Multimodal Hateful Memes Detection

Yanqin Tan, Yumeng Shen, Yile Su
Georgia Institute of Technology
ytan312, yshen345, ysu334@gatech.edu

Abstract

Hateful contents that hurt a benign cyber public space sometimes comes in as multimodal memes that involve both image and text. Recent advances in multimodal machine learning provide capable models to detect hateful contents across multiple modalities. Building upon Facebook's baseline models in the Hateful Memes Challenge, this project explored meaningful methods in deep learning including data augmentation, semi-supervised self-training and model architecture to improve the performance of baseline models.

1. Introduction

1.1 Motivation and Objective Nowadays, as the number of social media users increases and the format of online dialogue diversifies, hateful speeches that combine visual and linguistic contents sneak into different social medias, bypassing existing unimodal detectors that focus exclusively on texts or images. To cope with this new challenge, the Facebook AI team developed a multi-modal hateful memes dataset and designed "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes" to encourage exploration in this field in 2020. This project is an effort to improve the performance of hate speech detection in multimodal hateful memes.

Multimodual machine learning has been a rapidly emerging field since the 2000s. Mutimodual, meaning multiple modalities, is one of the features of our natural world, where text and image, audio and video are always fused together. Current multimodal applications includes audio-visual speech recognition, multimedia content indexing and retrieval, media description (such as image captioning and description, image question answering). The challenges in this field usually includes representation (e.g. representing audio and video as symbolic), translation (i.e. mapping data from one modality to another like image description), alignment (i.e. identifying relationship between elements from different modalities), fusion (i.e. order of joining different modalities) and co-learning across different modalities [1]. The Hateful Memes, which combine images and texts to convey hateful contents are tricky in a sense that the combination of image and text conveys hates while the image and the text alone is innocent. Therefore, multimodal hateful speech detection faces challenges in translating/captioning the image contents, learning the relationship between subelements in texts and images, as well as fusion between the two modalities.

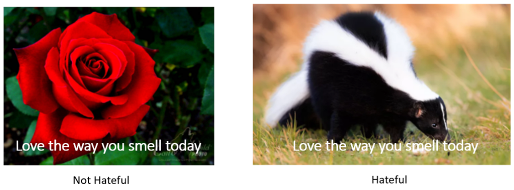


Figure 1: Multimodal Hateful Memes Examples

The Hateful Memes Challenge designed by Facebook AI research team is hateful in a sense that the textual and image contents might not align to express hate. Figure.1 is an example of this type of hateful memes across two modalities. Our goal in this project is to improve the baseline models' performance in multimodal hateful speech detection using deep learning frameworks including data augmentation and semi-supervised self-learning, with an ultimate goal to leverage machine learning algorithms in creating a cleaner, beginner and healthier cyber space that is friendly to people of all ages, genders, ethnicities and individualities in different social media.

1.2 Dataset The major dataset used in this project is the Hateful Memes dataset provided by Facebook AI (https://www.drivendata.org/competitions/64/hateful-memes/page/206). It contains 5 separate sub-datasets, one training set, two development sets (seen and unseen) and two testing sets (seen and unseen). Within the training set there is total 8500 memes. The development set contains 5000 memes, the unseen test set emcompasses 1000 memes. The dev and test set are fully balanced, they are comprised of memes using the following percentages: 40% multimodal hate, 10% unimodal hate, 20% benign text confounder, 20% benign image confounder, 10% random non-hateful. All the memes are labelled ready to use regarding hateful or not hateful by human annotators. Additional datasets used for experiments will be introduced in the Approach section.

1.3 Current Practice and Baseline Model Correctly classifying a meme in the dataset require reasoning and understanding about subelements and their relationship from multimodalities. Regardless of the challenge, The Hateful Memes Challenge paper [1] provided a few base lines which entails three types of models: unimodal pertained on vision or text models, multimodal pertained models with various fusion techniques such as concatenation of image and text features followed by MLP classifier (Concat BERT), taking the mean of the unimodal ResNet-152 and BERT output scores (Late Fusion), and more sophisticated fusion methods such as supervised multimodal bitransformers[5] (MMBT-Grid and MMBT-Region), and multimodal pretrained models. According to baseline results, text-only classifier performed slightly better than vision-only classifier and multimodal models performed better than unimodal models. Additionally, advanced fusion algorithm helps with performance, in addition to the advantage of early fusion model in comparison to middle and late fusion approaches. The baseline models entail two image encoders, 1) the standard ResNet-152[4] convolutional features from res-5c with average pooling(Image-Grid) 2) features from fc6 layer of Faster-RCNN[7] with ResNeXt-152 as its backbone[11]. And for the text encoder, BERT[3] is leveraged. The baseline models leverage transfer learning from pre-trained text and image models on multimodal tasks. The best baseline model achieves the AUROC of 0.7544 and Accuracy of 69.47% which is still far behind the reported human accuracy of 84.7% which leaves much room for improvement.

In this project, we used the MMBT model as a baseline, with most default parameters: batch_size = 32, lr = 1e-5, num_warmup_steps: 2000, loss function = Cross Entropy Loss, running a total of 5000 iterations with criteria of AUROC for early stopping. MMBT is chosen as the baseline model because that multimodal memes post big challenge in fusion, which learns the alignment or relationship between image and text. For these memes, simple concatination of pretrained text encoder and learned image features followed by classifiers such as MLP only gives moderate accuracy. Similarly, late fusion after text and image are pre-trained respectively doesn't provide state-of-art performance either. On the contrary, more sophisticated fusion such as Supervised Multimodel Bitransformers (MMBT) [5], which projects image embeddings to text token space and jointly fine-tunes unimodally pretrained text and image encoders achieves better accuracy. We believe that improvement based on this model has more merits in future applications.

1.4 Current Limit and Plan The challenge of hateful memes detection among the published baselines is that the classification knowledge needs to be transfer-learned from models pretrained on another dataset. The pretrained models might still experience difficulty to align with the specific task of the interest even if they are task-agnostic and learn the general features. This at some level explains the current baselines still have observable gaps from average human detection performance. It makes the transfer learning specific task figuring out even more challenging that there are only 8,500 memes (image text sets) in the training session, where 3,019 are labeled hateful, much smaller size than the pretrained source dataset. Therefore, our project focuses on leveraging modern deep learning techniques including data augmentation, semi-supervised self-training, hyper-parameter tuning and integration techniques to explore the room for improvements.

In addition, all base line models are using BERT text encoders, thus more advanced text encoders such as Roberta, the robustly optimized method for pretraining natural language processing (NLP) systems can be leveraged with the goal to boost the performance.

2. Approach

2.1 Data Augmentation

2.1.1 Memotion 7K Dataset

Our first attempt is to select additional hateful memes from other dataset to improve the learning of useful features. Memotion 7K Dataset (https://www.kaggle.com/williamscott701/memotion-dataset-7k) is originally used to predict memes sensitivities including humor, sarcasm, offense and motivation degree (see Appendix.8). We took a rigorous manual review to assess the labelling correctness and suitability for the hateful meme situation. One desirable aspect is that there are varied combinations of celebrities in the Memotion 7K dataset. For instance in Figure.2, though each sharing similar face expressions, they convey quite different semantics through encouraged, sarcastic and hateful. This might potentially boost the fine-tuning's capacity to differentiate the fine liner under the similar image features.

Mainly referring to the "offense" and "overall sentiment" dimensions, we collected memes labeled as "hateful offensive" or labeled "very negative" and added a total of 343 hateful memes to the training dataset of the hateful memes challenge.

2.1.2 Image Noise

Referring to fooling network interpretation in the image classification to bring system generality and robustness [9], we found it interesting to see if constructing a similar adversarial meme could fool the MMBT baseline. Below we took a hateful example Figure.3 which was initially classi-

NIGHTMARES
Are made of this.
Mr. Bean taught me one thing in life.
Enjoy your own company instead of expecting someone else to make you happy.
@DCMARVEL_MEMES
TAG A FRIEND SO THEY OPEN THEIR PHONE TO SEE THIS
THAT FACE YOUR FRIEND GIVES YOU
WHEN YOUR CRUSH WALKS INTO THE ROOM.

Figure 2: Memotion 7K Celebrity Meme Examples

fied correctly. We separately added Gaussian noise to this meme and went through the same network to find that it is possible to disturb the target prediction. As such, we decided to add Gaussian noisy memes to the training sample and test if would contribute to the model's accuracy and generalization.

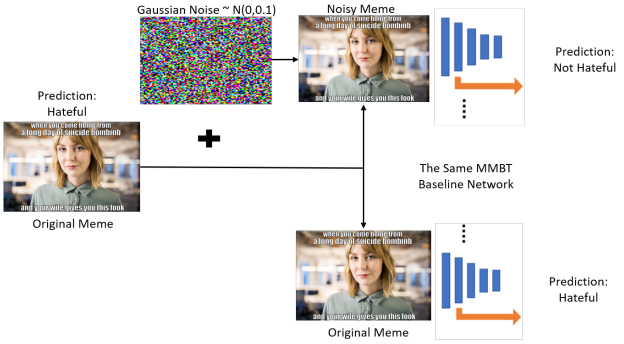

Gaussian Noise ~ N(0,0.1)
Noisy Meme
when you come home from a long day of suicide bombinb
and your wife gives you this look
Prediction: Not Hateful
Prediction: Hateful
Original Meme
The Same MMBT Baseline Network
Original Meme
Prediction: Hateful

Figure 3: Gaussian Noise Disturbance Example

Among the 8500 training memes, we randomly selected 4250 memes and added a Gaussian noise N(0,0.1), to these memes' corresponding images. A wide range of Gaussian variance was tested from 0.1 to 1.0 and we considered that 0.1 a best suit for the given memes challenge, where it did not twist the human visual perception of the image.

2.1.3 Text Replacement

From time to time reviewing the memes online, we may notice an occasion where the text is not written in a formal language. For example, the majority of the meme text is expressed in English yet several characters are from another language or even symbols like trademarks. However, these "special" characters still share much similar looks with those in English (Figure.4), namely it dose not influence human per understanding the semantic on English knowledge. While for machine training, different characters might make the network call for the respective language/symbol features and fail to leverage the overall semantic. Therefore, we decided to add the text "noise" to the training set and expect the network to gain more robustness, and learn more on how to use the major semantic of the image to come up with the replaced/twisted portion in the text.

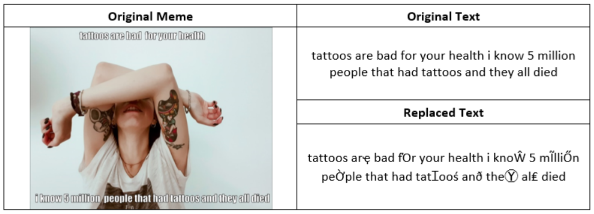

Original Meme
Original Text
tattoos are bad for your health i know 5 million people that had tattoos and they all died
Replaced Text
tattoos arę bad fOr your health i knoŴ 5 mĨlliŐn peỜple that had tatⵊooś anð theⓎ al₤ died

Figure 4: Text Replacement Example

Specifically, we randomly selected 4250 memes from the original 8500 training datasets and utilized the Facebook Augly tool to randomly replaced 80% of the text characters by similar characters in another "language" system. Then we added the newly created memes combinations to the training session to perform MMBT model training.

2.2 Semi-Supervised Self-Training

Besides the challenges about representation, translation, alignment and fusion in most multimodal tasks, one challenge that is unique in this task is its insufficiency in data size and data variety. Since the data needs to be collected, filtered and labeled manually, the cost of getting more labeled data is huge. In addition, since the relation between text and image is complex due to the nature of a meme itself, the dataset is far from being balanced regardless of class balance, which therefore results in lack of variety and robustness for real online memes classification. Here is where semi-supervised learning and self-training aids in.

Semi-supervised training is an approach to reduce the expanse of data collecting and labeling by supplementing additional unlabelled data. Recent advances show that deep self-training, as one type of semi-supervised training, can sometimes outperform pretraining significantly [8]. This method usually involves an iterative process of prediction and retraining. To be specific, in this project, we used the MMBT model that is pretrained on Hateful Memes Dataset to make predictions on the Reddit Memes Dataset (https://www.kaggle.com/sayangoswami/reddit-memes-dataset), which collected 3300+ latest memes from Reddit.com. The predictions are then used as pseudo-labels and were added to the training dataset together with the image data for a retraining. We believed that the strength of self-training resides in its capability to include extra data that share important relevant features with the target dataset and task while pretraining can sometimes introduce irrelevant information and bias [12]. Different from other memes datasets that have clear textual contents, the Reddit Memes Dataset contains memes that show a more complicated relationship between text and images, including sarcasm, irony and other implicit sentiments, which might contribute to the hateful memes detection if included with pseudo-labels.

However, one of the challenges in this approach is the proper use of confident prediction and the balance of class. As we anticipated, only of the Reddit Memes Dataset is predicted as harmful, which results in a class imbalance in auxiliary data to be added. Therefore, we added a label regularization term by setting a threshold for confident prediction: only non-hateful labels with 0.90+ confidence and hateful labels with 0.50+ confidence are used.

2.3 Integration and Hyperparameters Tuning

The selected 343 hateful memes from memotion 7K dataset, and the gaussian noise added 4250 memes and the text replaced noise added 4250 memes as well as the 3259 memes with labels generated from self-training pseudo labeling are included to the training set for an integrated training. We trained our final model based on the augmented and more pseudo label generated train set and validate on the dev set and evaluate the on the unseen test set. We report the area under the receiver operating characteristic curve(ROC AUC)[2] as the main metric, and also report the accuracy as the dev and test set is balanced and it's easier to interpret.

After integrating all enhancement, we aim to fine tune the hyperparameters, especially the learning rate, batch size, number of epochs, number of warmup steps, the optimizer and the learning rate scheduler on the extended training dataset to see if the enhancements add up or not.

2.4 Roberta Text Encoder and Supervised Multimodel Bitransformers Fusion

We planned to use RoBERTa[6] leaned text representation co to replace the original BERT text representation in combination with Supervised Multimodal Bitransformers Fusion (MMBT). As RoBERTa improves the undertrained BERT by training with bigger batches over more data, removing the next sentence prediction of BERT and dynamically changing the masking pattern applied to the training data, the improved training procedure of Roberta allows it to learn presentations to generalize even better to downstream tasks compared to BERT. Therefore, we attempted to use the enhanced text encoder RoBERTa to replace the text encoder of MMBT baseline model, with a hope to leverage the representation pre-trained by RoBERTa to gain model performance. However, during implementation, we encountered significant errors in the MMF framework, which prevented us from further exploration in this track.

3. Experiments and Results

3.1 Data Augmentation

Leveraging the three-version augmented dataset stated as above, we performed the training of the baseline model both on original training set and augmentations to collect performance variations. To be noted, the baseline model we selected in the experiment is MMBT provided by the Facebook MMF framework's "model zoo" (https://mmf.sh/docs/notes/\OT1\textquoterightmodel_zoo). The parameters of the MMBT baseline are all by default setting except that maximum iterations are set to 5000 to make the training time manageable for this course project.

The following Table.1 summarizes the performance comparison between MMBT baseline and three data augmentation versions. We also included the average loss curve comparison Figure.5 per iteration during the training process. As indicated in the testing set, there is barely improvement on the ROC metric, 0.10 point, per the Memion 7K version. Though during the first few iterations, there is a notable decrease on the loss level, given the added target samples, the two loss curves are quite converging approaching the end of iterations. This may be due to that the added samples number is still much smaller versus the baseline training data size to create obvious impact. Another potential reason is that as the MMBT is not fully pretrained on the celebrity faces, it would pick up slowly on the celebrity faces based hateful memes added from Memotion 7K.

For the image noise version, there is overall notable marginal benefit on both the validation and testing performance, 1.0 on testing accuracy and 2.0 on ROC metrics. During the initial training steps, it is shown that the augmentation version has a higher average loss level due to the noise disturbance. However, the loss curve is converging fast to the same level of the baseline and finally slightly underneath the baseline. Meanwhile, the benefit impact is not as sound as expected. After reviewing the unseen test set, we considered that the performance may be resulted from the testing images well-constructed in quality. Thus, there is limited space for the model to represent if it has learned to sustain the noise attacks.

As for the text replacement version, we were able to get the model performance as follows. It is indicated that there is a significant boost on both the validation and testing performance with around 6.00 metric point increase in testing ROC and 4.00 increase in validation ROC.This may be attributed that the twisted memes forced the model system to enhance the capacity to make use of the intersection part between image and text. During the training process, the average loss curve of the augmentation version is almost higher everywhere than the baseline. While leveraging the test outperformance, we could notice the text "noise" bring more generality and less overfitting on the existing data source.

Table 1: Performance - Baseline vs. Data Augmentation

Setup	Validation		Test	
	Acc.	AUROC	Acc.	AUROC
MMBT	61.30	61.69	63.80	62.33
MMBT_Add Memotion 7K	63.52	59.40	63.75	62.40
MMBT_Add Image Noise	62.78	60.37	64.85	64.22
MMBT_Add Text Replacement	65.00	65.83	65.90	68.19

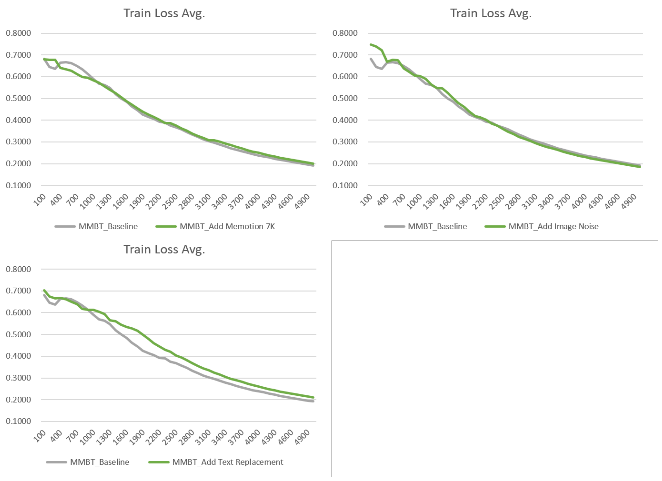


Figure 5: Loss Performance - Baseline vs. Data Augmentation

3.2 Semi-Supervised Self-Training

In the experiment of self-training, we tried two different settings based on prediction confidence. One of the problems we anticipated is that self-training might introduce noises because of the use of pseudo-labels, which are not generated by professional human annotators. Therefore, we applied a regularization term by filtering out non-hateful labels with 0.9 confidence and below as well as hateful labels with 0.5 confidence and below, in addition to the inclusive use of all pseudo-labels generated by the model.

Table 2: Performance - Baseline vs. Self-Training

Setup	Validation		Test	
	Acc.	AUROC	Acc.	AUROC
MMBT	61.30	61.69	63.80	62.33
MMBT_w/Self-Training	66.11	64.70	67.75	69.37

Table.2 shows the performance of self-training with these two settings in comparison to the baseline. For self-training without confidence regularization using all default hyperparameters in the baseline model, the validation accuracy is 64.81 and the test accuracy is 67.70, which is 3.51 and 3.90 higher than the baseline model. For self-training with confidence regularization, the difference is not obvious. These results show that self-training has notable improvement in both validation and test. The improvement is a result of self-training's capability in bringing in new data and, more importantly, features that is relevant to the task by generating pseudo-labels using the model itself. Figure. 6 presents more details regarding the training loss and validation loss of self-training in comparison to baseline with regards to each iteration. From the two charts, we can see that the training loss in self-training is always higher than the training loss in baseline by 0.05. However, the validation loss in self-training falls below the validation loss in baseline after 1000 iterations, and the reduction in validation loss increases as the training goes on. This shows that the newly added data with pseudo-labels helps the model to generalize better in sacrifice of some performance in training accuracy.

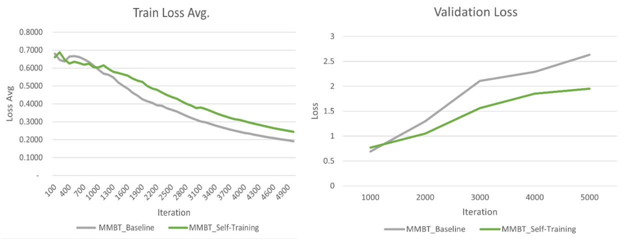


Figure 6: Loss Performance - Baseline vs. Semi-Supervised Self-Training

Although self-training brings some improvement in validation and testing, the improvement is not big given that the final test accuracy is still below 70. This might comes from many factors. Firstly, in this experiment, the Reddit Memes Dataset includes images that don't have much textual information and images that the textual information is not as easy to be parsed as those in the Facebook dataset. We, however, believe that the use of auxiliary dataset like this for self-training is still important in the task of hateful memes detection, because the real online memes are in different formats and are more challenging for the model to caption and decode. On the contrary, we believe that it is where self-training will help - to leverage the self-trained model to generate pseudo labels and capture relevant information among massive unlabelled data. Secondly, the lack of big improvement might be because the pseudo-labels are not balanced in classes. There are a total of 3249 images with pseudo-labels are added for retraining, but only 729 images are classified as hateful memes if label regularization is not applied. Another experiment showed that applying label regularization by setting threshold didn't help much. This might result in a class imbalance in the new training dataset, making the model less sensitive to hateful memes. To solve this, a better confidence regularization method such as model confidence regularization [13] might be beneficial in the future.

3.3 Enhancements Integration and hyper-parameters

tuning

In the experiment of enhancement integration and hyper parameter tuning, we integrated three augmented dataset to the train set: 1) 343 hateful memes from memotion dataset; 2) 4250 randomly selected Gaussian noise added memes 3) 4250 randomly selected text replaced noise memes. Additionally, 3259 pseudo labelled memes were included in the train set. In total, our training samples increased from 8500 to 20602, more than double the size of original training set. We grid search the hyperparametrs space including batch size, learning rate and number of warmup steps for the purpose of finetuning. The best selected parameters are batch size =16, peak learning rate=1e-5, and warm up epochs=2000 versus the MMBT base line parameters batch size =32, peak learning rate = 1e-5 and warm up epochs =2000. The results are shown in Table.3.

Table 3: Hyperparameters Tuning

Setup	Validation		Test	
	Acc.	AUROC	Acc.	AUROC
MMBT	61.30	61.69	63.80	62.33
MMBT_All Enhancements & Tuned	62.04	58.81	64.85	65.14

As shown in the graph of training loss, Figure.7, initial loss of our model was a bit higher than the base line loss. During the training proccess, the loss decreases slowly, which might be a result of loss plateau in around iteration 2000 to 2800. Since we setup adaptive learning rate and scheduler, we expected that the learning rate to decrease once the evaluation metric stop decreasing for a number of iterations and that after the optimizer jump out from the plateau, the loss continues to decrease thereafter. However, the validation loss for both baselines and all enhancements integrated one are increasing in a general train, which might indicates the issues of overfitting.

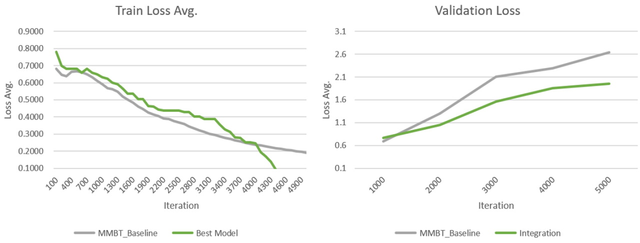


Figure 7: Loss Performance - Baseline vs. Hyper-parameters Tuning & Integration

It is also noteworthy that the final test accuracy 64.85 and AUROC 65.14 are lower than the self-training accuracy 67.75 and AUROC 68.20. This may because the self-training alone generated more labelled data that provides more desirable features from additional dataset. However, the 8500 memes of gaussian noise in images and textual noise posed a great challenge to the training since they weren't included in the original dataset for the first training in self-training. Moreover, since the test memes have high resolution in image and clear text, the strength of image or text noise in helping generalization is not obvious yet.

3.4 Text Representation and Fusion Experiment

In this experiment, we firstly tried to build some customized model using the existing encoders, modules and utilities from MMF. We set up pretrained Resnet152 for image encoders and extended fast sentence vector for encoders and concatenate the image and text tokens. We trained this customized model for 7000 iterations to check the performance of middle fusion and achieves 61.11 accuracy and 50.18 AUCROC on test set. The reason for the lack of improvement is that the text encoder may not leverage the pretrained representations. Rather, it simply projects to linear space. Further, concatenation is not advanced fusion to learn the combined representation from image and text.

4. Conclusion and Future Work

Data augmentation, semi-supervised self-training and hyper-parameter tuning approaches are used to boost multimodal hateful memes detection in this project. We concluded that both Gaussian image noise and text replacement delivered notable benefit on the model performance. Among these, our text replacement approach achieves 68.19 AUROC versus MMBT baseline 62.33 on the challenge testing set. Further direction on boosting data augmentation can be that utilizing GAN-derived model [10] to dynamically replace the text on the image without altering the style, which could enhance the knowledge on how the image and text link together in a semantic way.

In semi-supervised self-training, the model achieved 69.37 AUROC versus 62.33 AUROC in baseline model. However, there are still limits in this experiment. Finding higher-quality memes dataset, handling class imbalance with better confidence regularization term and improving image captioning and text captioning techniques are worthy of more researches. Moreover, since the combination of data augmentation and self-training didn' bring add-up result in our experiment, the combination of self-training and data augmentation is also a good starter for next step.

Last but not least, due to technical contraints in the MMF framework during our project period, we didn't succeed in explorations such as further pretraining models on additional dataset that might increase the model's capability to identify sensitive age, gender, race and ethnicity contents, as well as model architecture enhancement using RoBERTa, StructBERT and other newly developed text encoders. These are all important paths for further exploration in multimodal hateful memes detection.

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1, 2

[2] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997. 4

[3] 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. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[5] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019. 2

[6] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 4

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016. 2

[8] Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. Semi-supervised self-training of object detection models. 2005. 3

[9] Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2020–2029, 2019. 2

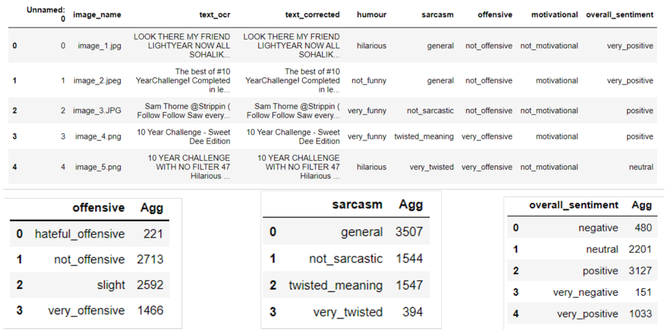
[10] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1500–1508, 2019. 6

[11] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 2

[12] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020. 4

[13] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5982–5991, 2019. 5

Appendix



	Unnamed: 0	image_name	text_ocr	text_corrected	humour	sarcasm	offensive	motivational	overall_sentiment
0	0	image_1.jpg	LOOK THERE MY FRIEND LIGHTYEAR NOW ALL SOHALIK...	LOOK THERE MY FRIEND LIGHTYEAR NOW ALL SOHALIK...	hilarious	general	not_offensive	not_motivational	very_positive
1	1	image_2.jpeg	The best of #10 YearChallenge! Completed in le...	The best of #10 YearChallenge! Completed in le...	not_funny	general	not_offensive	motivational	very_positive
2	2	image_3.JPG	Sam Thorne @Strippin (Follow Follow Saw every...	Sam Thorne @Strippin (Follow Follow Saw every...	very_funny	not_sarcastic	not_offensive	not_motivational	positive
3	3	image_4.png	10 Year Challenge - Sweet Dee Edition	10 Year Challenge - Sweet Dee Edition	very_funny	twisted_meaning	very_offensive	motivational	positive
4	4	image_5.png	10 YEAR CHALLENGE WITH NO FILTER 47 Hilarious ...	10 YEAR CHALLENGE WITH NO FILTER 47 Hilarious ...	hilarious	very_twisted	very_offensive	not_motivational	neutral

	offensive	Agg
0	hateful_offensive	221
1	not_offensive	2713
2	slight	2592
3	very_offensive	1466

	sarcasm	Agg
0	general	3507
1	not_sarcastic	1544
2	twisted_meaning	1547
3	very_twisted	394

	overall_sentiment	Agg
0	negative	480
1	neutral	2201
2	positive	3127
3	very_negative	151
4	very_positive	1033

Figure 8: Memotion 7K Description Preview