrics, along with a manual qualitative analysis that can provide insight into how to better direct future model improvements.

Another interesting approach is that proposed by Muti et al. (2022), which combines BERT and CLIP, achieving an F1 of 0.727 on sub Task A, and an F1 of 0.710 on sub Task B. Kalkenings and Mandl (2022) extends a similar approach by using BERT and FCNN, and testing it on the aforementioned Facebook AI's hateful meme challenge dataset for generalisability. An approach that is similar to ours to an extent is that of Sharma et al. (2022a), who test a variety of language models on the text part of the MAMI dataset. Our approach involves using such models to establish a comparative baseline of language-only models and combine them with vision-based models to analyze how that affects model performance. Finally, in another noteworthy experiment, Hakimov et al. (2022) proposes a CLIP text encoder and an LSTM for the text encoding part of the model. This model attains an F1 score of 0.834 on subtask A and an F1 score of

0.731 on subtask B.

In our work, we want to look beyond merely training a classifier that outperforms these methods. We are more interested in analyzing the finer details, and understanding *what* these models are doing wrong. Further, we are interested in establishing comparison baselines through text models and vision models to gain insight into which feature is more important, and to what extent. To the best of our knowledge, prior to this, there has been no experimentation to show the benefits of using pretrained models for multimodal misogyny detection.

3 Methodology

As discussed in Section 2, following the SemEval Task itself, we will refer to Task A as the classification of a meme as misogynistic and Task B as the subclassification of a misogynistic meme. Further, all models are finetuned on this dataset. As described later in the paper, BERT* benefits from pretraining on the hateful meme dataset.

3.1 MAMI: Dataset description

Although the MAMI dataset has been well described in the original paper (Fersini et al., 2022), we provide a summary of it here, for a holistic understanding of the experiments conducted in this study.

This dataset consists of 12,000 memes. The breakdown of these memes for train-test-dev is 10,000 - 1,000- 1,000 respectively. Further, the misogynistic memes are classified into four subclasses as mentioned above. The distribution across subclasses is shown below in Table 1.

The process of gathering pertinent memes for analysis involved searching popular social media platforms like Twitter and Reddit, as well as accessing dedicated meme creation and sharing websites such as 9GAG, Knowyourmeme, and Imgur. To ensure an adequate number of misogynous memes, the researchers undertook activities such as searching for meme threads focused on women, exploring discussions by individuals with anti-women or anti-feminist sentiments, investigating recent events. By employing these methods, a diverse dataset of relevant memes was successfully compiled for further examination in their study.

The authors found a coefficient of 0.5767 for agreement on misogynous vs. not misogynous annotations, and a coefficient of 0.3373 for the type of misogyny labeling. The dataset details are presented in Table 2. The Fleiss-k measure indicated moderate agreement for misogynous labeling, indicating a relatively straightforward task for humans. However, the agreement for the type of misogyny annotation was fair, suggesting a more challenging task.

Subclass	**Train**	**Test**
Shaming	1274	126
Stereotype	2810	350
Objectification	2202	348
Violence	953	153

Table 1: Distribution of misogynistic memes across subclasses

The dataset is evenly split between misogynistic and non-misogynistic memes with 5000 samples in the train and test set each,

3.2 Unimodal models

To establish baseline models, we experiment with a variety of language and vision models. For language models, we use the text from the memes and finetune the following models:

- BERT
- DeBERTa
- RoBERTa
- Hateful memes pre-trained BERT

Here, the last model is a model hosted on HuggingFace that has been pre-trained on the text from the hateful memes dataset released by Facebook AI [3]. Similarly, we train the following vision models on the memes to establish vision-only baselines:

- CNN
- Inception
- ViT

We showcase the performance of these models in Section 4.

[3] https://huggingface.co/am4nsolanki/autonlp-text-hateful-memes-36789092

	Misogyny Labelling (Sub-task A)			Type of Misogyny Labelling (Sub-task B)				
	Misogynous	Not Misogynous	Fleiss-k Agreement	Shaming	Stereotype	Objectification	Violence	Fleiss-k Agreement
Training Set	5000(50%)	5000(50%)	0.5767	1274(25.48%)	2810(56.20%)	2202(44.04%)	953(19.06%)	0.3373
Test Set	500(50%)	500(50%)	0.5767	146(29.20%)	350(70.00%)	348(69.60%)	153(30.60%)	0.3373

Table 2: Dataset Characteristics (Fersini et al., 2022)

3.3 Multimodal models

3.3.1 CLIP

CLIP (Contrastive Language-Image Pre-training) (Radford et al., 2021) is a state-of-the-art language and vision model developed by OpenAI. It is capable of understanding images and natural language text and can perform a range of tasks such as image classification, object detection, and captioning. The model has been trained on a large dataset of image-text pairs, allowing it to learn the correlations between visual and textual features. One of the unique features of CLIP is that it uses a contrastive learning approach, which means that it learns by comparing and contrasting similar and dissimilar image-text pairs.

The CLIP model consists of two parts: a vision encoder and a language encoder. The vision encoder is a convolutional neural network (CNN) that takes in an image and outputs a vector representation of the image. The language encoder is a transformer-based model that takes in natural language text and outputs a vector representation of the text. For our research, we finetuned the model with the Adam optimizer with a learning rate of 1e-4, weight decay of 0.01, and a batch size of 16. The maximum number of epochs is limited to 20, and early stopping is implemented with a patience of 3 epochs.

3.3.2 BERT + Inception

The BERT + Inception model (Guda et al., 2020) is a deep learning model that combines two different neural networks, BERT and Inception, to achieve better performance on image-text matching tasks. BERT, or Bidirectional Encoder Representations from Transformers (Devlin et al., 2018), is a pre-trained language model that excels at natural language processing (NLP) tasks, such as sentiment analysis and text classification. On the other hand, Inception is a convolutional neural network (CNN) (Szegedy et al., 2016) that is well-suited for image recognition and classification tasks. By combining these two models, the BERT + Inception model can effectively encode both text and image inputs and map them to a common latent space for matching.

The text encoder uses the BERT architecture, which is pre-trained on a large corpus of text data. The BERT model is used to encode textual descriptions into a fixed-size vector. The image encoder uses the InceptionV3 architecture, with the weights pre-trained on the ImageNet dataset. The InceptionV3 model is modified to remove the top classification layer and replace it with a global average pooling layer to generate a fixed-size feature vector for each input image. The sequence and pooled outputs from the text input are concatenated with the processed image input and passed through three dense layers with ReLU activation and dropout layers. The purpose of these dense layers is to combine the information from both the image and text encoders and generate a more informative representation for the final classification. The model is trained using a contrastive loss function (Alluri and Krishna, 2021) that encourages the image and text representations to be similar for positive pairs, and dissimilar for negative pairs. The batch size used in the training loop is 256. The model is trained for 10 epochs in each iteration of the training loop, and early stopping is implemented with a patience of 5 epochs, with a learning rate of 1e-4.

3.3.3 BERT + ViT

BERT is designed for processing text data and does not take into account the visual information present in many modern datasets. The Vision Transformer (Dosovitskiy et al., 2020) is a neural network architecture that has been specifically designed for processing visual information, such as images or videos. It uses a self-attention mechanism to analyze and process visual information, allowing it to learn complex patterns and relationships between different elements in an image. By combining BERT with the Vision Transformer (Velioglu and Rose, 2020), we can create a

powerful hybrid architecture that can process both text and visual information simultaneously.

The model has three input layers – one for the image input, one for the text input, and one for the input masks. The text input is processed using the BERT model to encode the input text into contextualized embeddings, and the image input is processed using the Vision Transformer (ViT) model to flatten images into patches to linearly project and combine with position encoding. The output features of the two models are combined using an attention mechanism and then passed through a 1D convolutional layer and a flatten layer to create joint features. The final output of the model is a probability distribution over the possible classes, which is obtained by passing the joint text and image embedding features through one or more fully connected layers with sigmoid activation. During training, the model is optimized using backpropagation and stochastic gradient descent with cross-entropy loss. The model was trained using Adam optimizer with a learning rate of 4e-5 for 5 epochs and a batch size of 16 was used.

3.3.4 VisualBERT

VisualBERT (Li et al., 2019)is a pre-trained model that combines the power of the BERT architecture with visual features to understand language in the context of images. The architecture of VisualBERT consists of two separate encoders, one for the visual modality and one for the textual modality. The visual encoder processes the image and extracts visual features, which are then combined with the textual features extracted by the textual encoder. These features are then fed into the BERT model for further processing, allowing the model to understand the relationship between the image and the text. VisualBERT uses a hierarchical approach to process the visual information, starting with low-level visual features and gradually moving up to more abstract concepts.

In VisualBERT, (Muennighoff, 2020) the image features extracted from pre-trained object proposal systems, such as Faster-RCNN, are treated as input tokens, just like words in a text. These image features are unordered, meaning they are not processed in any particular sequence or order. Along with the text, the image features are fed into the multi-layer Transformer architecture of VisualBERT, where they are processed and used to build a joint representation of the text and image. This allows the model to capture the intricate associations between text and image and enables it to perform tasks that require understanding the semantics of both modalities. The model was fine-tuned using an Adam optimizer with a learning rate of 2e-5, training for 10 epochs with a batch size of 32.

3.3.5 BERT* + ViT

Here, BERT* refers to a BERT model which is pre-trained on the hateful memes dataset released by Facebook AI. We propose using the model, BERT* + ViT which is a model that combines two powerful neural networks, a domain-specific pre-trained BERT model and ViT (Sohn and Lee, 2019). The image is passed through the Vision Transformer (ViT), while the text is passed through the BERT model. The text sequence embedding and image embedding are then combined using an attention mechanism, which attends to the relevant parts of the image based on the text input.

The model has achieved state-of-the-art performance on several benchmark datasets for hate speech detection (d'Sa et al., 2020). The use of both text and image information improves the model's ability to detect subtle nuances in hate speech and non-hate speech messages. The attention mechanism allows the model to attend to the most relevant features in both modalities and combine them to make a prediction. The convolutional layer further refines the joint features obtained from the two models, and the final dense layer predicts the probability of hate speech. The model has been pre-trained on a large corpus of hate speech data, making it highly effective at detecting hate speech in real-world scenarios. The training process utilized the Adam optimizer with a learning rate of 4e-5 for a duration of 5 epochs, and the training data were processed in batches of 16.

4 Results

4.1 Comparison of Baseline Models and Multimodal Models

Tables 3 and 4 answer our first two research questions about how each baseline compares to multimodal models, and how domain-specific pretraining may be useful to the model.

Model	Precision	Recall	F1	Modality
BERT	0.662	0.650	0.643	Lang
DeBERTa	0.684	0.682	0.681	Lang
RoBERTa	0.632	0.628	0.620	Lang
BERT*	0.685	0.695	0.690	Lang
CNN	0.571	0.782	0.616	Vision
ViT	0.611	0.659	0.632	Vision
Inception	0.511	0.672	0.623	Vision
BERT + Inception	0.623	0.778	0.694	Both
BERT + ViT	0.624	0.890	0.734	Both
BERT* +ViT	**0.862**	**0.881**	**0.874**	**Both**
CLIP	0.655	0.782	0.652	Both
VisualBERT	0.623	0.687	0.666	Both

Table 3: Performance of various finetuned models on subtask A. BERT* denotes the BERT model that has been pretrained on the hateful memes dataset.

Here, we find that BERT*+VIT outperforms even the top-ranking models described in Section 2 by a considerable margin. This indicates that domain-specific pretraining is indeed, quite useful in boosting model performance.

We also observe that apart from the model that has the advantage of domain-specific pretraining, the other models don't have a very large difference in terms of F1 scores. However, as one would expect, we see that the vision-only baselines are a bit lower than the text-only baselines. Multimodality can help significantly, for example, adding ViT improves the F1 score of BERT by 9 points. However, we find that some multimodal models actually perform worse than unimodal models (for instance, CLIP and DeBERTa). This implies a need for investigation as to what may be confounding the multimodal models, which we present our analysis for in Section 4.2.

Table 4 shows the performance of these models on subtask B. For this subtask, we record the F1 score for each class to ensure readability. Here, we see that the pre-trained multimodal model outperforms the others by a small margin. The table indicates that memes containing violence and objectification are easier to detect compared to the other classes, regardless of the model used. This is probably due to the fact that these are the most non-ambiguous memes, i.e., the classes where the memes (especially likely the text) often have only one meaning. This is discussed further below.

4.2 Qualitative Analysis

To answer our last research question, we present an extensive qualitative analysis of 200 randomly sampled memes from the test set. Our goal is to find potential patterns in errors made by the best-performing model. This is, to the best of our knowledge, the first time an error analysis is being performed for any model finetuned on the MAMI dataset. Our observations are as follows:

4.2.1 Visual Grounding in itself is not enough

For around 80 memes (out of the 200 randomly sampled ones), we find that the incorrect prediction might be owing to the fact that even visual grounding is not enough. This is particularly true for memes belonging to the stereotype class, but can occasionally apply to the shaming class too. We show an example in Figure 1. The idea here is that the model needs to be able to understand the stereotype behind the image/text combination, and simply looking at the memes may not provide that.

4.2.2 Lack of Context

Some memes (around 15) lack context in terms of exactly how they are offensive. These are memes that might prove challenging to classify even for humans. One such example is shown in Figure 2.

4.2.3 Lack of understanding of subtle objectification

Most of the memes that were randomly sampled from the objectification class are quite explicit in nature. However, the ones that have a lower degree of objectification/sexuality often go undetected by the model. For example, in Figure 3, it is not

Model	Stereotype	Shaming	Objectification	Violence
BERT	0.633	0.612	0.688	0.689
DeBERTa	0.651	0.627	0.655	0.678
RoBERTa	0.624	0.618	0.677	0.682
Pretrained BERT	0.644	0.683	0.685	0.691
CNN	0.551	0.582	0.577	0.613
ViT	0.541	0.608	0.591	0.612
Inception	0.595	0.605	0.613	0.599
BERT + Inception	0.644	0.621	0.685	0.690
BERT + ViT	0.653	0.631	0.695	0.710
BERT*+ViT	**0.697**	**0.695**	**0.693**	**0.721**
CLIP	0.648	0.628	0.655	0.684
VisualBERT	0.647	0.623	0.676	0.683

Table 4: Performance of various finetuned models on subtask B. BERT* denotes the BERT model that has been pretrained on the hateful memes dataset.

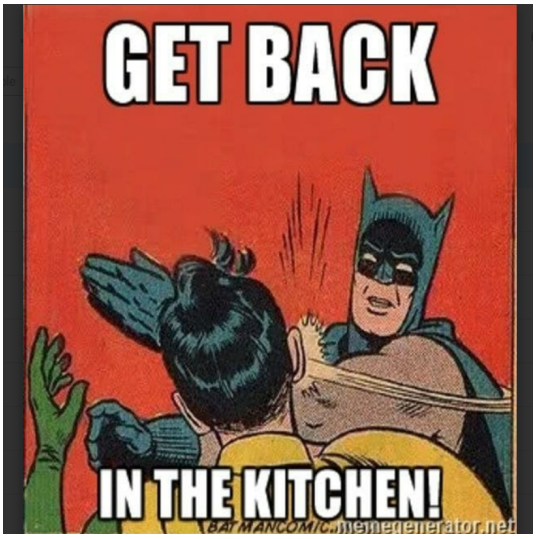

Figure 1: An example where it is necessary to understand the stereotype that "women belong in the kitchen" is misogynistic. The model misclassifies this as neutral.

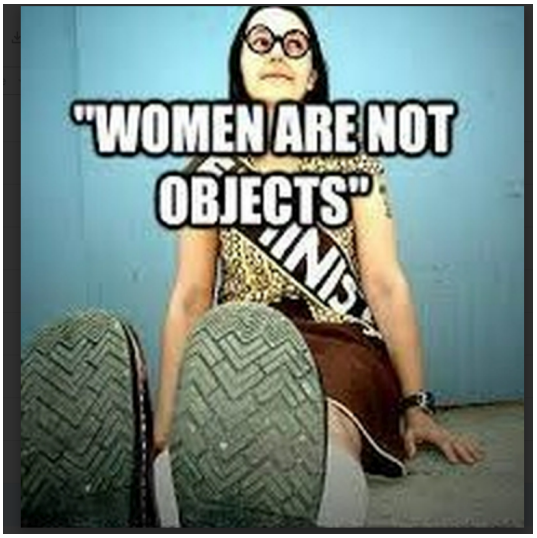

Figure 2: An example where context is unclear. The model marks this as violent.

enough to look at the image in itself, because all that is visible is a woman posing for the camera. It is also not enough to read the text. Some social context is needed to interpret that this meme could be hinting toward sex trafficking or other illicit similarities.

4.3 Analysis of benefits of pre-training

In this part of the paper, we provide examples of where pre-training helps BERT* classify complicated memes that require an understanding of visuolinguistic cues.

4.3.1 Understanding complex objectification

We find that by pre-training on hateful memes, the model is able to identify memes that objectify us-

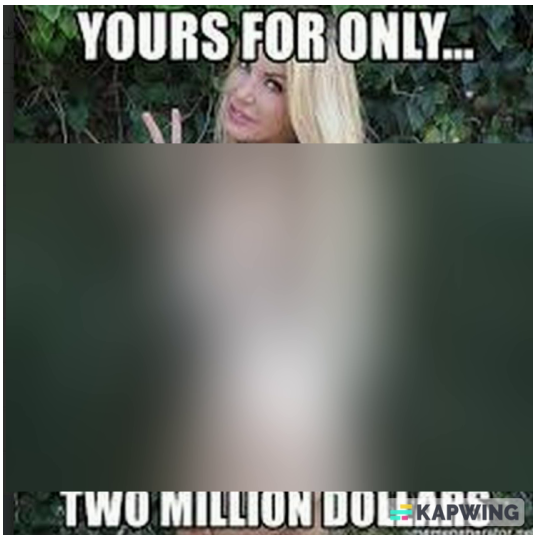

Figure 3: An example where objectification is subtle. The model misclassified this as shaming.

ing a combination of visual cues and complicated linguistic cues. Figure 4 shows an example of this phenomenon.

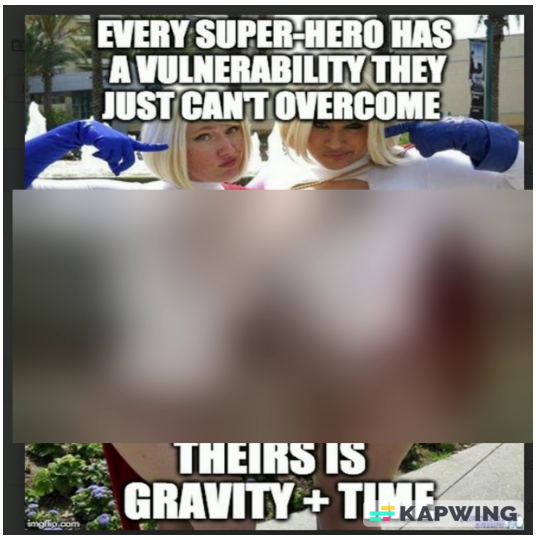

Figure 4: An example where the objectification is through a combination of visuo-linguistic cues. BERT* correctly classifies this as objectification.

4.3.2 Understanding lewd complex linguistic cues

Our results indicate that pretraining helps BERT* pick up on some complex linguistic innuendos that are offensive. The reason for this is somewhat unclear, but we hypothesize that the model benefits from a larger exposure to multimodal hateful content. For example, consider the meme shown in Figure 5. The model correctly identifies it as shaming, This means that it understands that losing a shoe in this case is suggestive of provocative shaming. This example doesn't rely on any visual cues, but the model is still able to classify it correctly.

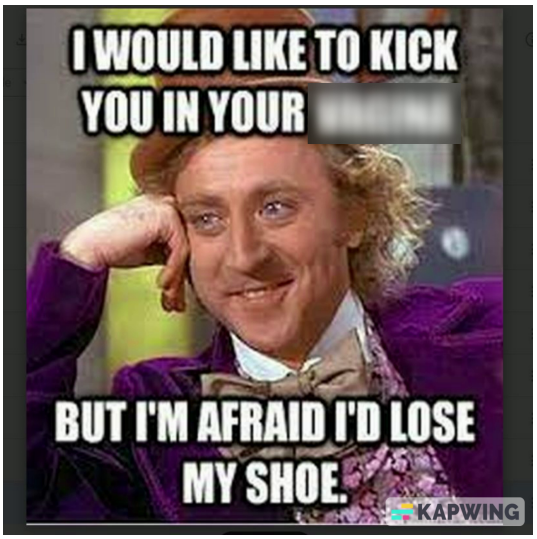

Figure 5: An example of BERT* correctly identifying complex linguistic cues

4.3.3 Connecting seemingly harmless text with objectifying images

While identifying misogynistic memes that are hateful through subtle visual cues and otherwise seemingly harmless text is still a challenge in this area, we find that BERT* benefits from pretraining to at least identify these memes correctly, if not subclassify them properly. For example, BERT* marks the meme shown in Figure 6 correctly as misogynistic in subtask A, but makes an error in classifying it as shaming instead of objectification, our hypothesis is that the model may further benefit from pretraining/finetuning on larger datasets that contain more examples of misogyny.

5 Conclusion

In this research, we have delved into the challenging task of identifying misogynistic memes online. By utilizing the MAMI dataset with 12,000 annotated memes, we have established baselines and conducted experiments with various models, including text-only, vision-only, and multimodal models. Our findings indicate that pretraining BERT on hateful memes and utilizing an attention-based approach with ViT performs better than the state-of-the-art models by more than 10% for subtask A, and by 2% on subtask B. This highlights the importance of domain-specific pretraining in identifying multimodal misogyny. Further, we have

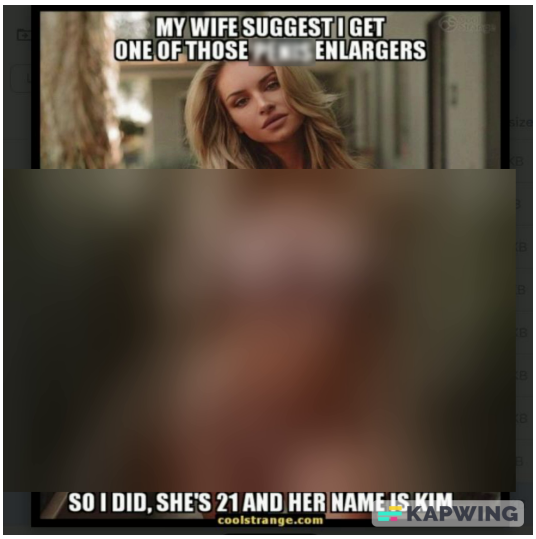

Figure 6: An example where BERT* benefits from pretraining is being able to identify this meme as misogynistic, but fails to subclassify it correctly.

provided a comprehensive qualitative analysis of random samples from the test set, which provided insight into the challenges of detecting multimodal misogyny. Our research emphasizes that identifying misogynistic memes online is a complex task that necessitates a thorough consideration of both visual and linguistic cues, and the significance of domain-specific pretraining in this area. Future work includes extending the dataset to multiple languages to evaluate the generalizability of the proposed approach beyond English. Additionally, a similar analysis could be performed on multimodal media such as reels and TikToks to assess the effectiveness of the proposed approach on these platforms. Further research is also needed to reduce the computational complexity of training and deploying these models for downstream tasks. Finally, investigating the interpretability of the proposed approach could shed light on which multimodal cues are most indicative of misogyny, thereby helping to better understand the underlying mechanisms of this phenomenon.

References

Nayan Varma Alluri and Neeli Dheeraj Krishna. 2021. Multi modal analysis of memes for sentiment extraction. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, volume 6, pages 213–217. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Ashwin Geet d'Sa, Irina Illina, and Dominique Fohr. 2020. Bert and fasttext embeddings for automatic detection of toxic speech. In *2020 International Multi-Conference on:"Organization of Knowledge and Advanced Technologies"(OCTA)*, pages 1–5. IEEE.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Bhanu Prakash Reddy Guda, Sasi Bhushan Seelaboyina, Soumya Sarkar, and Animesh Mukherjee. 2020. Nwqm: A neural quality assessment framework for wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8396–8406.

Sherzod Hakimov, Gullal S Cheema, and Ralph Ewerth. 2022. Tib-va at semeval-2022 task 5: A multimodal architecture for the detection and classification of misogynous memes. *arXiv preprint arXiv:2204.06299*.

Milan Kalkenings and Thomas Mandl. 2022. University of Hildesheim at SemEval-2022 task 5: Combining deep text and image models for multimedia misogyny detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 718–723, Seattle, United States. Association for Computational Linguistics.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33:2611–2624.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Shankar Mahadevan, Sean Benhur, Roshan Nayak, Malliga Subramanian, Kogilavani Shanmugavadivel, Kanchana Sivanraju, and Bharathi Raja Chakravarthi. 2022. Transformers at semeval-2022 task 5: A feature extraction based approach for misogynous meme detection. In *Proceedings of the 16th International*

Workshop on Semantic Evaluation (SemEval-2022), pages 550–554.

Niklas Muennighoff. 2020. Vilio: State-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Arianna Muti, Katerina Korre, and Alberto Barrón-Cedeño. 2022. UniBO at SemEval-2022 task 5: A multimodal bi-transformer approach to the binary and fine-grained identification of misogyny in memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 663–672, Seattle, United States. Association for Computational Linguistics.

Asriatun Nafiah and Dimas Teguh Prasetyo. 2021. Sexist memes related to covid-19 pandemic in social media, is it matter? In *Proceeding Conference on Genuine Psychology*, volume 1, pages 82–94.

Agnieszka Pluta, Joanna Mazurek, Jakub Wojciechowski, Tomasz Wolak, Wiktor Soral, and Michał Bilewicz. 2023. Exposure to hate speech deteriorates neurocognitive mechanisms of the ability to understand others' pain. *Scientific Reports*, 13(1):4127.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Gagan Sharma, Gajanan Sunil Gitte, Shlok Goyal, and Raksha Sharma. 2022a. IITR CodeBusters at SemEval-2022 task 5: Misogyny identification using transformers. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 728–732, Seattle, United States. Association for Computational Linguistics.

Mayukh Sharma, Ilanthenral Kandasamy, and Vasantha W B. 2022b. R2D2 at SemEval-2022 task 5: Attention is only as good as its values! a multimodal system for identifying misogynist memes. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 761–770, Seattle, United States. Association for Computational Linguistics.

Hajung Sohn and Hyunju Lee. 2019. Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In *2019 International Conference on Data Mining Workshops (ICDMW)*, pages 551–559. IEEE.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 32–41.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*.

Haris Bin Zia, Ignacio Castro, and Gareth Tyson. 2021. Racist or sexist meme? classifying memes beyond hateful. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 215–219.