Facebook Hateful Memes Challenge
First Assignment Report

Abhishek Das [*1] **Japsimar Singh Wahi** [*1] **Siyao Li** [*1]

1. Introduction

In today's world, social media platforms play a major role in influencing people's everyday life. Though having numerous benefits, it also has the capability of shaping public opinion and religious beliefs across the world. It can be used to attack people directly or indirectly based on race, caste, immigration status, religion, ethnicity, nationality, sex, gender identity, sexual orientation, and disability or disease. Hate Speech on online social media can trigger social polarization, hateful crimes. On large platforms such as Facebook and Twitter, it becomes practically impossible for a human to monitor the source and spreading of such malicious activities, thus it is the responsibility of the machine learning and artificial intelligence research community to address and solve this problem of detecting hate speech efficiently.

In tasks such as VQA and multimodal machine translation, it has been observed that baseline models using language domain perform really well without even exploiting the multimodal understanding and reasoning[15]. However, the Facebook Hateful Memes Challenge Dataset is designed in such a manner that unimodal models exploiting just the language or vision domain separately will fail, and only the models that are able to learn the true multimodal reasoning and understanding will be able to perform well.

They achieve this by the introduction of "Benign confounders" in the dataset, i.e. for every hateful meme, they find alternative image or caption which when replaced, is enough to make the meme harmless or non-hateful, thus flipping the label. Consider a sentence like "dishwasher for sale, missing parts". Unimodally, this sentence is harmless, but when combined with an equally harmless image of a girl without a hand, suddenly it becomes mean. See Figure 2 for an illustration. Thus, this challenge set is an excellent stage which aims to facilitate the development of robust multimodal models, and at the same time addresses an important real-world problem of detecting hateful speech on online social media platforms.

[1]Carnegie Mellon University, Pittsburgh, United States.

2. Experimental Setup

The objective of this challenge is that given an image and pre-extracted text, we need to classify the memes as hateful or benign. Thus this is essentially a binary classification problem, which requires a subtle reasoning to interpret the message and cues within the combination of image and caption used.

2.1. Dataset

We will use the Facebook Memes Challenge Dataset (Kiela et al., 2020) which comprises 10k memes. These memes are carefully designed for this task by annotators who are specially trained to employ hate-speech as defined by Facebook. The features in this dataset are the meme images themselves and string representations of the text in the image. The dataset comprises five different types of memes as shown in Figure 1: multimodal hate, where benign confounders were found for both modalities, unimodal hate where one or both modalities were already hateful on their own, benign image and benign text confounders and finally random not-hateful examples. The Training, Validation and Test split is 85, 5 and 10 respectively, and the individual sets are fully balanced.

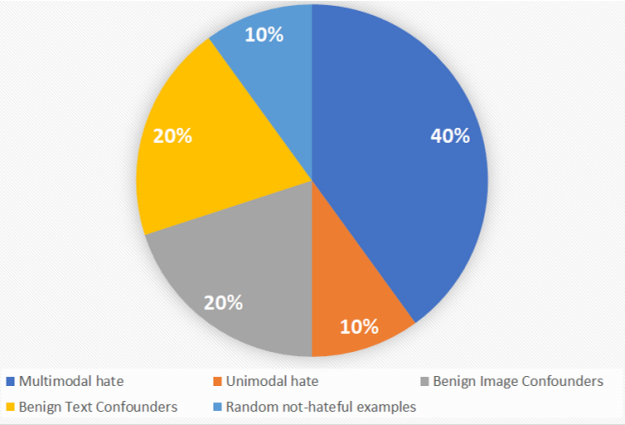

Figure 1. Types of memes in the Facebook Hateful Memes Challenge Dataset

Figure 2. Multimodal "mean" meme and Benign confounders. Mean meme (left), Benign image confounder (middle) and Benign text confounder (right).

2.2. Evaluation Metrics

We plan to evaluate the performance of our model using two metrics as suggested in the challenge

2.2.1. Area under the Receiver Operating Characteristics (AUC ROC)

The metric measures how well the binary classifier discriminates between the classes as its decision threshold is varied.(Bradley, 1997)

2.2.2. Classification Accuracy

We will find the accuracy of the predictions which is given by the ratio of correct predictions to the total number of predictions made, since it is easier to interpret.

3. Related Work

Hateful speech detection gains more and more attentions in recent years. Several text-only hate speech datasets have been released, mostly based on Twitter [(Waseem, 2016),(Waseem & Hovy, 2016),(Davidson et al., 2017)], and various architectures have been proposed for classifiers [(Kumar et al., 2018), (Malmasi & Zampieri, 2017)]. Some of the biases around hate speech have been prevalent like [(Dixon et al., 2018), (Sap et al., 2019)].

In the past few years there has been a surge in multimodal tasks and problems, ranging from visual question answering to image captioning and beyond. However, there has been surprisingly little work related to multimodal hate speech, with only a few papers including both images and text.

Gomez et al.(Gomez et al., 2020) highlighted the issue that most of the previous work on hate speech is done using textual data only and that hate-speech detection on Multimodal publications has not been addressed yet. So they have created and made available MMHS150k, a manually annotated multimodal hate speech dataset formed by 150,000 tweets, each one of them containing text and an image. The data points are labelled into one of the six categories: No attacks to any community, racist, sexist, homophobic, religion based attacks or attacks to other communities.They train a simple LSTM model which considers just the tweets text as a baseline for the task of detecting hate speech in multimodal publications. Their further objective is to exploit the information in the visual domain to outperform their baseline. They do this by proposing two models. The first one is the Feature Concatenation Model (FCM), which is a MLP that concatenates the image representation extracted by the CNN and the textual features of both the tweet text and the image text extracted by the LSTM.Their second model named Textual Kernels Model(TKM)is inspired by VQA tasks and is based on the intuition of looking for patterns

in the image corresponding to the associated texts. This is done by learning kernels from textual representations and convolving them with CNN feature maps.

Yang et al. [(Yang et al., 2019)] report that augmenting text with image embedding information immediately boosts performance in hate speech detection. In this paper, the image embeddings are formed by using the second last layer of the pre-trained ResNet neural network on ImageNet and then hashing these values for efficient photo indexing, searching and clustering. The hash takes advantage of the deep pre-trained image network which offers discriminative semantic representations. It preserves the similarity between original photos: the photos with smaller Hamming distance between their hashes look similar to each other. For each word in a piece of text, we retrieve the pre-trained embeddings, These embeddings are fixed during the model training, then apply a word-level MLP on each of the word embeddings, creating the new word embeddings. Next, they apply a 1D-convolution to the words. With proper padding, we ensure that the output of the convolution matches the length of the input for different Ngram-window sizes. This offers the convenience for executing attention operations. They then apply max-pooling and tanh to create a fixed-size vector representation for the piece of text. The most straightforward way of integrating text with photo features is to concatenate both image and text vectors. The concatenated vector is followed by dropout, MLP and softmax operations for the final hate speech classification. They also explore other fusion techniques like gated summation and bilinear transformation. They use attention mechanisms to further improve their results. The context vector is the 1D-convolution output from text, while the query vector is the photo vector. They use ROC-AUC as the performance metric in this paper, which measures the classifier's performance across all scoring points. Fusion using attention mechanism turn out to work pretty well.

Early work suffers from the limited size of the multimodal dataset collected. Singh et al. (Singh et al., 2017) tends to alleviate this issue by avoiding very deep network architectures. They use the dataset created by Hosseinmardi [(Hosseinmardi et al., 2015)]. After modification, the dataset is very small with 461 samples for training and 238 for testing. They use feature selection and the Bagging (Bootstrapping Aggregation) algorithm for classification. The text features are collected by using Linguistic Inquiry and Word Count (LIWC). They use Microsoft's Project Oxford to extract visual features. Example features that are useful for classification are 1) Visual: people in a performance, text signs present, outdoor scenes, 2) Textual: comparisons made, sadness, certainty, health related, sexual, informal non-fluencies. Their results show that using both text features and image features is better than only use unimodal features.

4. Research Ideas

When we explored the data, we realised that we should explore on getting the important information out of both the modalities and find the relationship between each important word or object in these modalities. Therefore, we plan to introduce Intermediate Representations. We can use these intermediate blocks to extract useful information out of each modality. We can generate this representation by performing various tasks like doing Sentiment Analysis on the image and the text modality. During our analysis on the data, we found that many images having a positive effect gets classified as hateful by just one negative keyword in the text or a positive text gets converted into hateful by presence of one negative object in the image. Sentiment analysis might help us in finding these positive and negative sentiments inside each modality and might help in improving the results. The idea here is to find the relationship between the objects present in each modality separately and learning the context behind it.

Then, we can concatenate the two modalities or use other fusion techniques with attention to get better results for binary classification. We can also explore the fusion techniques like concatenation, gated summation with early, moderate or late fusion and check which gives us the best results.

Some other ideas which we plan to work on includes using Bilinear Attention networks, which is very much used in VQA models such that we can extend and fine-tune it for our use case. We also found that majority of the text data set is divided in the top and bottom part of the image. We can explore on this to get a relationship between the two text embedings.

Pre-training models on image and texts on large-scale dataset of hate-speech like MMHS150K (Gomez et al., 2020) created from Twitter and other such platforms might help us in expanding the data we can work on and improving the results for our performance.

5. Unimodal Analysis

In this section, we will discuss our initial exploration of the individual modalities involved in the Hateful memes detection task. In specific, we use Glove (Pennington et al., 2014) and BERT (Devlin et al., 2018) to inspect the text component of the memes in their embedding space, and analyze the visual representations generated by ResNet (He et al., 2016).

5.1. Visual Modality

For visual modality, we look at the pre-trained image representations based on the ResNet (He et al., 2016). In specific, we encode the image with standard ResNet-152 convolu-

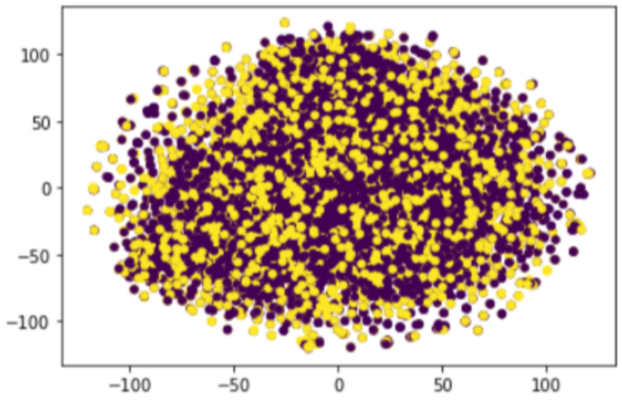

Figure 3. T-SNE results on the Visual modality.

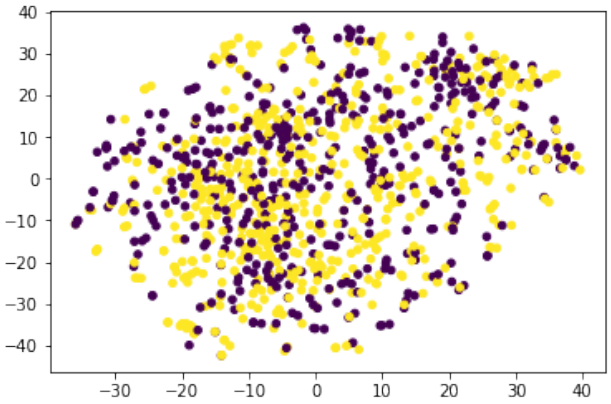

Figure 4. T-SNE results on the word-level representations.

tional features with average pooling and visualize the second to the last layer with T-SNE (Maaten & Hinton, 2008) algorithm, which is a technique for dimensionality reduction. The results are shown in Figure 3.

From figure 3, we can see that it is hard to separate the positive and negative labels. Some images are even labeled as both hateful and non-hateful. It is mainly because the dataset is constructed in a way to confound predictions from single modality. We run a classification task using the pre-trained representations. The test accuracy is 52.63 percent, which is not significantly higher than random. It is hard to get any useful information by simply looking at the visual component of the dataset.

5.2. Language Modality

We explore both word-level representation and contextual representations for our language data. For the word-level analysis, we use Glove (Pennington et al., 2014), which contains 6B tokens and is pre-trained on Wikipedia2014 and Gigaword5, to convert all the English word tokens into word embeddings with 50 dimensions. For simplicity, we do not include any out-of-vocabulary words. Visualizations are done by using the T-SNE (Maaten & Hinton, 2008). Since word tokens themselves are not directly related to the classification labels, we then denote the word tokens, which have higher probability to appear in a hateful meme than average, as hateful. By doing so, we plot 1000 the word embeddings with respect to their labels in Figure 4. In addition, we also inspect the representation at the sentence level by using BERT (Devlin et al., 2018). We use the bert-uncased-base version to generate our contextual embeddings and use T-SNE (Maaten & Hinton, 2008) again to map the representations to 2D plane. The results are shown in Figure 5. Similar to the visual representations, the positive and negative labels are still highly mixed. It is again due to the dataset itself. However, the labels are less messy in the

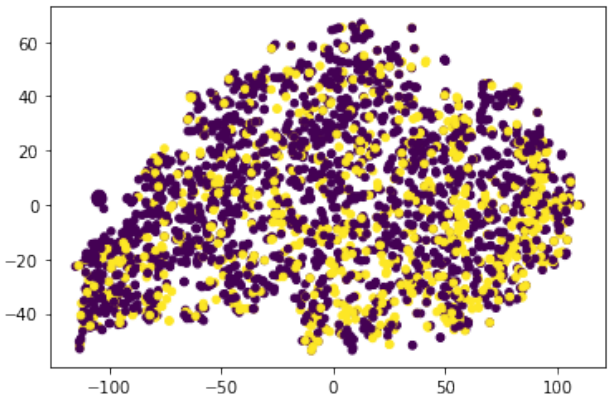

Figure 5. T-SNE results on the sentence-level representations.

text visualization. As we examine some of the tokens that are close to each other, we found that sometimes, tokens with the similar meanings tend to share the similar labels. For example, tokens related to countries are more likely to be presented in hateful memes, while tokens related to money tend to appear in non-hateful memes. The unimodal prediction results also align with our observation. The prediction based on BERT embeddings results in 59.20 percent accuracy, which is about 13 percent higher than those based on images. It shows that the text is a better indicator for the hateful meme detection, but the performance is still undesirable. As we experimented with multimodal representations (ViLBERT (Lu et al., 2019)), we observed another 5 percent increase in prediction accuracy; however, the results are still far away from human accuracy, which leads to a huge space for improvement.

Figure 6. False negative prediction by text only model

5.3. Error Cases

Some error cases can be draw from predictions based on text only. For examples, Figure 6 is actually hateful meme, but our unimodal model cannot predict correctly without accessing the image information. Similar cases are also found in image only prediction results. We can see that with only single modality, it is impossible to detect the hateful memes successfully.

5.4. Experimental Details

As we mentioned before, we carried out two unimodal baseline experiments and one multimodal basline experiment on the Facebook data set. For visual component, ResNet is used. Grid search was performed on batch size, learning rate and maximum number of updates to find the best hyperparameter configuration. 22000 iterations provided the best results for the same. We then did a detailed analysis on the results obtained by running these models and its predictions. The models were evaluated at an interval of 500 updates on the dev set and the model with best AUROC was taken as the final model to be evaluated on test set. Weighted Adam with cosine learning rate schedule and fixed 2000 warm up steps was used for optimization without gradient clipping. For text only and multimodal model, we used pretrained weights provided by Facebook Hateful Meme Challenge. The model used for text is BERT, while ViBERT pretrained on Conceptual Captions is used for multimodal prediction. Figure 7 shows results of other baseline models presented by the Facebook Hateful Meme Challenge.

References

Bradley, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, 1997. doi: 10.1016/S0031-3203(96)00142-2. URL https://eprints.qut.edu.au/114256/.

Davidson, T., Warmsley, D., Macy, M. W., and Weber, I. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009, 2017. URL http://arxiv.org/abs/1703.04009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. Measuring and mitigating unintended bias in text classification. 2018.

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. Detection of cyberbullying incidents on the instagram social network. *CoRR*, abs/1503.03909, 2015. URL http://arxiv.org/abs/1503.03909.

Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., and Testuggine, D. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pp. 1–11, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-4401.

Lu, J., Batra, D., Parikh, D., and Lee, S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

Malmasi, S. and Zampieri, M. Detecting hate speech in social media. In *Proceedings of the International*

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

Figure 7. Baseline Model Performance.

Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 467–472, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_062. URL https://doi.org/10.26615/978-954-452-049-6_062.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Sap, M., Card, D., Gabriel, S., Choi, Y., and Smith, N. A. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://www.aclweb.org/anthology/P19-1163.

Singh, V. K., Ghosh, S., and Jose, C. Toward multimodal cyberbullying detection. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 2090–2099, 2017.

Waseem, Z. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL https://www.aclweb.org/anthology/W16-5618.

Waseem, Z. and Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pp. 88–93, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-2013. URL https://www.aclweb.org/anthology/N16-2013.

Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E., and Predovic, G. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the Third Workshop on Abusive Language Online*, pp. 11–18, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3502. URL https://www.aclweb.org/anthology/W19-3502.