Facebook Hateful Memes Challenge
Midterm Assignment Report

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

1. Abstract

In the past few years, there has been a surge of interest in multi-modal problems, from image captioning to visual question answering and beyond. In this paper, we focus on hate speech detection in multi-modal memes wherein memes pose an interesting multi-modal fusion problem. Unimodally, the text on the image and the image itself are harmless but when combined, they can represent something different leading it to be either hateful or non-hateful memes. A crucial characteristic of the challenge is that it includes "benign confounders" to counter the possibility of models exploiting uni-modal priors. Thus, the challenge is designed such that it should only be solvable by models that are successful at sophisticated multi-modal reasoning and understanding. The majority of the baseline multi-modal models tend to find the relationship between the alignment of the two modalities and sometimes gives more preference to the hate speech (text modality). In our first approach, we try to explore the image modality by passing it through an image captioning model to fetch the "actual caption" and then combine it with the image embedding and the pre-extracted caption to get a binary classification. This approach might help in getting a better relationship of the two modalities and thus enhancing the results. Another approach we tend to try is to aid the prediction with sentiment analysis. Instead of only using multi-modal representations obtained from pre-trained neural networks, we also include the uni-modal sentiment to enrich the feature. The, we concatenate them with multi-modal representations and then use an MLP to make the final prediction.

2. 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 multi-modal machine translation, it has been observed that baseline models using the language domain perform well without even exploiting the multi-modal understanding and reasoning(Devlin et al., 2015). However, the Facebook Hateful Memes Challenge Dataset is designed in such a manner that uni-modal models exploiting just the language or vision domain separately will fail, and only the models that can learn the true multi-modal 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 an 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 1 for an illustration. Thus, this challenge set is an excellent stage that aims to facilitate the development of robust multi-modal models, and at the same time addresses an important real-world problem of detecting hateful speech on online social media platforms. Majority of the prior work baselines aim at solving this problem by finding an alignment between the two modalities, but it faces the hardship of not knowing the context behind the image and the text combination.

In this paper, we introduce two major ideas wherein we try to explore the two modalities using pre-trained Image captioning models and sentiment analysis to understand the context and relationship between the two modalities. Many of the baselines tend to focus more on the text modality for hate speech. Here, we try to balance the representations of the two modalities by fetching a deeper understanding of the image via captioning and using its image embedding as well and then using attention techniques. Also, sentiment analysis is another approach which will aid in understanding the positivity or negativity of the modalities and get a better representation along with the multi-modal models results.

Figure 1. Multi-modal "mean" meme and Benign confounders. Mean meme (left), Benign text confounder (middle) and Benign image confounder (right).

In what follows, we discuss the related prior work for such a problem in the next section (3), followed by defining the problem statement (4) and discussing the multi-modal baselines in section (5). We then present our Experimental methodology with its results and discussion in section (6) and (7) followed by explaining our new ideas related to the Hateful Meme Challenge in section (8).

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 multi-modal tasks and problems, ranging from visual question answering to image captioning and beyond. However, there has been surprisingly little work related to multi-modal 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 Multi-modal publications has not been addressed yet. So they have created and made available MMHS150k, a manually annotated multi-modal hate speech dataset formed by 150,000 tweets, each one of them containing text and an image. The data points are labeled 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 multi-modal 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 an 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.

Our first approach tries to extend this idea of introducing a deeper understanding of the visual domain. We plan to use pre-trained image captioning models to get the "actual caption" from the image along with the image embeddings and add these via attention models with the text embedding of the pre-extracted caption in the meme. This will help in finding the relationship between the image actual caption and the text in the meme which can tell whether both are related or not and thus should give better results.

Vision and language problems have gained a lot of traction is recent years [(Mogadala et al., 2019)], with great progress on important problems such as visual question answering [(Goyal et al., 2017)] and image caption generation and retrieval [(Sidorov et al., 2020),(Gurari et al., 2020)].

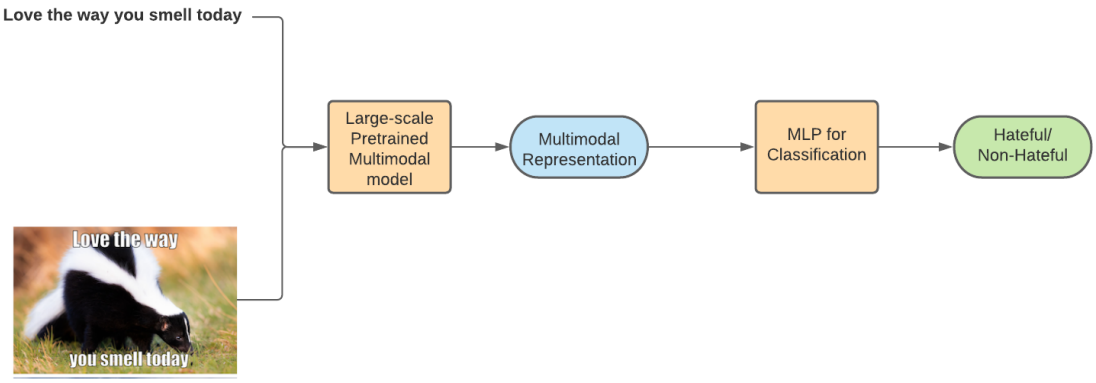

Figure 2. Multi-modal baseline models architecture. Pre-trained multi-modal model block can be VisualBERT or ViLBERT.

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 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 bi-linear transformation. They use attention mechanisms to further improve their results. The context vector is the 1D-convolution output from the 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.

There has been extensive research in multi-modal sentiment [(Soleymani et al., 2017)], but there is no agreed-upon standard dataset or benchmark task. We try to introduce a sentiment analysis approach as our second experiment wherein we pass both the textual and the visual modality via uni-modal sentiment analysis models to get the orientation of both the modalities. We then use these along with the large scale pre-trained multi-modal model. This will help in overcoming many of the error cases which can be seen in the later sections.

4. Problem Statement

The objective of this challenge is to classify memes as hateful or benign while considering their information from both text and visual modality. Denote the visual components of all memes by $X_1 = \{I_1, ..., I_i\}$ where i is the index of the memes, and in our case, the visual component I is the meme itself. Let $X_2 = \{T_1, ..., T_i\}$ denotes the text extracted from the memes. If phrases locate in multiple regions of a single meme, the corresponding T will include all the text information by concatenation. Let $Y = \{y_1, ..., y_i\}$ be the corresponding labels of all memes, where each $y \in \{0, 1\}$ with 0 means benign and 1 indicates a hateful meme. Thus, our task can be formulated as a binary classification problem with X_1 and X_2 as input. The goal of our paper is to model the $P(Y|X_1, X_2)$, denoted by p_θ, which minimize the following cost function:

$$J(\theta) = \sum_i -(Y log(p_\theta) + (1 - Y) log(1 - p_\theta)) \quad (1)$$

5. Multi-modal Baseline Models

In this section, we introduce two previous approaches for Facebook hateful memes challenge. In specific, we will discuss how multi-modal information of the input memes are fused through large-scaled pre-trained model named VisualBERT(Li et al., 2019) and ViLBERT(Lu et al., 2019) to better distinguish the hateful memes from the benign ones. The architecture of the two modals are very similar, which is shown in Figure 2.

5.1. VisualBERT

In order to utilize the VisualBERT, multiple region features $f_1, f_2, \ldots, f_n$ are first extracted from input image I using Faster RCNN (Ren et al., 2015). Each region feature f is then converted to visual embedding e_v by following equation.

$$e_v = f + e_s \quad (2)$$

where e_s stands for segment embedding, which indicates whether the input is text or image.

For the text input, the textual embedding e_t is obtained in a

similar way:

$$e_t = f_t + e_s + e_p \tag{3}$$

where f_t is the token embedding for each token in the sentence, and e_p is the positional embedding to indicate the relative position of each token.

After concatenating e_v and e_t, the embedding is sent into pre-trained VisualBERT model for further processing.

VisualBERT (Li et al., 2019) is a pre-trained model for learning joint contextualized representations of vision and language. It contains multiple transformer blocks on top of the visual and text embedding. It is pre-trained on Microsoft COCO captions (Chen et al., 2015) with two objectives: masked language modelling and sentence-image prediction task. The masked language modelling is very similar to the approach in sentence BERT (Devlin et al., 2018), where some input text tokens are masked randomly, and the model needs to predict what are the original tokens. The sentence-image prediction requires the model to decide whether the input text matches the image.

The VisualBERT output of the first token is used as the multi-modal representations e_m. An MLP is then used to make the final prediction. The model is fine-tuned for the current task by using the following loss function.

$$l(\theta) = CrossEntropyLoss(W \cdot e_m, y) \tag{4}$$

where e_m is a vector of size h. h is the hidden size of VisualBERT. W, which has a shape of 2 by h, is the learnable matrix of the MLP. θ denotes the parameters of the entire model, including the W.

5.2. ViLBERT

Following a similar way in the last subsection, the e_v and e_t are extracted. After adding spatial information of the extracted region into each e_v, the two embeddings are passed through ViLBERT to obtain the multi-modal representation e_m.

The ViLBERT (Lu et al., 2019) is another pre-trained model for learning contextual representations, but has a different architecture from VisualBERT. Embedding from each modality is sent into separate transformers with self-attention and co-attention. The co-attention is introduced as a mechanism where the key and values used for attention are obtained from the other modality. The ViLBERT is pre-trained on Conceptual Captions(Sharma et al., 2018). In addition to masked language modelling and sentence-image matching objectives, the model also needs to predict the semantic class of masked image regions.

Same as the approach in the last subsection, an MLP is appended after the ViLBERT, and the entire model is fine-tuned using cross-entropy loss as shown in equation 4.

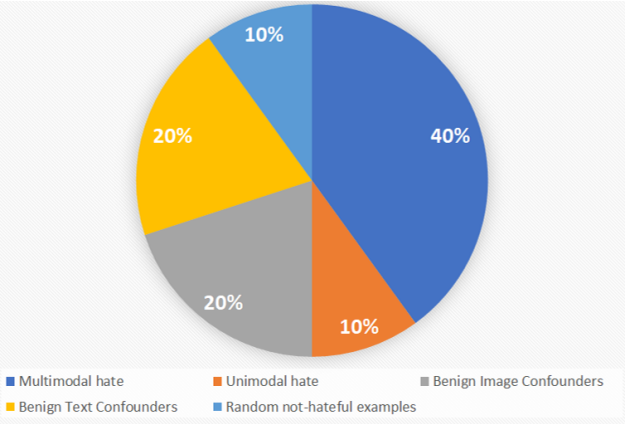

Figure 3. Types of memes in the Facebook Hateful Memes Challenge Dataset

6. Experimental Methodology

6.1. Dataset

We have used 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 3: multi-modal hate, where benign confounders were found for both modalities, uni-modal 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. Each image in the training and validation set are annotated as either 1 or 0 which corresponds to the hateful and benign classes respectively.

6.2. Evaluation Metrics

We have evaluated the performance of our classifier using the following two metrics as suggested in the challenge

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

Receiver Operating Characteristics curve is a graph of True Positive Rate (TPR) v/s False Positive Rate (FPR). It measures how well the binary classifier discriminates between the classes as its decision threshold is varied.(Bradley, 1997). A perfect classifier will have an area under the curve of 1, where the top left corner in the plot is the ideal point with a TPR of 1 and a FPR of 0. Thus, a larger area under the curve is desirable for any classifier to maximize TPR and minimize FPR.

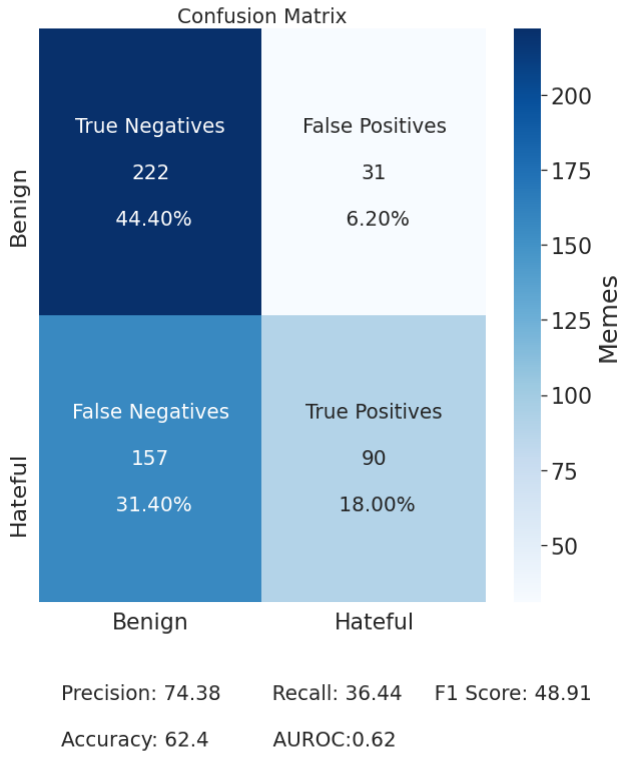

Figure 4. Confusion Matrix for VisualBERT COCO model.

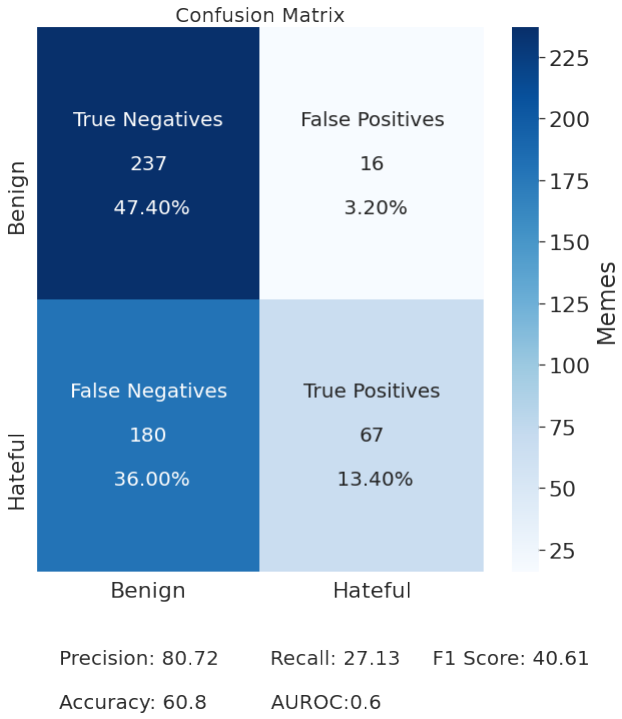

Figure 5. Confusion Matrix for ViLBERT CC model.

6.2.2. 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.

For analysis, we selected two multi-modal baseline models mentioned in (Kiela et al., 2020) namely the VisualBERT pretrained on COCO dataset and ViLBERT pretrained on the Conceptual Captions dataset. We fine-tune the models on our dataset following the same training guidelines as mentioned in the original challenge paper. We then evaluated both the models on the validation set comprising of 500 memes.

7. Results and Discussion

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 uni-modal 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.

About 40% of the predictions in both the multi-modal models are False Positives or False Negatives as depicted by the confusion matrices for VisualBERT COCO model in figure 4 and for ViLBERT CC model in figure 5. Also, both the models have low Precision and Recall. The difference in accuracies by multi-modal baselines and uni-modal baselines is relatively small, where both perform slightly better than random and majority-class baselines, since the dataset is balanced. The baseline models are very far from the human accuracy on test set which is 84.7%. All of this indicate that the multi-modally pretrained models can be improved much further by incorporating techniques and knowledge in the classifier which will help it to understand the underlying relations between the image and captions better. For code reference here.

When we look at the most confidently predicted False Positives by both the classifiers, we believe that the benign memes are labeled as hateful just by the presence of certain keywords in the captions which target a particular community, race, religion, sexual orientation etc. e.g. Country names in figure 6 or words like 'blackhead' in 7. Similarly, when we study the most confidently predicted False negatives, we find that the models fail to identify the sentiment in the image domain as seen in figure 8 and the irony behind the combination of image and caption used in figure 9.

Annotated label: Benign Predicted label: Hateful Probability: 0.9995

Figure 6. Most confident False Positive predicted by VisualBERT COCO model.

Annotated label: Benign Predicted label: Hateful Probability: 0.9998

Figure 7. Most confident False Positive predicted by ViLBERT CC model.

Annotated label: Hateful Predicted label: Benign Probability: 0.0002

Figure 8. Most confident False Negative predicted by VisualBERT COCO model.

Annotated label: Hateful Predicted label: Benign Probability: 0.0

Figure 9. Most confident False Negative predicted by ViLBERT CC model.

8. New Research Ideas

8.1. Using Image Captioning

As we can see in the Figure 10, We first pass the pre-extracted text caption ($X_2 = \{T_1, ..., T_i\}$ denotes the text extracted from the memes.) from a Language model like BERT, which provides us with a text embedding (h1, h2, h3,...). Parallelly, we pass the image features ($X_1 = \{I_1, ..., I_i\}$ where i is the index of the memes, and in our case, the visual component I is the meme itself) via a Visual model and get the image representation (can use the second last layer of RESNET) . To find a better relationship between the image and the text caption, we pass the image representation via a pre-trained image captioning model to fetch deeper understanding of the image modality and get an "actual caption" of the image (a1,a2,a3,...) . We then plan to combine these three outputs, i.e., the pre-extracted text representation (h1,h2,h3,...), the image representation (second last layer of ResNet for example) and the "actual caption" representation (a1,a2,a3,...) and concatenate them (using attention mechanism) to find a better relationship between the two modalities. We then pass the output via an Multi-layer perceptron to get a binary classification of hateful and non-hateful memes (0/1).

As we saw from our error analysis in the previous section, most of the multi-modal baselines tend to focus more on the text modality for the hate speech. Our focus in using this approach is to find a deeper relationship between the text modality and the image modality by bringing the image modality and finding its "actual caption" and also parallelly sending the image representation for concatenation step. This will help in bringing the focus on both the modalities and thus could define a better representation for the problem statement. Moreover, comparing the "actual caption" with the "pre-extracted caption" of the meme will help in understanding whether both are aligned or not because in

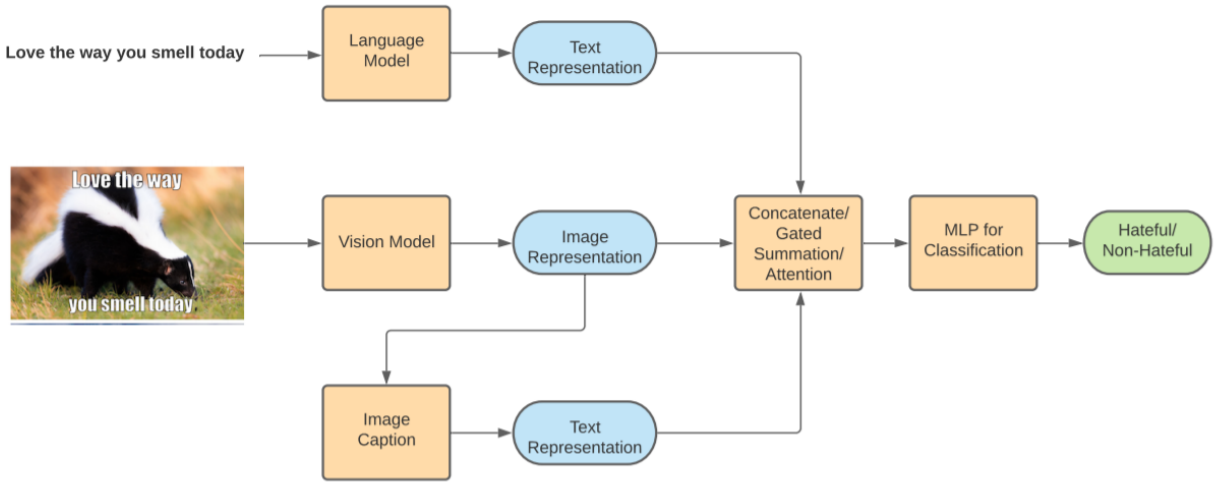

Figure 10. Approach 1 - Using Image captioning.

many cases a hateful image is turned benign just by declaring what is happening in the image. Thus, this approach takes into account both the image representation and its caption as an input to the attention. Exploring other fusion techniques like gated summation and attention fusion methods will help in improving the results even further. Also, pre-training models on image and texts on large-scale data set 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.

8.2. Using Sentiment Analysis

Another approach we tend to try is to aid the prediction with sentiment analysis. Instead of only using multi-modal representations obtained from pre-trained neural networks, we also include the uni-modal sentiment to enrich the feature. Both text and image sentiment analysis models are also pre-trained and transferred to our task as there is no annotated data we could use to train those models from scratch. After getting the uni-modal sentiment analysis results, we concatenate them with multi-modal representations and then use an MLP to make the final prediction. The model architecture is shown in Figure 11.

The intuition for this idea is that current pre-trained representations, like VisualBERT and ViLBERT, have the objective of predicting the semantic correlation between image and text, but semantic information is difficult to capture and may not be enough for solving our task. We try to include high-level features like text and image sentiments because sentiment analysis is a related but relatively simple task. Several error cases can be alleviated by incorporating sentiment information. We used some of the state-of-the-art models with Flair which allows you to apply state-of-the-art natural language processing (NLP) models to sections of text. Flair utilizes a pre-trained model to detect positive or negative comments and print a number in brackets behind the label which is a prediction confidence. TextBlob is another such method which is based on the polarity and subjectivity and determines the positive, negative or neutral text. Using these uni-modal sentiment models and combining it with the multi-modal representation would hopefully aid the contextual representation with sentiment information will improve the model performance.

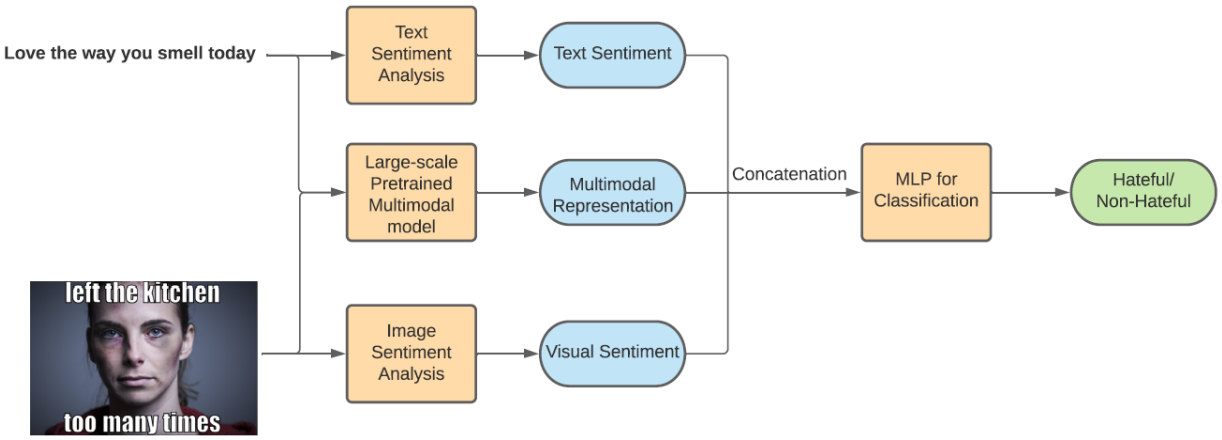

Figure 11. Approach 2 - Using Sentiment analysis

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/.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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., Gupta, S., Girshick, R., Mitchell, M., and Zitnick, C. L. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

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.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Gurari, D., Zhao, Y., Zhang, M., and Bhattacharya, N. Captioning images taken by people who are blind, 2020.

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.

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., and Chang, K.-W. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

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.

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

Mogadala, A., Kalimuthu, M., and Klakow, D. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *CoRR*, abs/1907.09358, 2019. URL http://arxiv.org/abs/1907.09358.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

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.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. Textcaps: a dataset for image captioning with reading comprehension, 2020.

Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S.-F., and Pantic, M. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3 – 14, 2017. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2017.08.003. URL http://www.sciencedirect.com/science/article/pii/S0262885617301191. Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

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.