Hateful Meme Classification

Aditi Ramesh Athreya, Shanthi Lekkala

University of Colorado Boulder

Abstract. Every day, millions of memes are created and shared on social media platforms. Although it is a fantastic way to spread humor, some individuals abuse it by creating offensive memes to disparage particular individuals or groups. Therefore, in order to stop the spread of hate, it is crucial to identify hate memes. To categorize these memes, one needs to have a working understanding of both the meme's visual and linguistic components. In this paper, we use VisualBERT to explore the Visio-linguistic features of the meme, Image captioning model with BERT to generate an interpretable textual representation of the image and combine them to perform classification. When the model's performance is assessed using the AUROC score and accuracy, it is found that our model performs better than the baseline model, VisualBERT, demonstrating how the performance was enhanced by image captioning.

Keywords: Meme Classification, Image Captioning, BERT, VisualBERT, AUROC, Accuracy

1 Introduction

Memes are one of the best tools available today for spreading hatred on the internet. The widespread use of social media has aided in the dissemination of hate speech and messages that are prejudiced against members of minority groups. Memes are frequently used to spread anti-Semitic, alt-right, and even neo-Nazi propaganda on Reddit, 4chan, and other well-known social media platforms. TheDonald and 4chan produced a sizable portion of the recent wave of racist memes that spread from niche to mainstream media platforms. Over 600,000 racist memes were shared from this community alone over a 13-month period, accounting for about 5% of all memes posted. Flagging hateful memes before they spread can help prevent future acts of violence and harm, as well as help reduce the growing community division caused by such hateful content online.

In 2020, Facebook created a hate meme classification dataset and developed baseline unimodal and multimodal models. On the same dataset, trained humans had an AUROC score of 82.65 for classifying hate memes, but the baselines performed poorly and had a significant performance gap with humans. Later, they hosted a challenge on the Drivendata platform for the same, where many approaches were explored and many models outperformed the baseline. One example is Zhu's[1] model, which topped the leaderboard with an AUROC score of 84.50 by using race tags for the face. This means that the current state-of-the-art model outperforms humans. But, most of the existing solutions are an ensemble of multiple Visio-linguistic transformer models, and deploying them in real-world applications can be expensive.

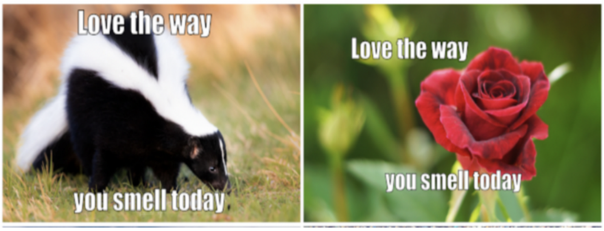


Fig. 1. Examples of hate and benign memes

The Hateful Meme Challenge leaderboard solutions that achieved high performance are large ensemble models that cannot be deployed in the real world. Our plan is to suggest a cutting-edge model with excellent performance and low computational cost. We plan to use a fusion of VisualBERT and BERT to classify hateful memes, this model will have high performance as it is a transformer-based model using both image understanding and meme text, and it should be computationally less expensive compared to an ensemble model. Though, we also have other models which don't use ensemble which achieved reasonable performance, they don't use image captioning with a transformer-based model, which makes this proposed model different from others. In this model, we begin by feeding the image captions generated from meme image to language model BERT, then feeding the same meme with image and text to multi-modal VisualBERT, the output from both the models are combined and sent to Dense model with fully connected layers to classify memes as hateful or not. By the end of this project, we hope to implement this model for hateful meme detection and evaluate how well it performs in comparison with the baseline model, VisualBERT and also the current state-of-the-art models.

2 Related Work

As the use of social media has grown, detecting hate speech has become more important. There are several text-based datasets and a substantial amount of research on hate speech detection, with cutting-edge models achieving very high performance using machine learning[2, 3] and deep learning[4, 5] techniques. While these models focus on text-based hate, detecting hate when the language is combined with images as in memes is also important.

Gomez et al.[6] proposed a novel task of detecting hate in multimodal publications for which they manually annotated dataset, and labeled it into six categories. They extracted features from text and images and fused them to classify the data points. Initially, image classifiers or transformers for language are used unimodally to classify memes. However, the text or image by itself is insufficient to properly categorize the meme. The later approaches used multimodal methods with either early or late fusion[7, 8] of image and text features from the unimodal models. Additionally, they improved models by directly fine-tuning large-scale pre-trained multimodal models like VisualBERT[9] trained on COCO. The Hateful Memes Challenge Paper by Kiels et al.[7] presented the results for all three of the models mentioned above showing that multimodal models outperform unimodal models and performance of the best multimodal model was very far away from that of human performance.

In 2020, Facebook organized the Hateful Meme Challenge through Drivendata and NeurIPS-2020 to build models using a dataset of 10,000 images, which aided many researchers in developing novel solutions to the problem. The winning strategies and outcomes were outlined in the competition report[10] for the same. Zhu[1] came out on top with a performance that was nearly human-like, using additional labels for the dataset and creating an ensemble model with VL-BERT, UNITER-ITM, VILLA-ITM, and ERNIE-Vil. Niklas Muenninghoff[11] won second prize by ensembling different multimodal models. As can be seen, this challenge's solutions heavily drew from an ensemble of numerous large transformer-based models. Despite these models' high performance, their high computational cost raise problems in applying and deploying these solutions in real life. Therefore, it is necessary to look for models that are both highly performant and computationally inexpensive.

In paper[12], they used LSTM and decision tree rather than high-capacity transformers to build lightweight models and achieved comparable results to transformer baselines. In paper[13], they propose a novel multimodal learning method with image captions, objects, and memes and use a triple connection cross-modality network to achieve high performance. Instead of using a cross-modality network, our proposal is to build two transformer-based multimodal models using the meme's image captions, meme, and text and it will be less computationally expensive than the ensemble models from the leaderboard.

3 Methods

3.1 Dataset

For our experiment, we use Hateful Meme Classification Dataset provided by Facebook. There are 11,040 total memes in it, with 8,500, 540, and 2000 memes serving as the train, validation, and test sets, respectively. Each meme is labeled as 1 or 0 which corresponds to hateful or benign class respectively and to enable quicker processing, the dataset also includes meme text separately for each sample. The train set contains 36% hateful memes and 64% non-hateful memes, while the validation set and test set are balanced. There are five different types of memes in the dataset: unimodal hate memes, where the text or image is hateful on its own, multimodal hate memes, where only the combination of the memes makes it hateful, benign confounders made by modifying the text or image of multimodal hate memes to make them non-hateful, and some random non-hateful memes.

3.2 Model Architecture

Figure.2 shown below illustrates the architectural framework of our model where textual representation is generated from BERT using image caption and Multimodal representation is generated from VisualBERT using meme image and meme text. These two representations are concatenated together and sent to a Multi-layer perceptron for binary classification. Image captioning comes from the idea that if the meme is hateful, the caption generated from the image to describe it will be completely different from the original meme text; however, if the meme is not hateful, they might be somewhat similar. We use BERT because, unlike other pretrained models such as word2vec[14] or glove[15], it generates word embeddings based on context, and words can have multiple representations with BERT based on context, and for Multi-modal representation, we are using VisualBERT because it grounds the elements of image and language together without explicit supervision.

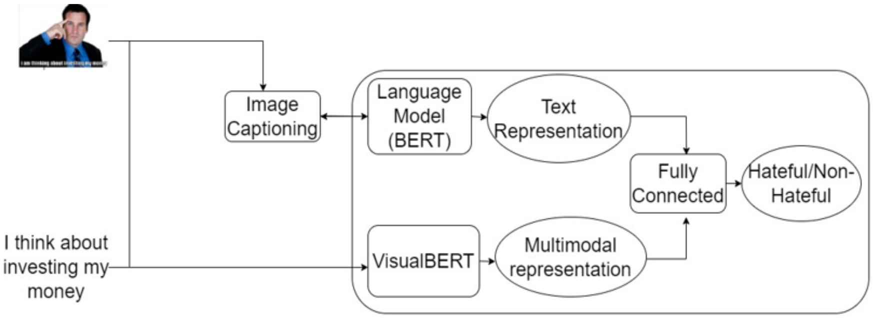


Fig. 2. Architectural Diagram of the model

First, all of the train images are resized to 224x224 and normalized during data preprocessing. Second, image captions are generated using a pre-trained image caption model which uses bottom-up and top-down attention approach[16] as it is the state-of-art model and this image captioning model is not trained as a part of the original model. To generate textual representation, we must fine-tune the BERT model on the caption corpus, for which we created captions for all of the meme images using the image captioning model and stored them. For fine-tuning BERT, we first used a tokenizer to obtain the maximum sentence length and built a vocabulary of words using the captions corpus. Then, we padded all the samples in it to the maximum length and used a tokenizer to encode them.

We load pre-trained BERT architecture which has gelu activation function at hidden layers and trained the model using the AdamW optimizer and NLLLoss for at least 20 epochs on the encoded caption corpus dataset. Similar to this, we finetuned the VisualBERT model with gelu activation by training it on the visual embeddings of the meme image and word embeddings of the meme text with attention mask using AdamW optimizer.

Once we have all the pre-trained models i.e. Image Captioning model, BERT, Visual BERT. The captions are sent to a pretrained BERT model to get a 768-dimensional textual representation of the caption. Parallelly, we also pass the meme text and meme image to the Visual BERT model to fetch another 768-dimensional multi-modal representation of the two modalities. Textual and multimodal representations are combined to create a final one-dimensional output vector with a size of 1536, which is then passed through two fully connected layers to obtain binary classification. This fully connected layer is trained final representations and their labels for over 30 epochs with a batch size of 64 using categorical cross-entropy loss and Adam optimizer. As a regularization technique, we used Dropout and Earlystopping with a patience of 5 monitoring the validation loss. We fine-tuned and trained the model on Google Colab which provides a free Tesla K80 GPU for a limited time per day to run DL applications and we also hosted Google Compute Engine VM with NVIDIA T4 GPU for fine-tuning the BERT and VisualBERT models as they are transformer models which take a lot of time to train.

Finally, after all the models are fine-tuned and trained. While testing, first we resize the images to 224x224, followed by generating an image caption for the meme. Next, generate textual and multimodal representations and concatenate them. Finally, send this vector to Fully connected network to get the prediction. We tested the model on all the test images and returned the results in terms of performance metrics i.e. AuROC and accuracy.

4 Experimental Design

For evaluating the model performance, we used two metrics as suggested in the challenge. The primary metric is the area under the receiver operating characteristic curve i.e. AUROC[17], which quantifies how well the classifier distinguishes between the classes as the decision threshold varies. Accuracy is used as a secondary metric as it is easily interpretable for a binary classification model. We also sampled some multimodal hate memes in order to evaluate how well the model performed when applied to the hate meme and its benign confounders. This enabled us in determining whether the model is classifying memes based on both text and image.

For every sample, a meme image is sent to an image captioning model which generates a caption describing the image. This caption is sent to the BERT language model to get textual representation. Parallelly, the meme image & meme text are also passed to Visio-linguistic model i.e. VisualBERT to get multimodal representation of the meme. Concatenating these two representations and sending it to our fully connected model will enable us to determine whether the meme is hateful or benign. The process is repeated over all the memes in the test dataset and the performance metrics i.e. accuracy, AUROC values are stored. Later, a comparison table is then created with existing state-of-the-art models from the hateful meme challenge and baseline model, VisualBERT, to see if the model has succeeded in its aim to design a simple model that would perform better than the non-ensemble-based model as it is capturing the context of the image as well.

5 Experimental Results

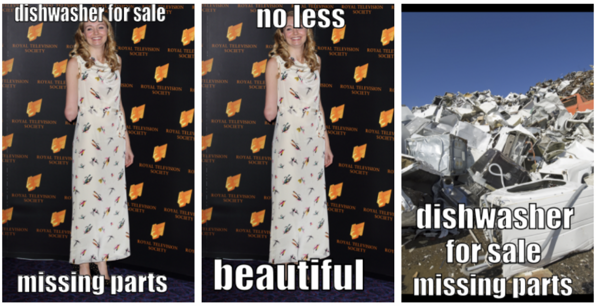
Table 1 compares the performance of our model with that of VisualBERT, the baseline model, and the best ensemble model from the Facebook hateful meme classification challenge. As we can see, when compared to the baseline model, the accuracy and AUROC score significantly increased after image captioning was incorporated into the model architecture. When compared to the baseline model, the model's accuracy has increased by 4.3% and the AUROC score has increased by 3%. This demonstrates that adding image captioning improves the results by providing a better representation of the image modality. Additionally, the performance of the current model is significantly lower than the performance of the best solution[1] to the Facebook challenge, but the current architecture is less complex than the ensemble models from the challenge, so we assume that our model is computationally more affordable than the best model.

Model	**AUROC**	**Accuracy**
Baseline Model - VisualBERT	71.41	64.73
Top performing ensemble model	84.5	73.2
Our Model	74.3	69

Table 1. Comparing performance metrics of our model

Testing the model's performance on a hate meme and its benign confounders is the best way to ensure that it works as intended. The figure5 depicts the same with three memes. The first image is the original hateful meme; the second is its text confounder, in which text has been replaced; and the third is an image confounder of the first meme, in which its image is replaced. Our model is put to the test on a few sets of hate memes and their benign confounders, and it successfully classifies the first meme to hate and its benign confounders as benign. This demonstrates how our model performs classification using both image and text modalities from the memes.

Analyzing the memes, we discovered some recurring patterns that the model had trouble identifying. One such pattern is that, when a meme contains multiple subimages as shown in figure5, the generated captions may be inaccurate

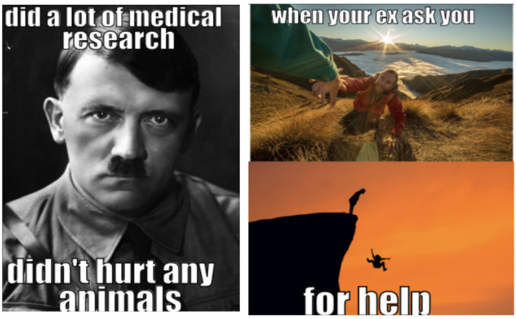


Correctly classified : **Hate**
Caption: actor at the premiere screening of film

Correctly classified : **Benign**
Caption: actor at the premiere screening of film

Correctly classified : **Benign**
Caption: a dumpster full of trash

Fig. 3. Performance of Model on Hate meme and its benign confounders



Caption: military commander
Actual Label: Hateful
Predicted Label: Hateful

Caption: digital art
Actual Label: Benign
Predicted Label: Hateful

Fig. 4. First meme demonstrates how captioning aided in correct prediction, while the second meme with subimages results in incorrect prediction

and cause incorrect predictions. Another finding is that, occasionally, even if a meme is not hateful, its text may not at all correspond to what the meme's accompanying image depicts. While as humans we might be able to connect, an algorithm with limited world knowledge will not be able to do so, leading to erroneous predictions.

As the image captioning model is not able to generate descriptions for memes with multiple subimages, in future, we must investigate methods to generate captions for these subimages separately and then combine them. A few approaches from the Facebook challenge used race, ethnicity, and gender labels on top of visio-linguistic models. Though this may help increase accuracy, it will be computationally expensive. As a result, we must seek ways to incorporate this information into the model without increasing its complexity. Apart from this, generally differentiating hate and benign memes is difficult for humans itself due to their complexity, and the dataset provided is also small, with only 12k+ images. As a result, the small dataset may have caused the model to be underfit.

6 Conclusions

For classifying a meme to hate or benign it is important to understand both its image and text. We added an image captioning model with BERT on top of base line model for better interpretation of the image and our model performed as expected i.e. better than the VisualBERT model. So, one key takeaway is that imagecaptioning helped understand the image better and thereby improved its accuracy. Finally, if the performance of this model is improved, it can be used to flag hate memes online and stop hate spread.

From an ethical standpoint, the definition of hate speech is quite arbitrary, and it is frequently challenging to distinguish between what is hateful and what is not. One which appears to be a hateful meme to one person may not be to another. Therefore, it's crucial to generalize the dataset to include data from all castes, races, genders, and conditions. However, gathering data necessitates the use of more hateful memes, which could imply the inclusion of images of members of these groups and thus offend some.

References

1. Zhu, R.: Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. arXiv preprint arXiv:2012.08290 (2020)
2. Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., Vakali, A.: Mean birds: Detecting aggression and bullying on twitter. In: Proceedings of the 2017 ACM on web science conference. (2017) 13–22
3. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive language detection in online user content. In: Proceedings of the 25th international conference on world wide web. (2016) 145–153
4. Cao, R., Lee, R.K.W., Hoang, T.A.: Deephate: Hate speech detection via multifaceted text representations. In: 12th ACM conference on web science. (2020) 11–20
5. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on World Wide Web companion. (2017) 759–760
6. Gomez, R., Gibert, J., Gomez, L., Karatzas, D.: Exploring hate speech detection in multimodal publications. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. (2020) 1470–1478
7. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems **33** (2020) 2611–2624
8. Suryawanshi, S., Chakravarthi, B.R., Arcan, M., Buitelaar, P.: Multimodal meme dataset (multioff) for identifying offensive content in image and text. In: Proceedings of the second workshop on trolling, aggression and cyberbullying. (2020) 32–41

9. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
10. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Fitzpatrick, C.A., Bull, P., Lipstein, G., Nelli, T., Zhu, R., et al.: The hateful memes challenge: competition report. In: NeurIPS 2020 Competition and Demonstration Track, PMLR (2021) 344–360
11. Muennighoff, N.: Vilio: state-of-the-art visio-linguistic models applied to hateful memes. arXiv preprint arXiv:2012.07788 (2020)
12. Deshpande, T., Mani, N.: An interpretable approach to hateful meme detection. In: Proceedings of the 2021 International Conference on Multimodal Interaction. (2021) 723–727
13. Zhou, Y., Chen, Z., Yang, H.: Multimodal learning for hateful memes detection. In: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE (2021) 1–6
14. Church, K.W.: Word2vec. Natural Language Engineering **23**(1) (2017) 155–162
15. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). (2014) 1532–1543
16. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, PMLR (2015) 2048–2057
17. Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern recognition **30**(7) (1997) 1145–1159