Multimodal Detection of Hateful Memes by Applying a Vision-Language Pre-Training Model

Yuyang Chen (ychen@putnamscience.org)
Putnam Science Academy, 18 Maple St, Putnam, CT, USA

Feng Pan
Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Jiefang Ave 1277, Wuhan 430022, China https://orcid.org/0000-0003-2820-768X

Method Article

Keywords: Artificial Intelligence, Deep Learning, Multimodal, Hate Speech, Self-attention Mechanism

Posted Date: April 11th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1414253/v2

Multimodal Detection of Hateful Memes by Applying a Vision-Language Pre-Training Model

Yuyang Chen[1], Feng Pan[2]

[1] Putnam Science Academy, Putnam, CT 06260, USA

[2] Department of Radiology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430022, China.

Abstract

Detrimental to individuals and society, online hateful messages have recently become a major social issue. Among them, one new type of hateful message, named "hateful meme", has emerged and brought difficulties in traditional deep learning-based detection. Because hateful memes were formatted with both text and image to express users' intents, they cannot be accurately identified by singularly analyzing embedded text or images. In order to effectively detect a hateful meme, the algorithm must possess strong vision and language fusion capability. In this study, we move closer to this goal by feeding a triplet by stacking the visual features, object tags, and textual features of memes generated by the object detection model VinVL and the optical character recognition (OCR) technology into a Transformer-based Vision-Language Pre-Training model OSCAR+ to perform the cross-modal learning of memes. After fine-tuning and connecting to a random forest classifier (RF), our model (OSCAR+RF) achieved a 0.768 AUROC score on the hateful meme detection task in a public dataset, which was higher than the published baselines. In conclusion, this study has demonstrated that Vision-Language PTMs with the addition of anchor points can improve the performance of deep learning-based detection of hateful memes by involving a more substantial alignment between the textual and visual information.

Keywords — Artificial Intelligence, Deep Learning, Multimodal, Hate Speech, Self-attention Mechanism.

1. Introduction

The hateful message, more commonly known as hate speech, has unfortunately almost become a ubiquitous phenomenon on social media. After reviewing the definitions of hateful message from some prominent organizations or individuals, such as YouTube, Facebook, Twitter, etc., the hateful message can be defined as a statement that explicitly or implicitly expresses hatred or violence against people with protected characteristics (**S1**). This definition distinguishes hateful messages from an offensive by their targets: though offensive language can be directed at either individuals or groups (i.e., "Get out of this place!"), it does not target them due to their protected characteristics. In sum, hateful messages can be regarded as content that expresses hatred against an individual or

a group due to their protected characteristics. They are detrimental to both individuals and the society, leading to prejudice against individuals, depreciation of minority's abilities, alienation of minorities, degrade in individual mental health, a rise of the suicide rate, an increase of offline hate crimes, and discriminatory practices when allocating public resources [1-6].

Though the U.S. government is legally prohibited by the first amendment from restricting hateful messages, as analyzed by the American Bar Association, privately-owned digital platforms still have unparalleled agency in taking a more active stance against hateful messages [7, 8]. Most major digital social media platforms, including Facebook, Twitter, and YouTube, have developed respective moderation policies against hateful messages. However, the outdated manual moderation is considered insufficient, as human moderators are slow and expensive. Besides, content moderators are known to suffer PTSD-like syndromes after repetitively reviewing violent and exploitative content [9]. As a result, attempts to automate the identification-removal process have been carried out to reduce cost while increasing speed.

From generation to generation, different automatic hateful message detection methods were developed. As the most straightforward ways, block-word-list and dictionary-based methods can filter out hateful messages but are ineffective because hateful users can use typical expressions to circumvent the detection. With the development of machine-learning and deep-learning technology since 2009, these dilemmas have been changed [10-16]. As early as 2009, there had been a study on detecting online harassment using a support vector machine (SVM), yet its detection accuracy was only 0.44 [15]. Over time, the detection accuracy has been dramatically improved with the f1 score reaching 0.95, since the continuous algorithm optimization and applications of variable networks, such as long short term memories (LSTM), gradient boosted decision tree (GBDT), convolutional neural network (CNN), and gated recurrent unit (GRU) [10-14, 16]. Though their works had primarily surpassed traditional machine learning methods, they focused on only one modality - textual messages; so, their work could not be applied to the real-world situation where a new type of "hateful messages" named "hateful memes" emerged and have become rampant, because they are more subtle [17].

Hateful memes have a format that leverages text and images to express users' intents. Specifically, many sentences or images that are harmless by themselves may become hateful when combined (**Figure 1**). So, hateful users intentionally publish content in which neither text nor image alone can determine whether this piece of content is hateful [17]. These subtle references are easy for humans to understand yet difficult for machines to detect. Traditional automatic textual detection completely ignoring a substantial proportion of visual features cannot be applied in the images, let alone the memes involving both text and image simultaneously. In order to develop a detection model to capture the complexity of the memes, the model should be not only able to process every single modality but also be capable of fusing the two modalities.

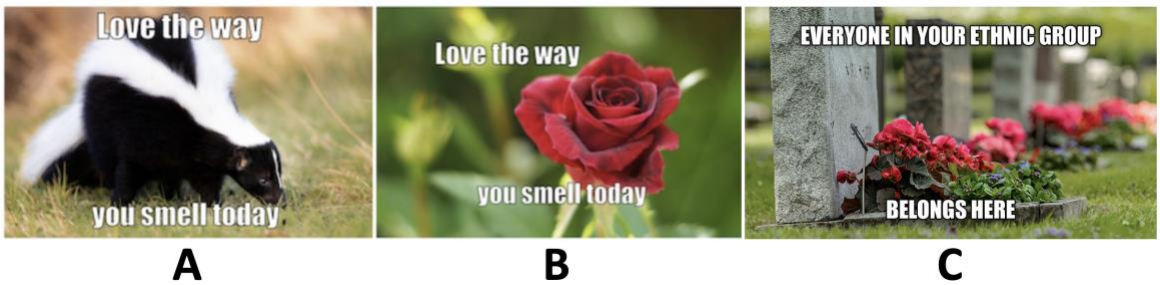

Figure 1: Examples of hateful and non-hateful memes. In meme A (hateful), the image of a skunk and the sentence that "Love the way you smell today" are both neutral, but the combined version of them intends to insult the viewer; in meme B (non-hateful), with the same caption inscribed as meme A, but the image part changed from a skunk to a rose, has a reversed implication from insulting to complimenting as non-hateful; in meme C (hateful), neither the roses and cemetery in the image nor the caption expresses hatred alone, yet

when combined, this meme expresses racism. (Original images from the "Facebook Hateful Meme Dataset", https://www.kaggle.com/parthplc/facebook-hateful-meme-dataset)

However, the advent of Large-scale Pre-trained Models (PTM) has shed hope into overcoming the challenge mentioned above. PTMs such as GPT (Generative Pre-trained Transformers) and BERT (Bidirectional Encoder Representation from Transformer) have recently achieved great success in many complex natural language processing (NLP) tasks and become a milestone in the wider machine learning community [18]. PTMs are now generally used as backbones for downstream tasks. The rich knowledge stored implicitly in the tremendous amount of model parameters could be leveraged by fine-tuning them for specific tasks. Following the success of PTMs on language tasks, the community has proposed various PTMs for vision tasks, such as VilBERT and VLBERT [19, 20].

Furthermore, they could be optimized to directly combine learned vector representations of both the text and image to reach state-of-the-art results in several Vision-Language tasks [21, 22]. As a result, the Vision-Language PTMs are developed and pre-trained using text orientation for individual language tasks, image orientation for individual image tasks, and cross-modal loss for Vision-Language fusions. Most models adopt the Image-Text-Matching task for cross-modal pre-training. Some models add other pre-training tasks to complete their multimodal knowledge. These large models are shown to be able to capture more complex information and perform better on many Vision-Language tasks. However, ambiguity is a big issue, which means these models cannot understand the different image region features well in the same Vision-Language semantic space due to the oversampling, leading to the wrong classification in the downstream tasks [23].

To overcome this ambiguity issue as mentioned above, we built a novel hateful meme detector by applying a Transformer-based Vision-Language Pre-Training model with a triplet input. Unlike previous models, our model additionally takes object tags of the image into account, serving as the anchor points. The thought of anchor point is motivated by the observation that modern object detectors can accurately detect the salient objects in an image, which are often mentioned in the paired text [24]. Involving the object tags can not only help to facilitate visual feature classifications but also bridge the textual caption and related visual image [23]. In this study, we aimed to test the accuracy of this novel model in detecting hateful memes in a public dataset and compare its performance with other published methods.

2. Materials and Equipment

Our model was trained and tested in a public dataset named "Facebook Hateful Meme Dataset" (https://www.kaggle.com/parthplc/facebook-hateful-meme-dataset). In constructing this dataset, researchers at Facebook first reconstruct online memes by placing meme text over a new underlying licensed image without the loss of meaning. Then, they hired annotators from a third-party annotation company rather than a crowd-sourcing platform. The annotators were trained for 4 hours in recognizing hateful memes according to Facebook guidelines (https://www.facebook.com/communitystandards/hatespeech) (**S1**). These annotators reconstructed the memes and made the classification. For every hateful meme, there is always a non-hateful alternative whose caption or image is changed from the original one. This kind of substitution is called "benign confounders", a technique similar to current strategies of using counterfactual or contrastive examples [22, 24, 25]. After the memes were labeled hateful, "benign confounders" were constructed. Finally, a dataset with a total of 10k memes was set up and categorized into hateful or non-hateful. A dev and test set constituted 5% and 10% of the data, respectively, and the rest served as a training set.

3. Methods

3.1 Detection pipeline

In this model, an optical character recognition (OCR) module is applied to extract the textual part of the memes, and an object detection (OD) module VinVl (Visual features in Vision-Language) is used to encode the correlated image part of memes. Then, we choose a Vision-Language PTM named OSCAR+ (Object-SemantiCs Aligned Re-training) to encode the extracted textual part and encoded image part of memes [24, 25]. OSCAR+ was established from the basis of multi-layer Transformers like most other PTMs [26]. However, unlike most existing PTMs, which simply concatenate image region features and textual features as input and resort to the self-attention mechanism to learn semantic alignments between image and text in a brute force manner, in OSCAR+, we can additionally input the embeddings of object tags to bridge the textual and related visual image.

We used the Google Colab platform to provide accelerators for inference and training, 208 the hardware accelerators used are an Nvidia P100 GPU and an 8-core TPU (Tensor 209 Processing Unit) V3. We first fine-tune OSCAR+ on the hateful meme dataset. In this stage, a Fully-Connected Neural Network (OSCAR+FCNN) is connected to OSCAR+ output (OSCAR+FCNN). The mini-batch gradient descent is carried out on the training set of 8500 images with a batch size set to 50 and a learning rate of 0.000002; the loss function is set to Binary Cross-Entropy Loss with Logits $L(x, y) = -(y \ln \sigma(x) + (1 - y) \ln (1 - \sigma(x))$. After OSCAR+ is successfully fine-tuned, its output is connected to a random forest classifier. Then, the random forest is further optimized, consisting of ten decision trees whose maximum depth is set to 10. The trained random forest is the final classifier (OSCAR+RF) for recognizing hateful memes. The pipeline of our model construction is illustrated in **Figure 2**.

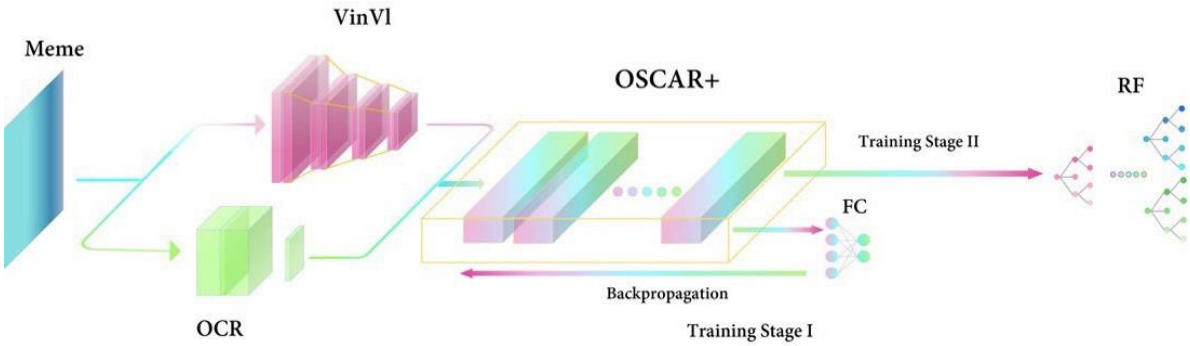

Figure 2. Full pipeline. FC denotes Fully Connected Layers; RF denotes Random Forest.

3.2 Meme preprocessing

Because the meme itself can not be fed directly into OSCAR+, it must be preprocessed into a suitable data format (**Figure 3**). In preprocessing, the meme is first input into a VinVl object detector which uses the ResNeXt 152-C4 as the backbone feature extractor [27, 28]. In VinVl, the backbone network transforms the input meme into a feature map. A Region Proposal Network (RPN) outlines regions of interest (ROIs) on the feature maps containing predefined categories of objects. The features of ROIs are then pooled into ROI-Pooling vectors of 2048 dimensions. Afterwards, those vectors are passed through two paralleled FCNN pathways. The first one is used to predict

the bounding box's position and size for each ROI-Pooling vector. Each Bounding-Box-Regression vector is then concatenated with the corresponding ROI-Pooling vector to form an intermediate vector of size 2054. These intermediate vectors are further passed through another FCNN to produce image feature embedding vectors with a size of 768. The other one is used to predict the category of the object (text object tags) in the corresponding ROIs. In general, the VinVL object detector will produce two sequences of vectors after meme input: the textual object tags and the corresponding image feature embeddings (**Figure 3A**) [25].

On the other hand, the meme is input into an OCR module to extract all the caption text that appears in the memes [29]. Then, both texted object tags and extracted text caption were tokenized by OSCAR+, forming embedding vectors with a size of 768. At last, the text embeddings, the object tag embeddings, and the image feature embeddings are then further concatenated in sequence, inserting between which the embedding of special token **[SEP]** that denotes different sections of OSCAR+ input and adding an embedding of the special token **[CLS]** is then appended at the start of the whole sequence. Afterwards, a meme was transferred into an embedded triplet for further OSCAR+ encoding (**Figure 3B**) [24].

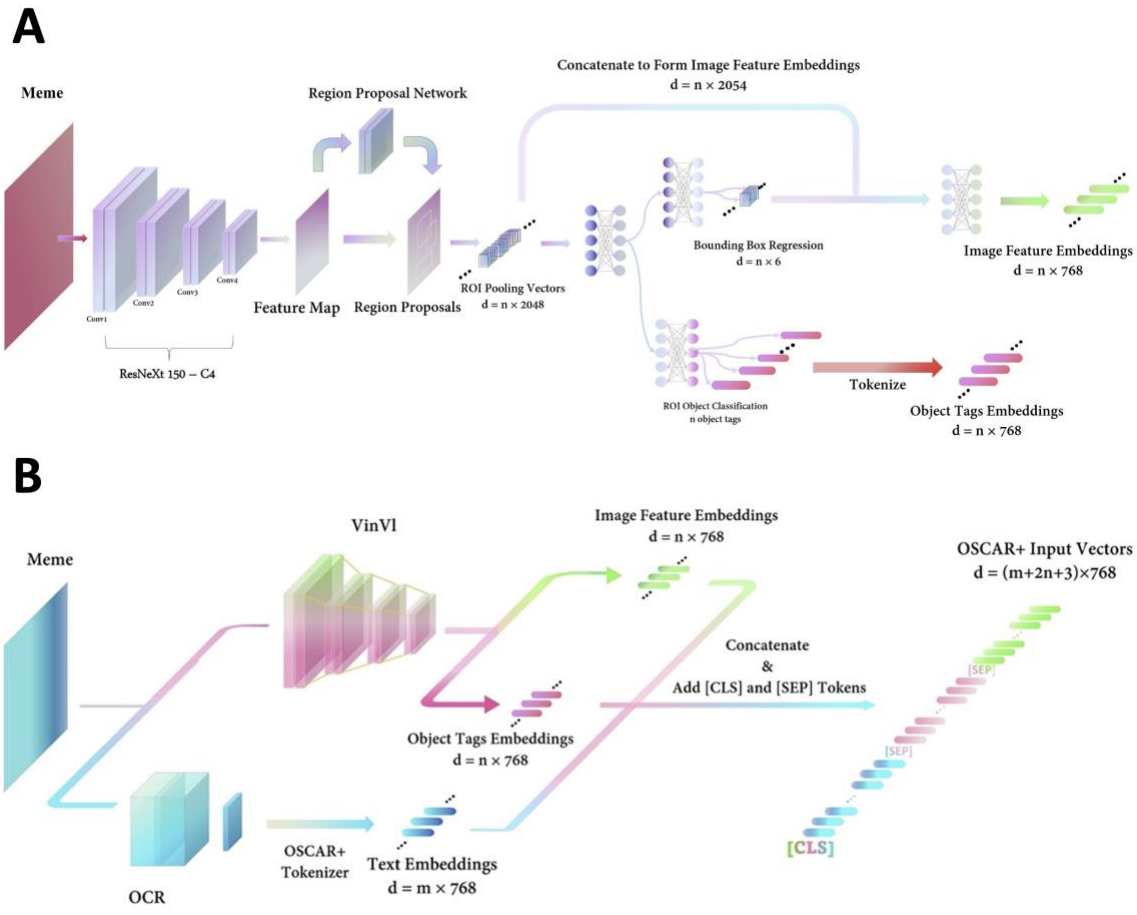

Figure 3. Meme preprocessing. A. Embeddings of image features and object tags in VinVL. B. Meme preprocessing by VinVL and OCR in parallel results in a triplet input of image feature embeddings, object tags embeddings, and text (or caption) embeddings to OSCAR+.

3.3 Encoding process in Vision-Language PTM

The OSCAR+ model consists of 12 same encoder blocks (**Figure 4A**) [24]. The first encoder takes the preprocessed meme embeddings with a matrix of dimension $(m + 2n + 3) \times 768$ as input, which consists of the image features embeddings (m), the object tag embeddings (n), the text embeddings (n) and the embedding of special tokens (**[CLS]**, **[SEP]**, and **[SEP]).** The followed encoder block takes the output embedding sequence produced by the previous encoder as input (**Figure 4A**). In every encoder block (**Figure 4B**), the input embedding sequence is passed through 12 self-attention heads in parallel, each outputting a smaller matrix of dimension $(m + 2n + 3) \times 64$. More specifically, in each self-attention head, the input matrix will pass through three separate FCNN to produce three smaller matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ of dimension $(m + 2n + 3) \times 64$. Then, the standard dot product attention operation was carried out as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \qquad (1)$$

where d_k equals 64 (**Figure 4C**) [26]. These intermediate output matrices are again concatenated to form a matrix of the original size, passed through an FCNN, then added with the original matrix, and normalized by rows. This normalized matrix is given through an FCNN, added with itself, and normalized by rows again. Finally, the encoder block will produce a matrix of dimension $(m + 2n + 3) \times 768$.

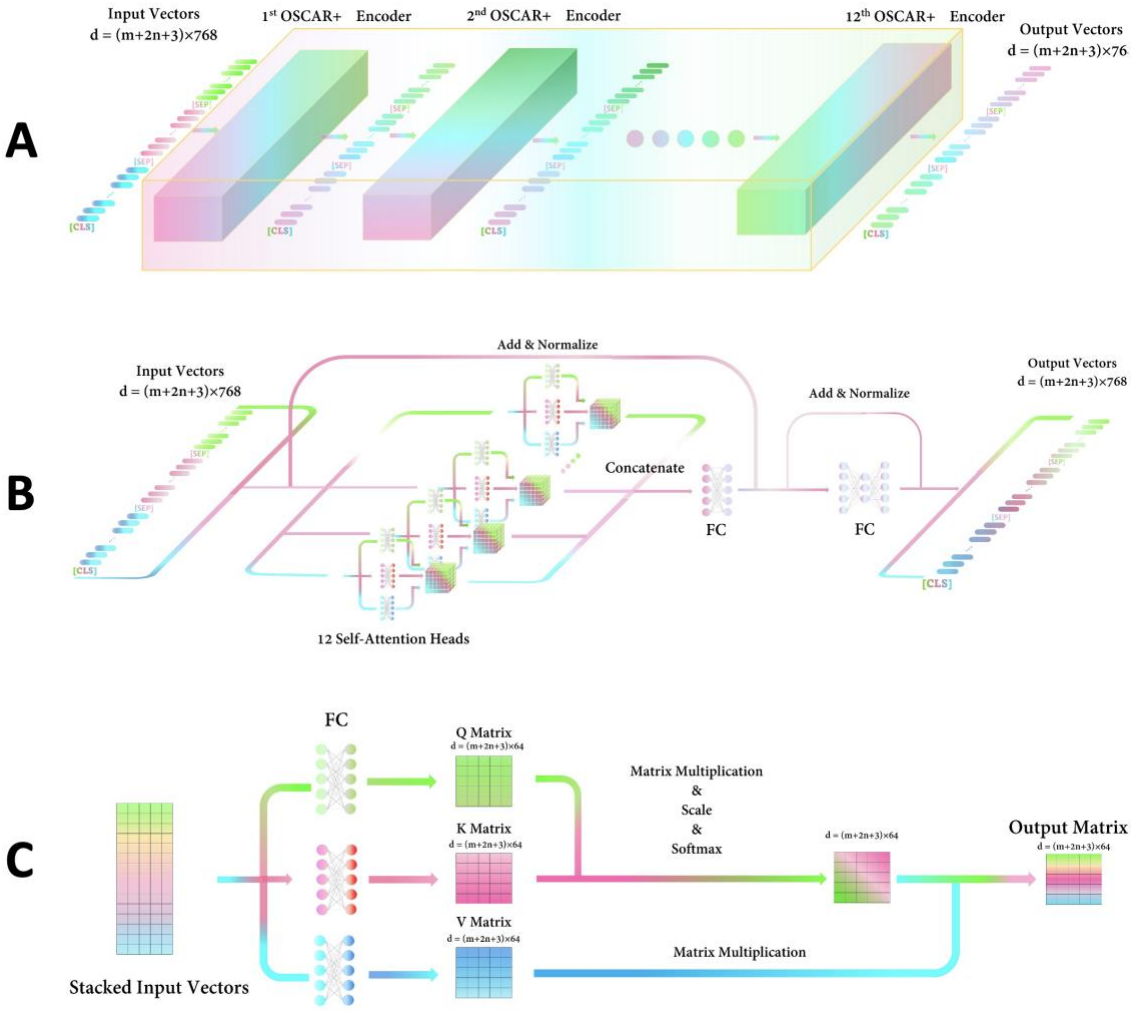

Figure 4. Architectures of OSCAR+ and its encoder. A. OSCAR+ consists of 12 tandem encoders; B. the architecture of an encoder; C. structure of the self-attention head.

3.4 Comparisons with baselines

We compared the hateful meme detection accuracy (Acc.) and area under the receiver operating characteristic (AUROC) between our model to other published methods as baselines trained and tested on the same datasets [17]. The baselines included both unimodal PTMs and multimodal PTMs. The unimodal PTMs pre-trained in text data was BERT (Text BERT) [14]. The unimodal PTMs pre-trained in image data included standard ResNet-152 with average pooling (Image-Grid) and with fine-tuned FC6 layer by using weights of the FC7 layer (Image-Region) [30], ViLBERT [20], and Visual BERT [31]. The multimodal baseline methods learning both text and related image in pre-training include supervised multimodal bi-transformers using either Image-Grid or Image-Region features (MMBT-Grid and MMBT-Region) [32], ViLBERT trained on Conceptual Captions (ViLBERT CC) [33], Visual BERT trained on COCO dataset (Visual BERT COCO) [34]. The codes of baselines were published on the website: https://github.com/facebookresearch/mmf/tree/main/projects/hateful_memes.

4. Results

The results are shown in **Table 1**. Our proposed OSCAR+ with random forest classifier (OSCAR+ RF) performed better than all other baselines. Compared to OSCAR+ with a simple linear classifier (OSCAR+ LC), OSCAR+RF also achieved both higher accuracy and AUROC (Area Under the Receiver Operating Characteristic) in the validation set (dev set) and test set. Besides, we observe that the text-only classifier performs slightly better than the vision-only classifier, and multimodal PTMs performed much better than the unimodal models.

Table 1. Comparisons among OSCAR+RF with other published baselines.

Model	Pre-trained modalities	Validation Acc. (%)	Validation AUROC (%)	Test Acc. (%)	Test AUROC (%)
Image-Grid[17]	Image	50.67	52.33	52.73	53.71
Image-Region[17]	Image	52.53	57.24	52.36	57.74
Text BERT[17]	Text	58.27	65.05	62.80	69.00
MMBT-Grid[17]	Image&Text	59.59	66.73	62.83	69.49
MMBT-Region[17]	Image&Text	64.75	72.62	67.66	73.82
ViLBERT[17]	Image	64.75	72.62	67.66	73.82
Visual BERT[17]	Image	65.01	74.14	66.67	74.42
ViLBERT CC[17]	Image&Text	61.40	70.07	61.10	70.00
Visual BERT COCO[17]	Image&Text	65.93	74.14	69.47	75.44
OSCAR+FCNN	Image&Tag&Text	66.32	75.37	68.52	75.56
OSCAR+RF	Image&Tag&Text	**66.58**	**76.83**	**68.73**	**77.14**

Footnotes: Acc., accuracy; AUROC, area under the receiver operating characteristic.

5. Discussion

In this study, we show that the learning of cross-modal representations can be significantly improved by introducing object tags detected in images as anchor points to ease the understanding of semantic alignments between images and texts. Our results demonstrate that our multimodal PTMs by intaking object tags are better than unimodal and multimodal models without anchor points in previous studies.

Traditionally, researchers mainly focused on detecting hateful textual messages by applying deep-learning methods such as FCNN and CNN due to their better performance than traditional machine learning methods [10, 16]. However, their detection accuracy was restricted due to their consideration of only one modality because they could not understand the subtlety behind the memes. Our study shows that the unimodal CNNs (Image-Grid & Image-Region) demonstrate the lowest accuracy and AUROC. On the other hand, we can find adding multimodal learning of both text and image brings a pronounced increase of the detection accuracy in these two models (MMBT-Grid & MMBT-Region): about 10 percent. It indicates the importance of multimodal learning in hateful meme detection.

Even so, some hateful memes employ a lot of background knowledge which makes the relations between visual and textual elements they used very complex and diverse, and therefore difficult to

learn by a traditional neural network [21, 35, 36]. Just like mentioned above, many sentences or images that are harmless by themselves may become hateful when combined; or changing the image part in the meme could quickly reverse the hateful intention. However, with the advent of Large-scale Pre-trained Models (PTM), a better detection model with a better multimodal fusion capability of text and images emerged. Thanks to the immensity of training data (for BERT, the pre-training corpus contains 3,300 million words) and the massive number of model parameters (the base version of BERT has 110 million parameters while the large version of BERT contains 340 million parameters), PTMs demonstrated higher potentials in complex learning because of the rich knowledge reserve in the massive amount of model parameters, some of which have even surpassed human performance on multiple language understanding benchmarks, such as GLUE [18, 30, 37-39]. Not surprisingly, our result shows most PTMs showed higher accuracy than traditional CNNs, especially the PTMs with multimodal learning (ViLBERT CC and Visual BERT COCO). However, these models cannot classify the different image region features well because of the oversampling among the ROIs with overlapping.

By intaking object tags as anchor points, PTM can achieve the highest accuracy and AUROC than conventional Vision-Language PTM and CNNs. This benefit is because the hateful memes generally involve visual and textual cues that could only be identified when considering them simultaneously. The explicit representation of object tags provides the model with clues of the features needing more attention. Besides, in order to achieve better visual embeddings, an object detection (OD) module VinVl (Visual features in Vision-Language) was used to encode the image part in our study. Because VinVl can encode a more diverse collection of visual objects and concepts than typical OD models, it can extract much richer semantics, richer visual concepts, and attribute information [25].

At last, a random forest classifier was connected to the feature vectors produced by OSCAR+ to classify if a given meme is hateful or not. After fine-tuning, specifically, the random forest classifier had better performance than the linear classifier because it had a more remarkable degree of freedom to separate two groups of data points in high-dimensional representation space. In contrast, a linear classifier simply divides the representation space into two sections. However, some of the models' classification false results were observed (**Figure-E1** and **E2**). We found that the memes that invoked simultaneously the visual and textual cues that complemented each other were challenging to be classified correctly. And we also found that the memes that involve objects with specific external knowledge, such as the symbols of ethnic groups, nations, and religions, can also pose a challenge for accurate detection.

Our detection model also has some limitations. First, hateful messages are evolving quickly, so the model cannot keep its detection accuracy if not re-trained in time. For example, modern online communication heavily employs non-standard features, such as emojis and other irregular tokens such as $; and hateful users often try to evade detection by substituting the characters in their messages with very symbols which are very different in terms of machine encodings yet look or sound similar to human beings. One future improvement for the hateful message detection system is to take advantage of these underutilized visual or audio aspects of the textual information in order to include more real-life scenarios. Second, the detection accuracy of our model, or the same kind of our model, largely depends on the knowledge base of the PTM that we used. As we showed that detecting underlying hateful metaphors requires the system to possess the ability to relate visual and linguistic entities in the image or captions to the real-world knowledge base, we expect future Vision-Language PTMs that are supplemented with external knowledge bases, such as Dbpedia [40] and wikidata [41] to achieve better performance on this task. Third, a limited capacity of object recognition in our model. Most contemporary pre-trained visual language models only take image and text into account, and the training dataset contains only common objects. The hateful memes often invoke unusual and specific objects connected to historical or social events not presented in those training sets. Thus, a sufficient dataset related to social or historical events is also in great demand.

5. Conclusion

This study has demonstrated that Vision-Language PTMs with the addition of anchor points can improve the detection of hateful memes that involve strongly correlated textual and visual information. As a result, our proposed model shows the best detection performance compared to previous unimodal and multimodal baselines.

Data Availability Statement

Publicly available datasets were analyzed in this study. This data can be found here: https://www.kaggle.com/parthplc/facebook-hateful-meme-dataset. The code of our model can be found here: https://github.com/SPQRXVIII/hateful_meme_detection.

Author Contributions

Conception and design of study: YC and FP. Acquisition, and analysis of data: YC and FP. Drafting manuscript or figures: YC and FP.

Conflict of Interest

All co-authors declare there is no conflict of interest or financial support of this study.

Acknowledgments

I would like to express my sincere gratitude to Prof. Xingang Wang at School of Electronic Information and Communications (EIC), Huazhong University of Science and Technology for many professional suggestions on the artificial intelligence model establishment.

References

1. Fasoli, F., A. Maass, and A. Carnaghi, *Labelling and discrimination: Do homophobic epithets undermine fair distribution of resources?* British Journal of Social Psychology, 2015. **54**(2): p. 383-393.
2. Greenberg, J. and T. Pyszczynski, *The effect of an overheard ethnic slur on evaluations of the target: How to spread a social disease.* Journal of Experimental Social Psychology, 1985. **21**(1): p. 61-72.
3. Mullen, B. and D.R. Rice, *Ethnophaulisms and exclusion: the behavioral consequences of cognitive representation of ethnic immigrant groups.* Pers Soc Psychol Bull, 2003. **29**(8): p. 1056-67.

4. Mullen, B. and J.M. Smyth, *Immigrant suicide rates as a function of ethnophaulisms: Hate speech predicts death.* Psychosomatic Medicine, 2004. **66**(3): p. 343-348.
5. Soral, W., M. Bilewicz, and M. Winiewski, *Exposure to hate speech increases prejudice through desensitization.* Aggressive behavior, 2018. **44**(2): p. 136-146.
6. Williams, M.L., et al., *Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime.* The British Journal of Criminology, 2020. **60**(1): p. 93-117.
7. Schieb, C. and M. Preuss. *Governing hate speech by means of counterspeech on Facebook*. in *66th ica annual conference, at fukuoka, japan*. 2016.
8. Konikoff, D., *Gatekeepers of toxicity: Reconceptualizing Twitter's abuse and hate speech policies.* Policy & Internet, 2021.
9. Newton, C., *The secret lives of Facebook moderators in America.* The Verge, 2019. **25**.
10. Badjatiya, P., et al. *Deep learning for hate speech detection in tweets*. in *Proceedings of the 26th international conference on World Wide Web companion*. 2017.
11. Gitari, N.D., et al., *A lexicon-based approach for hate speech detection.* International Journal of Multimedia and Ubiquitous Engineering, 2015. **10**(4): p. 215-230.
12. Mehdad, Y. and J. Tetreault. *Do characters abuse more than words?* in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2016.
13. Nobata, C., et al. *Abusive language detection in online user content*. in *Proceedings of the 25th international conference on world wide web*. 2016.
14. Rodriguez, A., C. Argueta, and Y.-L. Chen. *Automatic detection of hate speech on facebook using sentiment and emotion analysis*. in *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. 2019. IEEE.
15. Yin, D., et al., *Detection of harassment on web 2.0.* Proceedings of the Content Analysis in the WEB, 2009. **2**: p. 1-7.
16. Zhang, Z., D. Robinson, and J. Tepper. *Detecting hate speech on twitter using a convolution-gru based deep neural network*. in *European semantic web conference*. 2018. Springer.
17. Kiela, D., et al., *The hateful memes challenge: Detecting hate speech in multimodal memes.* Advances in Neural Information Processing Systems, 2020. **33**: p. 2611-2624.
18. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.
19. Lu, J., et al., *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.* Advances in neural information processing systems, 2019. **32**.
20. Su, W., et al. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. in *International Conference on Learning Representations*. 2019.
21. Antol, S., et al. *Vqa: Visual question answering*. in *Proceedings of the IEEE international conference on computer vision*. 2015.
22. Zellers, R., et al. *From recognition to cognition: Visual commonsense reasoning*. in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
23. Anderson, P., et al. *Bottom-up and top-down attention for image captioning and visual question answering*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
24. Li, X., et al. *Oscar: Object-semantics aligned pre-training for vision-language tasks*. in *European Conference on Computer Vision*. 2020. Springer.
25. Zhang, P., et al. *Vinvl: Revisiting visual representations in vision-language models*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
26. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.
27. Ren, S., et al., *Faster r-cnn: Towards real-time object detection with region proposal networks.* Advances in neural information processing systems, 2015. **28**.

28. Xie, S., et al. *Aggregated residual transformations for deep neural networks*. in *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
29. Mithe, R., S. Indalkar, and N. Divekar, *Optical character recognition.* International journal of recent technology and engineering (IJRTE), 2013. **2**(1): p. 72-75.
30. He, P., et al. *Deberta: Decoding-enhanced bert with disentangled attention*. in *International Conference on Learning Representations*. 2020.
31. Li, L.H., et al., *Visualbert: A simple and performant baseline for vision and language.* arXiv preprint arXiv:1908.03557, 2019.
32. Kiela, D., et al., *Supervised multimodal bitransformers for classifying images and text.* arXiv preprint arXiv:1909.02950, 2019.
33. Sharma, P., et al. *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)*. 2018.
34. Lin, T.-Y., et al. *Microsoft coco: Common objects in context*. in *European conference on computer vision*. 2014. Springer.
35. Specia, L., et al. *A shared task on multimodal machine translation and crosslingual image description*. in *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*. 2016.
36. Mogadala, A., M. Kalimuthu, and D. Klakow, *Trends in integration of vision and language research: A survey of tasks, datasets, and methods.* Journal of Artificial Intelligence Research, 2021. **71**: p. 1183-1317.
37. Liu, Y., et al., *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692, 2019.
38. Raffel, C., et al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research, 2020. **21**: p. 1-67.
39. Wang, A., et al. *Glue: A multi-task benchmark and analysis platform for natural language understanding*. in *7th International Conference on Learning Representations, ICLR 2019*. 2019.
40. Auer, S., et al., *Dbpedia: A nucleus for a web of open data*, in *The semantic web*. 2007, Springer. p. 722-735.
41. Vrandečić, D. and M. Krötzsch, *Wikidata: a free collaborative knowledgebase.* Communications of the ACM, 2014. **57**(10): p. 78-85.

Electronic Supplementary

S1. Definitions of hateful message (or hate speech)

Defining hateful hate messages (or hate speech) is the foundation of our work, yet there are a few candidates competing for this task; to even further obscure our challenge, there is a neighboring category of expression, namely offensive language, that is very similar to hateful messages but being considered more acceptable online. For example, "All Chinese should get out of this place!" is a hateful message while "Get out of this place!" is simply an angry announcement. Therefore, we may first review the definitions from some prominent organizations or individuals to determine the meaning of hateful messages that is distinguishable from that of offensive language:

1. **YouTube, "Hate Speech Policy: YouTube Community Guidelines"**: Hate Speech is not allowed on YouTube. We remove content promoting violence or hatred against members of protected groups including but not limited to 'race, gender, sexual orientation or religious affiliation. We may allow content that includes hate speech like news coverage of world events if their primary purpose is educational, documentary, scientific and artistic in nature.

2. **Facebook, "Community Standard, III. Objectionable Content, Hate Speech"**: We define hate speech as a direct attack against people on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and serious disease. We define attacks as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation. We consider age a protected characteristic when referenced along with another protected characteristic. We also protect refugees, migrants, immigrants and asylum seekers from the most severe attacks, though we do allow commentary and criticism of immigration policies. Similarly, we provide some protections for characteristics such as occupation, when they're referenced along with a protected characteristic. (https://www.facebook.com/communitystandards/hatespeech))

3. **Twitter, "Rules and Policies, Hateful Conduct Policy"**: You may not promote violence against or directly attack or threaten other people based on race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others based on these categories.

4. **Encyclopedia of American Constitution**: "Hate speech is speech that attacks a person or group based on attributes such as race, religion, ethnic origin, national origin, sex, disability, sexual orientation, or gender identity.

5. **Fortuna et al.**: Hate speech is the language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humor is used.

6. **Davidson et al.**: we define hate speech as the language that is used to expresses hatred towards a targeted group or is intended to be derogatory, to humiliate, or to insult the members of the group.

The commonalities of hate speech in these definitions, are their expressed "hatred", "violence", "attack", "threaten", "diminish", "humiliation", "insults" and so on, against "members of protected groups". And these "protected characteristics" include but not limited to 'race', 'gender', 'sexual

orientation' and 'religious affiliation'. One may refer to section 3.1 for further specifications and examples.

Therefore, we summarize hate speech as such: hate speech is a statement that explicitly or implicitly expresses hatred or violence against people with protected characteristics. This definition distinguishes hate speech from an offensive by their targets: though offensive language can be directed at either individuals or groups, it does not target them due to their protected characteristics. However, this definition alone is not sufficient, as praising can also be a hate speech, i.e., praising KKK or Nazis.

Another challenge for this kind of definition arises when annotating the data: identifying and agreeing whether a specific text is hateful is difficult when annotators are subjective and definitions cannot contain that subjectivity; Ross, et al. studied the reliability of hate speech annotations and suggest that annotators are unreliable. Agreement between annotators, measured using Krippendorff's α, was very low (up to 0.29). However, they also pointed out that the low score is due to their definition not being specific enough. In the following section, we provide further specifications and examples that enable better understandings of hate speech and data labeling.

Twitter, YouTube and Facebook have already provided ample specifications and examples of hate speech on their respective community guidelines web pages; though some of those specifications of hate speech are rather ambiguous and controversial, many of them are clear, illuminating and have already become consensus. Therefore, we integrate those guidelines together to form a more comprehensive and useful instruction that specifies hate speech. Hate speech can be first divided into two general categories: a) hate speech that targets individuals or groups because of their protected characteristics, b) hate speech that doesn't specify the characteristics of its targets. Next, hate speech can be divided into more subcategories. For the sake of space, the detailed specification is moved to the appendix section. Please see the appendix section for further details.

However, hateful messages should not be confused with profane and offensive messages. Here are the explanations and examples of the aforementioned terms:

1. **Norma**l: When a speech is not considered hate speech, offensive speech, or profane speech, it is normal.

2. **Profanity**: Profanity is a socially offensive use of language, which may also be called cursing, cussing or swearing, cuss words (American English vernacular), curse words, swear words, bad words, dirty words, or expletives. Accordingly, profanity is language use that is sometimes deemed impolite, rude, or culturally offensive. It can show a debasement of someone or something, or be considered an expression of strong feeling towards something. Some words may also be used as intensifiers. Five possible functions of profanity include:

   (a) Abusive swearing, intended to offend, intimidate, or otherwise cause emotional or psychological harm; e.g., Go to hell, you damn bast**d!

   (b) Cathartic swearing, used in response to pain or misfortune; e.g., I failed my exam, I am f***ed up.

   (c) Dysphemistic swearing, used to convey that the speaker thinks negatively of the subject matter and to make the listener do the same; e.g., N***ers are stupid.

   (d) Emphatic swearing, intended to draw additional attention to what is considered to be worth paying attention to; e.g., This show is damn f***ing cool!

   (e) Idiomatic swearing, used for no other particular purpose, but as a sign that the conversation and relationship between speaker and listener is informal; e.g., These are my ni***.

3. **Offensive**: Offensive language is similar to hate speech; it acts against the target; yet it doesn't contain strongly malignant intent against the target; e.g., "Get out of this place!" "F*** off!"

(Refer to website: openprofanitylist.com/Download/List for more examples.)

Hate speech promotes violence or hatred against individuals or groups based on any of the following **target classes** of protected characteristics:

1. **Age**
2. **Caste**
3. **Disability**
4. **Ethnicity**
5. **Gender Identity and Expression**
6. **Nationality**
7. **Race**

8. **Immigration Status**
9. **Religion**
10. **Sex/Gender**
11. **Sexual Orientation**
12. **Victims of a major violent event and their kin**
13. **Veteran Status**
14. **No specific target**

Hate speech against the members of these groups contains the following categories of actions:

1. **Violent Threats**

Hate speech that contains violent threats against an identifiable target. Violent threats are declarative statements of intent to inflict injuries that would result in serious and lasting bodily harm, where an individual could die or be significantly injured, e.g., "I will kill you." Note that threats such as "Get off or I would kick your ass!" would not be considered hate speech, though they are offensive. At the same time, this definition applies whether or not the target is with protected categories.

2. **Calling Serious Harms**

Hate speech contains wishing, hoping or calling for serious harm on a person or group of people. This includes, but is not limited to:

(a) Hoping that an entire protected category and/or individuals who may be members of that category die as a result of a serious disease, e.g., "I hope all [nationality] get COVID and die."

(b) Wishing for someone to fall victim to a serious accident, e.g., "I wish that you would get run over by a car next time you run your mouth."

(c) Saying that a group of individuals deserves serious physical injury, e.g., "If this group of [slur] don't shut up, they deserve to be shot."

(d) Encouraging others to commit violence against an individual or a group based on their perceived membership in a protected category, e.g., "I'm in the mood to punch a [racial slur], who's with me?"

3. **Hateful Slurs as Degradation** Hate speech that targets individuals or groups with repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category. This includes repeated and/or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone.

4. **Incitement Against Protected Categories** Hate speech that intends to:

(a) incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities, e.g., “all [religious group] are terrorists.”

(b) to incite others to harass members of a protected category on or off the platform, e.g., “I’m sick of these [religious group] thinking they are better than us, if any of you see someone wearing a [religious symbol of the religious group], grab it off them and post pics!”

(c) to incite others to discriminate in the form of denial of support to the economic enterprise of an individual or group because of their perceived membership in a protected category, e.g., “If you go to a [religious group] store, you are supporting those [slur], let’s stop giving our money to these [religious slur].”

5. **Dehumanilzation**

Hate speech that degrades the members of these groups in the form of comparisons, generalizations or unqualified behavioral statements to or about:

(a) Insects

(b) Animals that are culturally perceived as intellectually or physically inferior

(c) Filth, bacteria, disease and feces

(d) Inferior humans

(e) Sexual predators, thieves, bank robbers, and other criminals Specific examples include:

(a) Black people and apes or ape-like creatures

(b) Black people and farm equipment

(c) Caricatures of Black people in the form of blackface

(d) Jewish people and rats

(e) Jewish people running the world or controlling major institutions such as media networks, the economy or the government

(f) Denying or distorting information about the Holocaust

(g) Muslim people and pigs

(h) Muslim person and sexual relations with goats or pigs

(i) Mexican people and worm-like creatures

(j) Women as household objects or referring to women as property or "objects"

(k) Transgender or non-binary people referred to as "it"

(l) Dalits, scheduled caste or "lower caste" people referred to as menial laborers

6. **Denial of existence**

Hate speech that asserts the non-existence of the individual or groups with the aforementioned classes of protected characteristics, or an event that hurts some groups of people is non-existent

7. **Insulting Victims**

Hate speech that mocks, insults, harass, or deny the existence of, the concept, events or victims of mass murder, violent events, or specific means of violence

8. **Exaggerating Inferiority**

Hate speech that exaggerates physical deficiencies, mental deficiencies and moral deficiencies of the protected groups by degrading them with derogatory words:

(a) Derogatory terms related to sexual activity, including, but not limited to: whore, slut, perverts

(b) Expressions about being less than adequate, including, but not limited to: worthless, useless

(c) Expressions about being better/worse than another protected characteristic, including, but not limited to: "I believe that males are superior to females."

(d) Expressions about deviating from the norm, including, but not limited to: freaks, abnormal

9. **Support of Hate Crime** Hate speech includes the following types of actions that target the protected groups:

   (a) Supporting groups that commit hate crimes, including but not limited to: supporting Nazi, supporting genocides

   (b) Supporting segregation in the form of calls for action, statements of intent, aspirational or conditional statements

   (c) Supporting explicit economical, social and political exclusion, by denying access to economic entitlements and limiting participation in the labor market, denying access to physical and online spaces and social services, and denying the right to political participation

   (d) Promoting hateful logos, symbols, imageries and sayings of hateful groups

   (e) Self-admission to intolerance on the basis of protected characteristics, including, but not limited to: homophobic, islamophobic, racist

   (f) Expressions that a protected characteristic shouldn't exist

   (g) Expressions of hate, including, but not limited to: despise, hate

   (h) Expressions of dismissal, including, but not limited to: don't respect, don't like, don't care for

   (i) Expressions suggesting that the target causes sickness, including, but not limited to: vomit, throw up

   (j) Expressions of repulsion or distaste, including, but not limited to: vile, disgusting, yuck

1.1.1 what are considered offensive languages instead of hate speech

1. Referring to the target as genitalia or anus, including, but not limited to: cunt, dick, asshole

2. Profane terms or phrases with the intent to insult, including, but not limited to: fuck, bitch, motherfucker

3. Terms or phrases calling for engagement in sexual activity, or contact with the genitalia, anus, feces or urine, including but not limited to: suck my dick, kiss my ass, eat shit

4. Content that describes or negatively targets people with slurs, where slurs are defined as words that are inherently offensive and used as insulting labels for the above-listed characteristics.

Figure E1. An exemplary illustration of a false positive identification.

Figure E2. An exemplary illustration of a false negative identification.