Hateful Memes Classification: Project Idea

Akshay Goindani (20171108) Preet Thakkar (20171068) Vaibhav Garg (20171005)
Sagar Joshi (2020701007)
Team 6

February 15, 2021

1 Problem Outline

Problem Statement: The goal is to predict whether a meme is hateful or non-hateful. This is essentially a binary classification problem with multimodal input data consisting of the the meme image itself (the image mode) and a string representing the text in the meme image (the text mode).

Input and Output: Given a rgb meme image, and a string representing the english text in the meme image, the trained model will output the probability that the meme is hateful.

Motivation:

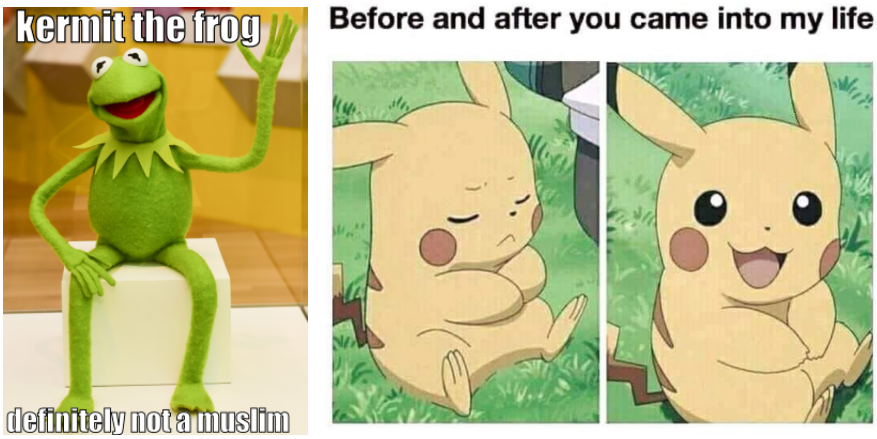


Figure 1: Both the images are fairly popular memes but clearly, one induces hate and negativity whereas the other one promotes love and positivity.

Internet memes have become a major form of communication and expression, they are an essential part of social media's popular culture. Modern-day social media platforms are full of memes as they are very easy to consume and also have the associated humor aspect. Figure 1 contains examples of such commonly shared memes. Moreover, these platforms have made them incredibly easy to share, two clicks are literally all it takes to make sure all your friends/followers have the same meme on their feeds. But like with any other type of content in social media, they also can lead to the spread of hate speech and propaganda. At the end of the day, these memes become an indirect source for our information consumption and play a significant role in shaping our belief systems.

The scale at which such content is produced makes it impossible for humans to go through all the memes and filter out the distasteful ones and hence there is an urgent need for fast and scalable automated solutions. Facebook AI's Hateful Memes Challenge [6] was one of the very first large-scale attempts to explore this problem and the results prove just how challenging this problem is, interpreting a meme requires a lot of cultural, political, and contextual knowledge and multimodal understanding since it is both the text and the image combined that brings out the real meaning of the meme. Another important thing to note here that even the human annotators themselves had a hard time classifying the memes and an average time of 27 minutes was spent per meme in the annotation process. And lastly, even though the results from previous research might look promising, but at the scale of the internet that still equates to a lot of hateful memes still being shared openly.

2 Problem Scope

The core objective of our Hateful Memes Classification Project is to come up with a classifier system to be able to spot memes bearing hateful content. The scope of this project is restricted to the following goals:

1. Exploring literature for existing work [1][3][9][11][12][13][16][18][19][20] in hateful meme classification, exploring quality datasets [6][14] that encourage experimentation for obtaining better results.
2. In implementation, try to replicate the performance of some of the existing state-of-the-art published methods, focusing on the use of multimodal transformer [15] architectures [2][7][8][10][17].
3. Approaches involving ensemble of classifier models [11][13][16] have produced some of the best results in hateful meme classification. So, we plan to design an ensemble approach after experimenting with some of the ensembling approaches in literature that have given good results for this problem.
4. Optionally, we will explore to see if existing machine translation techniques can be leveraged with our classification model to identify hateful content in languages other than English.

As described in section 5.3, we plan to use AUROC as our primary evaluation metric for performance of the classifier. As a part of the pre-processing required for the classification model, we intend to extract image features using a pre-trained Mask-RCNN [5] model, and generate text embeddings using BERT [4] as specified in section 5.1. In implementation, there might potentially be some changes at architectural level based on our on-the-ground observations and feasibility of implementation in stipulated time. We plan to stick to the above outline, and do not intend to cover techniques that can be experimented with ahead of the classification process, like identifying and blurring offensive content in memes or trying to convert a hateful meme to a non-hateful one. We also won't be focusing much on the scalability and efficiency aspect of our solution.

3 Expected Outcome

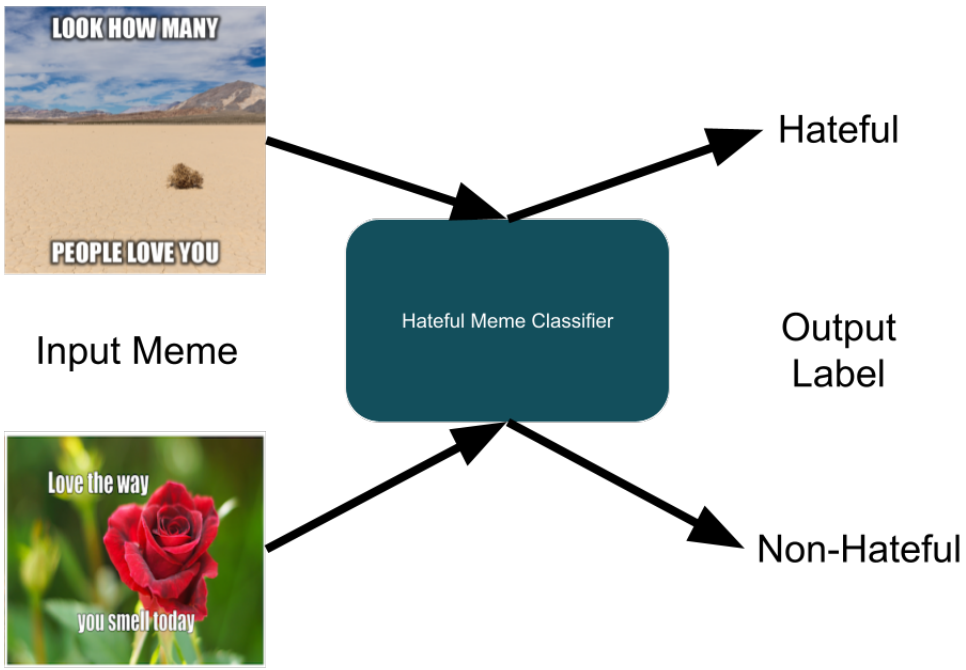


Figure 2: Input and Expected Output of Hateful Meme Classifier

Our goal for the next deliverable is to implement a multi-modal hateful meme classification model which takes the image and text in the meme as input and outputs a label which determines whether the input meme is hateful or non-hateful. For implementation we will follow the procedure and architecture mentioned in the Section 5. We

expect that our implemented model performs as good as the models proposed in [11] and [16]. After achieving this goal, the plan for the next deliverables is mentioned in the Secition 6. Figure 2 shows the expected outcome of our project on a given input.

4 Dataset Analysis

4.1 Facebook Hateful Memes Dataset

As part of the Hateful Memes Challenge [6], Facebook AI released a novel Hateful Memes Dataset consisting of $10k$ memes for the classification task to finetune on pre-trained multimodal architectures. The dataset is carefully constructed with the help of human annotators such that unimodal architectures focusing on just the image or text modality will fail if tried to be applied individually, so only multimodal architectures processing both the modalities together will work to give acceptable results. In the context of the dataset, the creators define hate as

"A direct or indirect attack on people based on characteristics, including ethnicity, race, nationality, immigration status, religion, caste, sex, gender identity, sexual orientation, and disability or disease."

The process of dataset creation can be summarized in the following steps:

- The objective was to construct new memes from the original memes already available on public social media groups, but so that they preserve the meaning and intent of the original meme. So, around $46k$ memes were filtered out from a source dataset of 1 million images and similar images to the meme image were found from Getty Images for constructing new memes.
- Memes were constructed using the new underlying image and text extracted from the original meme using an off-the-shelf OCR system.
- Human annotators were trained for 4 hours in recognizing hate speech, after which they were asked to annotate the memes on a scale of 1 (hateful) to 3 (non-hateful), which resulted in inter-annotator agreement (Cohen's kappa score) of 67.2%, implying moderate level of agreement. After moderation by expert annotators, binary classification labels were applied to the memes.
- To make the dataset challenging, benign confounders were added, i.e. memes with minimum replacement image or text component that can flip the classification label. An illustration of such benign confounders is shown in figure 3.

Figure 3: The original hateful memes are in the left column. Image confounders formed by replacing the image are in the middle column. Text confounders formed by minor replacements to the original text are in the right column.

Of the 10k memes that resulted out of the process, 8.5k memes are available as part of the training set and 0.5k and 1k respectively for the dev and test sets. They comprise of five different types of memes:

- Multimodal hate (40%) - Benign confounders exist for both the modalities.
- Unimodal hate (10%) - Either of image or text (or both) are hateful on their own.

- Benign image (20%) - Benign confounder exists for the image modality.
- Benign text (20%) - Benign confounder exists for the text modality.
- Random non-hateful memes (10%)

The creators have established baseline scores using several unimodal and multimodal classifiers, as shown in table 1.

Type	Model	Validation		Test	
		AUROC	Acc.	AUROC	Acc.
	Human	-	-	82.65	84.70
Unimodal	Image-Grid	58.79	52.73	52.63 ± 0.20	52.00 ± 1.04
	Image-Region	57.98	52.66	55.92 ± 1.18	52.13 ± 0.40
	Text-BERT	64.65	58.26	65.08 ± 0.87	59.20 ± 1.00
Multimodal (Unimodal Pretraining)	Late Fusion	65.97	61.53	64.75 ± 0.96	59.66 ± 0.64
	Concat BERT	65.25	58.60	65.79 ± 1.09	59.13 ± 0.78
	MMBT-Grid	68.57	58.20	67.92 ± 0.87	60.06 ± 0.97
	MMBT-Region	71.03	58.73	70.73 ± 0.66	60.23 ± 0.87
	ViLBERT	71.13	62.20	70.45 ± 1.16	62.30 ± 0.46
	Visual BERT	70.60	62.10	71.33 ± 1.10	63.20 ± 1.06
Multimodal (Multimodal Pretraining)	ViLBERT CC	70.07	61.40	70.03 ± 1.77	61.10 ± 1.56
	Visual BERT COCO	73.97	65.06	71.41 ± 0.46	64.73 ± 0.50

Table 1: Performance of state-of-the-art unimodal and multimodal classifiers on the Facebook Hateful Memes Dataset.

As observed from the benchmark scores and as is clear from the structure of the dataset, multimodal architectures beat the state-of-the-art unimodal architectures in Accuracy and AUROC scores on the meme classification task. Human accuracy on the dataset, however, stands at 84.40% with AUROC being 82.65%, both of which are far above the best multimodal model scores. This proves the challenging nature of the dataset and the gap to be filled in with better fine tuning on models, also possibly improvising with an ensemble approach.

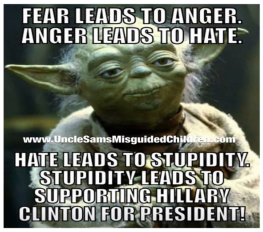
4.2 Multimodal (MultiOFF) Meme Dataset For Identification of Offensive Content

With the view to create a dataset for identifying offensive content in memes, a very specific event of US 2016 Presidential Elections was used to collect the corresponding memes. From an existing Kaggle dataset, the MultiOFF dataset [14] was created having images and preprocessed captions along with *offensive/non-offensive* as the classification labels. The annotations were performed using 8 annotators which had an agreement of Fleiss' kappa score in the range of 0.4-0.5, implying *moderate agreement*. With consideration of both the modalities, i.e. image and text, memes that could be perceived to be personal attack, homophobic/racial/minority abuse were labelled offensive. Figure 4 shows example memes in such categories.

The dataset contains a total of 743 samples with 445 training samples (187 offensive v/s 258 non-offensive) and 149 samples (59 offensive v/s 90 non-offensive) each in the validation and test sets. Several approaches were tested for the classification task, separately dealing with text-based classification, image-based classification, along with three multimodal approaches that use an early fusion of image and text embeddings.

Table 2 shows the precision, recall and F1-scores obtained on the classifiers used. Among all approaches, CNN on text was seen to have the best recall value of 84% and a decent precision score of 39%. A classifier having just VGG-16 for image data had the lowest recall score of 16%. Multimodal approaches were seen to have a fair precision v/s recall balance as compared to other models, but recall not reaching as high as CNN on the text. The best precision scores, however, were produced by basic Logistic Regression and Naive Bayes models dealing with just text data.

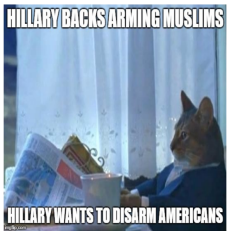
Considering the multimodal nature of data, and also substantial overlap between the objectives of the dataset with that of Facebook Hateful Memes, it can be worthwhile to see how state-of-the-art multimodal transformer architectures fined on Facebook dataset perform on the MultiOFF dataset. Since the training data available as a part of this dataset is just 445 samples, overfitting issues restrict the direct use of multimodal architectures on this dataset.



(a) Personal attack

(b) Homophobic abuse

(c) Racial abuse



(d) Minority abuse

Figure 4: Examples of memes with different types of offensive intents.

Text	Classifier	Precision	Recall	F1-score
Text	Logistic Regression	0.58	0.40	0.48
	Naive Bayes	0.52	0.45	0.49
	Deep Neural Network	0.47	0.54	0.50
	Stacked LSTM	0.39	0.42	0.40
	BiLSTM	0.42	0.23	0.30
	CNN	0.39	0.84	0.54
Image	VGG16	0.41	0.16	0.24
Multimodal	Stacked LSTM + VGG16	0.40	0.66	0.50
	BiLSTM + VGG16	0.40	0.44	0.41
	CNNText + VGG16	0.38	0.67	0.48

Table 2: Precision, recall and F1-score for unimodal and multimodal classifiers on MultiOFF dataset.

5 Surface Level View of our Approach

In our literature survey, we realized just how challenging this problem is due to the multimodal aspect of memes and the huge amount of contextual knowledge required to really understand and interpret a meme. The dataset mentioned above isn't large and diverse enough to really encompass all the world knowledge required for this task and hence taking advantage of large pre-trained language and vision models becomes a key in tackling this problem. Also, the dataset is made in such a way that unimodal approaches that use only the image features or the text features, will not be able to achieve good accuracy, therefore, using a multimodal approach is a very important part of this task. We found that [11] and [16] achieve AUROC score comparable to humans and therefore, we have decided to use key components from their proposed models. Previous research also showed that an *ensemble* of deep models radically impacted performance and therefore, we also intend to experiment with the newer ensembling techniques.

5.1 Feature Extraction

For extracting image features, we will use ResNeXT-152 based Mask-RCNN model [5]. For each image, this model gives 100 boxes of 2048 dimensional region based features. These extracted visual features are then projected into the textual embedding space. The text features are extracted using the BERT model [4].

5.2 Architecture

After getting the image and text features, the aim is to combine these features to get a multi-modal representation and use that representation for final classification. One approach to combine image and text features is to pass the image and text features through transformer layers where self-attention will discover implicit relation between image and text. The output of the transformer will be then passed through a fully connected layer followed by softmax which will output the probability of the input being hateful and non-hateful. This approach is similar to

[7]. Another approach is to use the image and text features and pass these features through an ensemble of different models such as ERNIE-VIL Large [17], ERNIE-VIL Small [17], UNITER [2], OSCAR [8] and VisualBERT [7]. The output features of these models are averaged. Simplex Optimization is used to make the final classification. This approach is similar to [11]. For the first approach mentioned above, we will also use ensembling. We will try different ensembling methods like majority voting, mixture of experts gating model etc. Based on the performance of these approaches we will select the method which gives better Accuracy and AUROC score.

5.3 Evaluation Metrics

The performance of our classification model will be tested using Area Under the Receiver Operating Characteristic curve (AUROC) metric, which was the one used by Facebook AI in the Hateful Memes Challenge. It is a measure of how much the model is capable of distinguishing between classes. An excellent model has AUROC near to 1 which means it has a good measure of separability. A poor model has AUROC near 0 which means it has the worst measure of separability. For additional information, we will also report the accuracy of our model which is simply the fraction of meme labels of the test set predicted correctly.

$$AUROC = \int_{\infty}^{-\infty} TPR(T)FPR^{'}(T)\,dT \tag{1}$$

$$Accuracy = \frac{1}{N}\sum_{i=0}^{N} I(y_i = \hat{y_i}) \tag{2}$$

AUROC score is also interpreted as the expectation that a uniformly drawn random positive is ranked before a uniformly drawn random negative. In the hateful memes challenge, we can say that intuitively AUROC score penalizes models that are bad at ordering memes by hatefulness. Whether the classification probability values themselves are high or low does not matter, only how they are ranked. The goal of our classification model will be to maximize AUROC.

6 Next Deliverables

After the next deliverable, we are planning to explore two directions:

- Improve the AUROC score on the Facebook hateful meme dataset [6] by enriching both the image and text features. We will also try to propose a novel approach for combining the image and text features in order to enhance the performance of the multi-modal architecture.
- Another direction is to create a hateful meme classification model for different languages for e.g., Hindi. Since, enough training data for Hindi is not available to train a model from scratch therefore, our idea is to use a translation module, which will first translate the text in the meme to English and then we will use our implementation of the hateful meme classification model for english, to make the final classification. In this approach we will first extract memes which have text in the foreign language (not English) and then we will do inference on those memes. We will not use the extracted memes for training.

References

[1] Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. A Multimodal Memes Classification: A Survey and Open Research Issues. *arXiv e-prints*, page arXiv:2009.08395, September 2020.

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[3] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. Detecting Hate Speech in Multi-modal Memes. *arXiv e-prints*, page arXiv:2012.14891, December 2020.

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

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[6] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

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

[8] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *arXiv e-prints*, page arXiv:2004.06165, April 2020.

[9] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A Multimodal Framework for the Detection of Hateful Memes. *arXiv e-prints*, page arXiv:2012.12871, December 2020.

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv e-prints*, page arXiv:1908.02265, August 2019.

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

[12] Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro-i-Nieto. Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation. *arXiv e-prints*, page arXiv:1910.02334, October 2019.

[13] Vlad Sandulescu. Detecting Hateful Memes Using a Multimodal Deep Ensemble. *arXiv e-prints*, page arXiv:2012.13235, December 2020.

[14] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 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, Marseille, France, May 2020. European Language Resources Association (ELRA).

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv e-prints*, page arXiv:1706.03762, June 2017.

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

[17] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

[18] Weibo Zhang, Guihua Liu, Zhuohua Li, and Fuqing Zhu. Hateful Memes Detection via Complementary Visual and Linguistic Networks. *arXiv e-prints*, page arXiv:2012.04977, December 2020.

[19] Xiayu Zhong. Classification of Multimodal Hate Speech – The Winning Solution of Hateful Memes Challenge. *arXiv e-prints*, page arXiv:2012.01002, December 2020.

[20] Yi Zhou and Zhenhao Chen. Multimodal Learning for Hateful Memes Detection. *arXiv e-prints*, page arXiv:2011.12870, November 2020.