Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset

Hannah Rose Kirk[1†‡], **Yennie Jun**[1†], **Paulius Rauba**[1†], **Gal Wachtel**[1†], **Ruining Li**[2†], **Xingjian Bai**[2†], **Noah Broestl**[3†], **Martin Doff-Sotta**[4†], **Aleksandar Shtedritski**[4†], **Yuki M. Asano**[4†]

[1]Oxford Internet Institute, [2]Dept. of Computer Science, [3]Oxford Uehiro Centre for Practical Ethics
[4]Dept. of Engineering Science, [†] Oxford Artificial Intelligence Society
[‡]hannah.kirk@oii.ox.ac.uk

Abstract

Hateful memes pose a unique challenge for current machine learning systems because their message is derived from both text- and visual-modalities. To this effect, Facebook released the Hateful Memes Challenge, a dataset of memes with pre-extracted text captions, but it is unclear whether these synthetic examples generalize to 'memes in the wild'. In this paper, we collect hateful and non-hateful memes from Pinterest to evaluate out-of-sample performance on models pre-trained on the Facebook dataset. We find that memes in the wild differ in two key aspects: 1) Captions must be extracted via OCR, injecting noise and diminishing performance of multimodal models, and 2) Memes are more diverse than 'traditional memes', including screenshots of conversations or text on a plain background. This paper thus serves as a reality check for the current benchmark of hateful meme detection and its applicability for detecting real world hate.

1 Introduction

Hate speech is becoming increasingly difficult to monitor due to an increase in volume and diversification of type (MacAvaney et al., 2019). To facilitate the development of multimodal hate detection algorithms, Facebook introduced the Hateful Memes Challenge, a dataset synthetically constructed by pairing text and images (Kiela et al., 2020). Crucially, a meme's hatefulness is determined by the combined meaning of image and text. The question of likeness between synthetically created content and naturally occurring memes is both an ethical and technical one: Any features of this benchmark dataset which are not representative of reality will result in models potentially overfitting to 'clean' memes and generalizing poorly to memes in the wild. Thus, we ask the question: How well do Facebook's synthetic examples (`FB`) represent memes found in the real world? We use Pinterest memes (`Pin`) as our example of memes in the wild and explore differences across three aspects:

1. **OCR.** While `FB` memes have their text pre-extracted, memes in the wild do not. Therefore, we test the performance of several Optical Character Recognition (OCR) algorithms on `Pin` and `FB` memes.
2. **Text content**. To compare text modality content, we examine the most frequent n-grams and train a classifier to predict a meme's dataset membership based on its text.
3. **Image content and style**. To compare image modality, we evaluate meme types (traditional memes, text, screenshots) and attributes contained within memes (number of faces and estimated demographic characteristics).

After characterizing these differences, we evaluate a number of unimodal and multimodal hate classifiers pre-trained on `FB` memes to assess how well they generalize to memes in the wild.

2 Background

The majority of hate speech research focuses on text, mostly from Twitter (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018; Zampieri et al., 2019). Text-based studies face challenges such as distinguishing hate speech from offensive speech (Davidson et al., 2017) and counter speech (Mathew et al., 2018), as well as avoiding racial bias (Sap et al., 2019). Some studies focus on *multimodal* forms of hate, such as sexist advertisements (Gasparini et al., 2018), YouTube videos (Poria et al., 2016), and memes (Suryawanshi et al., 2020; Zhou and Chen, 2020; Das et al., 2020).

While the Hateful Memes Challenge (Kiela et al., 2020) encouraged innovative research on multimodal hate, many of the solutions may not generalize to detecting hateful memes at large. For

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 26–35
August 6, 2021. ©2021 Association for Computational Linguistics

example, the winning team Zhong (2020) exploits a simple statistical bias resulting from the dataset generation process. While the original dataset has since been re-annotated with fine-grained labels regarding the target and type of hate (Nie et al., 2021), this paper focuses on the binary distinction of hate and non-hate.

3 Methods

3.1 Pinterest Data Collection Process

Pinterest is a social media site which groups images into collections based on similar themes. The search function returns images based on user-defined descriptions and tags. Therefore, we collect memes from Pinterest[1] using keyword search terms as noisy labels for whether the returned images are likely hateful or non-hateful (see Appendix A). For hate, we sample based on two heuristics: synonyms of hatefulness or specific hate directed towards protected groups (e.g., 'offensive memes', 'sexist memes') and slurs associated with these types of hate (e.g., 'sl*t memes', 'wh*ore memes'). For non-hate, we again draw on two heuristics: positive sentiment words (e.g., 'funny', 'wholesome', 'cute') and memes relating to entities excluded from the definition of hate speech because they are not a protected category (e.g., 'food', 'maths'). Memes are collected between March 13 and April 1, 2021. We drop duplicate memes, leaving 2,840 images, of which 37% belong to the hateful category.

3.2 Extracting Text- and Image-Modalities (OCR)

We evaluate the following OCR algorithms on the Pin and FB datasets: Tesseract (Smith, 2007), EasyOCR (Jaded AI) and East (Zhou et al., 2017). Previous research has shown the importance of prefiltering images before applying OCR algorithms (Bieniecki et al., 2007). Therefore, we consider two prefiltering methods fine-tuned to the specific characteristics of each dataset (see Appendix B).

3.3 Unimodal Text Differences

After OCR text extraction, we retain words with a probability of correct identification ≥ 0.5, and remove stopwords. A text-based classification task using a unigram Naïve-Bayes model is employed to discriminate between hateful and non-hateful memes of both Pin and FB datasets.

3.4 Unimodal Image Differences

To investigate the distribution of *types* of memes, we train a linear classifier on image features from the penultimate layer of CLIP (see Appendix C) (Radford et al., 2021). From the 100 manually examined Pin memes, we find three broad categories: 1) traditional memes; 2) memes consisting of just text; and 3) screenshots. Examples of each are shown in Appendix C. Further, to detect (potentially several) human faces contained within memes and their relationship with hatefulness, we use a pre-trained FaceNet model (Schroff et al., 2015) to locate faces and apply a pre-trained DEX model (Rothe et al., 2015) to estimate their ages, genders, races. We compare the distributions of these features between the hateful/non-hateful samples.

We note that these models are controversial and may suffer from algorithmic bias due to differential accuracy rates for detecting various subgroups. Alvi et al. (2018) show DEX contains erroneous age information, and Terhorst et al. (2021) show that FaceNet has lower recognition rates for female faces compared to male faces. These are larger issues discussed within the computer vision community (Buolamwini and Gebru, 2018).

3.5 Comparison Across Baseline Models

To examine the consequences of differences between the FB and Pin datasets, we conduct a preliminary classification of memes into hate and non-hate using benchmark models. First, we take a subsample of the Pin dataset to match Facebook's dev dataset, which contains 540 memes, of which 37% are hateful. We compare performance across three samples: (1) FB memes with 'ground truth' text and labels; (2) FB memes with Tesseract OCR text and ground truth labels; and (3) Pin memes with Tesseract OCR text and noisy labels. Next, we select several baseline models pretrained on FB memes[2], provided in the original Hateful Memes challenge (Kiela et al., 2020). Of the 11 pretrained baseline models, we evaluate the performance of five that do not require further preprocessing: Concat Bert, Late Fusion, MMBT-Grid, Unimodal Image, and Unimodal Text. We note that these models are not fine-tuned on

[1] We use an open-sourced Pinterest scraper, available at https://github.com/iamatulsingh/pinterest-image-scrap.

[2] These are available for download at https://github.com/facebookresearch/mmf/tree/master/projects/hateful_memes.

Pin memes but simply evaluate their transfer performance. Finally, we make zero-shot predictions using CLIP (Radford et al., 2021), and evaluate a linear model of visual features trained on the FB dataset (see Appendix D).

4 Results

4.1 OCR Performance

Each of the three OCR engines is paired with one of the two prefiltering methods tuned specifically to each dataset, forming a total of six pairs for evaluation. For both datasets, the methods are tested on 100 random images with manually annotated text. For each method, we compute the average cosine similarity of the joint TF-IDF vectors between the labelled and cleaned[3] predicted text, shown in Tab. 1. Tesseract with FB tuning performs best on the FB dataset, while Easy with Pin tuning performs best on the Pin dataset. We evaluate transferability by comparing how a given pair performs on both datasets. **OCR transferability is generally low**, but greater from the FB dataset to the Pin dataset, despite the latter being more general than the former. This may be explained by the fact that the dominant form of Pin memes (i.e. text on a uniform background outside of the image) is not present in the FB dataset, so any method specifically optimized for Pin memes would perform poorly on FB memes.

Table 1: Cosine similarity between predicted text and labelled text for various OCR engines and prefiltering pairs. Best result per dataset is bolded.

	FB	**Pin**	$\|\Delta\|$
Tesseract, FB tuning	**0.70**	0.36	0.34
Tesseract, Pin tuning	0.22	0.58	0.26
Easy, FB tuning	0.53	0.30	0.23
Easy, Pin tuning	0.32	**0.67**	0.35
East, FB tuning	0.36	0.17	0.19
East, Pin tuning	0.05	0.32	0.27

4.2 Unimodal Text Differences

We compare unigrams and bigrams across datasets after removing stop words, numbers, and URLs. The bigrams are topically different (refer to Appendix E). A unigram token-based Naïve-Bayes classifier is trained on both datasets separately to distinguish between hateful and non-hateful classes. The model achieves an accuracy score of 60.7% on FB memes and 68.2% on Pin memes (random guessing is 50%), indicating mildly different text distributions between hate and non-hate. In order to understand the differences between the type of language used in the two datasets, a classifier is trained to discriminate between FB and Pin memes (regardless of whether they are hateful) based on the extracted tokens. The accuracy is 77.4% on a balanced test set. The high classification performance might be explained by the OCR-generated junk text in the Pin memes which can be observed in a t-SNE plot (see Appendix F).

4.3 Unimodal Image Differences

While the FB dataset contains only "traditional memes"[4], we find this definition of 'a meme' to be too narrow: **the Pin memes are more diverse**, containing 15% memes with only text and 7% memes which are screenshots (see Tab. 2).

Table 2: Percentage of each meme type in Pin and FB datasets, extracted by CLIP.

Meme Type	**FB**	**Pin**	$\|\Delta\|$
Traditional meme	95.6%	77.3%	18.3%
Text	1.4%	15.3%	13.9%
Screenshot	3.0%	7.4%	4.4%

Tab. 3 shows the facial recognition results. **We find that Pin memes contain fewer faces than FB memes**, while other demographic factors broadly match. The DEX model identifies similar age distributions by hate and non-hate and by dataset, with an average of 30 and a gender distribution heavily skewed towards male faces (see Appendix G for additional demographics).

Table 3: Facial detection and demographic (gender, age) distributions from pre-trained FaceNet and DEX.

	FB		**Pin**	
metric	Hate	Non-Hate	Hate	Non-Hate
Images w/ Faces	72.8%	71.9%	52.0%	38.8%
Gender (M:F)	84:16	84:16	82:18	88:12
Age	$30.7_{\pm 5.7}$	$31.2_{\pm 6.3}$	$29.4_{\pm 5.5}$	$29.9_{\pm 5.4}$

4.4 Performance of Baseline Models

How well do hate detection pipelines generalize? Tab. 4 shows the F1 scores for the predictions of hate made by each model on the three samples: (1)

[3] The cleaned text is obtained with lower case conversion and punctuation removal.

[4] The misclassifications into other types reflect the accuracy of our classifier.

FB with ground-truth caption, (2) FB with OCR, (3) Pin with OCR.

Table 4: F1 scores for pretrained baseline models on three datasets. Best result per dataset is bolded.

	FB		Pin
Text from:	Ground-truth	OCR	OCR
Multimodal Models			
Concat BERT	0.321	0.278	0.184
Late Fusion	0.499	0.471	0.377
MMBT-Grid	0.396	0.328	0.351
Unimodal Models			
Text BERT	0.408	0.320	0.327
Image-Grid*	0.226	0.226	0.351
CLIP Models			
CLIP$_{\text{Zero-Shot}}$*	0.509	0.509	0.543
CLIP$_{\text{Linear Probe}}$*	**0.556**	**0.556**	**0.569**

* these models do not use any text inputs so F1 scores repeated for ground truth and OCR columns.

Surprisingly, we find that the CLIP$_{\text{Linear Probe}}$ generalizes very well, performing best for all three samples, with superior performance on Pin memes as compared to FB memes. Because CLIP has been pre-trained on around 400M image-text pairs from the Internet, its learned features generalize better to the Pin dataset, even though it was fine-tuned on the FB dataset. Of the multimodal models, Late Fusion performs the best on all three samples. When comparing the performance of Late Fusion on the FB and Pin OCR samples, **we find a significant drop in model performance of 12 percentage points**. The unimodal text model performs significantly better on FB with the ground truth annotations as compared to either sample with OCR extracted text. This may be explained by the 'clean' captions which do not generalize to real-world meme instances without pre-extracted text.

5 Discussion

The key difference in text modalities derives from the efficacy of the OCR extraction, where messier captions result in performance losses in Text BERT classification. This forms a critique of the way in which the Hateful Memes Challenge is constructed, in which researchers are incentivized to rely on the pre-extracted text rather than using OCR; thus, the reported performance overestimates success in the real world. Further, the Challenge defines a meme as 'a traditional meme' but we question whether this definition is too narrow to encompass the diversity of real memes found in the wild, such as screenshots of text conversations.

When comparing the performance of unimodal and multimodal models, we find multimodal models have superior classification capabilities which may be because the combination of multiple modes create meaning beyond the text and image alone (Kruk et al., 2019). For all three multimodal models (Concat BERT, Late Fusion, and MMBT-Grid), the score for FB memes with ground truth captions is higher than that of FB memes with OCR extracted text, which in turn is higher than that of Pin memes. Finally, we note that CLIP's performance, for zero-shot and linear probing, surpasses the other models and is stable across both datasets.

Limitations Despite presenting a preliminary investigation of the generalizability of the FB dataset to memes in the wild, this paper has several limitations. Firstly, the errors introduced by OCR text extraction resulted in 'messy' captions for Pin memes. This may explain why Pin memes could be distinguished from FB memes by a Naïve-Bayes classifier using text alone. However, these errors demonstrate our key conclusion that the pre-extracted captions of FB memes are not representative of the appropriate pipelines which are required for real world hateful meme detection.

Secondly, our Pin dataset relies on noisy labels of hate/non-hate based on keyword searches, but this chosen heuristic may not catch subtler forms of hate. Further, user-defined labels introduce normative value judgements of whether something is 'offensive' versus 'funny', and such judgements may differ from how Facebook's community standards define hate (Facebook, 2021). In future work, we aim to annotate the Pin dataset with multiple manual annotators for greater comparability to the FB dataset. These ground-truth annotations will allow us to pre-train models on Pin memes and also assess transferability to FB memes.

Conclusion We conduct a reality check of the Hateful Memes Challenge. Our results indicate that there are differences between the synthetic Facebook memes and 'in-the-wild' Pinterest memes, both with regards to text and image modalities. Training and testing unimodal text models on Facebook's pre-extracted captions discounts the potential errors introduced by OCR extraction, which is required for real world hateful meme detection. We hope to repeat this work once we have annotations for the Pinterest dataset and to expand the analysis from comparing between the binary categories of hate versus non-hate to include a comparison across different types and targets of hate.

References

M. Alvi, Andrew Zisserman, and C. Nellåker. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *ECCV Workshops*.

Wojciech Bieniecki, Szymon Grabowski, and Wojciech Rozenberg. 2007. Image preprocessing for improving ocr accuracy. In *2007 international conference on perspective technologies and methods in MEMS design*, pages 75–80. IEEE.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Facebook. 2021. Community standards hate speech. https://www.facebook.com/communitystandards/hate_speech. Accessed on 12 June 2021.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Francesca Gasparini, Ilaria Erba, Elisabetta Fersini, and Silvia Corchs. 2018. Multimodal classification of sexist advertisements. In *ICETE (1)*, pages 565–572.

Jaded AI. Easy OCR. https://github.com/JaidedAI/EasyOCR.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *ArXiv*, abs/2005.04790.

Julia Kruk, Jonah Lubin, Karan Sikka, X. Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. *ArXiv*, abs/1904.09073.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Shaoliang Nie, Aida Davani, Lambert Mathias, Douwe Kiela, Zeerak Waseem, Bertie Vidgen, and Vinodkumar Prabhakaran. 2021. Woah shared task fine grained hateful memes classification. https://github.com/facebookresearch/fine_grained_hateful_memes/.

Pinterest. 2021. All about pinterest. https://help.pinterest.com/en-gb/guide/all-about-pinterest. Accessed on 12 June 2021.

Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, 174:50–59.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. Dex: Deep expectation of apparent age from a single image. In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 252–257.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41.

P. Terhorst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, N. Damer, A. Morales, Julian Fierrez, and Arjan Kuijper. 2021. A comprehensive study on face recognition biases beyond demographics. *ArXiv*, abs/2103.01592.

Laurens Van Der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. Technical report.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*.

Xiayu Zhong. 2020. Classification of multimodal hate speech–the winning solution of hateful memes challenge. *arXiv preprint arXiv:2012.01002*.

Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560.

Yi Zhou and Zhenhao Chen. 2020. Multimodal learning for hateful memes detection. *arXiv preprint arXiv:2011.12870*.

A Details on Pinterest Data Collection

Tab. 5 shows the keywords we use to search for memes on Pinterest. The search function returns images based on user-defined tags and descriptions aligning with the search term (Pinterest, 2021). Each keyword search returns several hundred images on the first few pages of results. Note that Pinterest bans searches for 'racist' memes or slurs associated with racial hatred so these could not be collected. We prefer this method of 'noisy' labelling over classifying the memes with existing hate speech classifiers with the text as input because users likely take the multimodal content of the meme into account when adding tags or writing descriptions. However, we recognize that user-defined labelling comes with its own limitations of introducing noise into the dataset from idiosyncratic interpretation of tags. We also recognize that the memes we collect from Pinterest do not represent all Pinterest memes, nor do they represent all memes generally on the Internet. Rather, they reflect a sample of instances. Further, we over-sample non-hateful memes as compared to hateful memes because this distribution is one that is reflected in the real world. For example, the FB dev set is composed of 37% hateful memes. Lastly, while we manually confirm that the noisy labels of 50 hateful and 50 non-hateful memes (see Tab. 6), we also recognize that not all of the images accurately match the associated noisy label, especially for hateful memes which must match the definition of hate speech as directed towards a protected category.

Table 5: Keywords used to produce noisily-labelled samples of hateful and non-hateful memes from Pinterest.

Noisy Label	**Keywords**
Hate	"sexist", "offensive", "vulgar", "wh*re", "sl*t", "prostitute"
Non-Hate	"funny", "wholesome", "happy", "friendship", "cute", "phd", "student", "food", "exercise"

Table 6: Results of manual annotation for noisy labelling. Of 50 random memes with a noisy hate label, we find 80% are indeed hateful, and of 50 random memes with a noisy non-hate label, we find 94% are indeed non-hateful.

	Noisy Hate	**Noisy Non-Hate**
Annotator Hate	40	3
Annotator Non-Hate	10	47

B Details on OCR Engines

B.1 OCR Algorithms

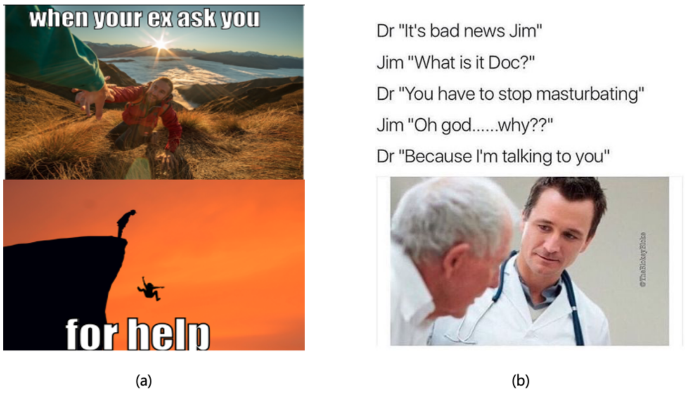
We evaluate three OCR algorithms on the Pin and FB datasets. First, Tesseract (Smith, 2007) is Google's open-source OCR engine. It has been continuously developed and maintained since its first release in 1985 by Hewlett-Packard Laboratories. Second, EasyOCR (Jaded AI) developed by Jaded AI, is the algorithm used by the winner of the Facebook Hateful Meme Challenge. Third, East (Zhou et al., 2017) is an efficient deep learning algorithm for text detection in natural scenes. In this paper East is used to isolate regions of interest in the image in combination with Tesseract for text recognition.

B.2 OCR Pre-filtering

Figure 4 shows the dominant text patterns in FB (a) and Pin (b) datasets, respectively. We use a specific prefiltering adapted to each pattern as follows.

FB Tuning: FB memes always have a black-edged white Impact font. The most efficient prefiltering sequence consists of applying an RGB-to-Gray conversion, followed by binary thresholding, closing, and inversion. **Pin Tuning:** Pin memes are less structured than FB memes, but a commonly observed meme type is text placed outside of the image on a uniform background. For this pattern, the most efficient prefiltering sequence consists of an RGB-to-Gray conversion followed by Otsu's thresholding.

The optimal thresholds used to classify pixels in binary and Otsu's thresholding operations are found so as to maximise the average cosine similarity of the joint TF-IDF vectors between the labelled and predicted text from a sample of 30 annotated images from both datasets.



Dr "It's bad news Jim"
Jim "What is it Doc?"
Dr "You have to stop masturbating"
Jim "Oh god......why??"
Dr "Because I'm talking to you"

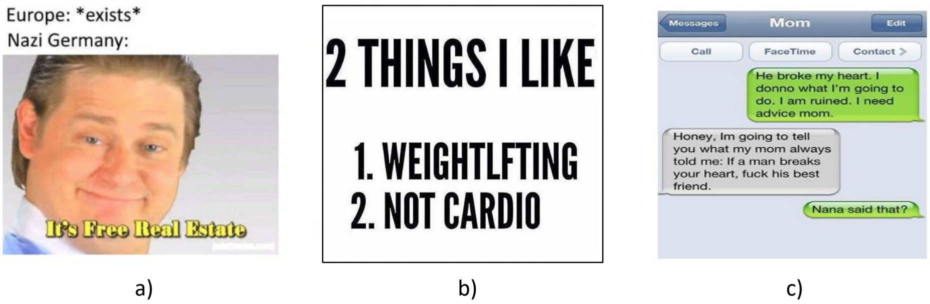
(a) (b)

Figure 1: Dominant text patterns in (a) Facebook dataset (b) Pinterest dataset.

C Classification of Memes into Types

C.1 Data Preparation

To prepare the data needed for training the ternary (i.e., traditional memes, memes purely consisting of text, and screenshots) classifier, we annotate the `Pin` dataset with manual annotations to create a balanced set of 400 images. We split the set randomly, so that 70% is used as the training data and the rest 30% as the validation data. Figure 2 shows the main types of memes encountered. The `FB` dataset only has traditional meme types.



He broke my heart. I donno what I'm going to do. I am ruined. I need advice mom.

Honey, Im going to tell you what my mom always told me: If a man breaks your heart, fuck his best friend.

Nana said that?

a) b) c)

Figure 2: Different types of memes: (a) Traditional meme (b) Text (c) Screenshot.

C.2 Training Process

We use image features taken from the penultimate layer of CLIP. We train a neural network with two hidden layers of 64 and 12 neurons respectively with ReLU activations, using Adam optimizer, for 50 epochs. The model achieves 93.3% accuracy on the validation set.

D Classification Using CLIP

D.1 Zero-shot Classification

To perform zero-shot classification using CLIP (Radford et al., 2021), for every meme we use two prompts, "`a meme`" and "`a hatespeech meme`". We measure the similarity score between the image and text embeddings and use the corresponding text prompt as a label. Note we regard this method as neither multimodal nor uni-modal, as the text is not explicitly given to the model, but as shown in (Radford et al., 2021), CLIP has some OCR capabilities. In a future work we would like to explore how to modify the text prompts to improve performance.

D.2 Linear Probing

We train a binary linear classifier on the image features of CLIP on the FB train set. We train the classifier following the procedure outlined by (Radford et al., 2021). Finally, we evaluate the binary classifier of the FB dev set and the Pin dataset.

In all experiments above we use the pretrained ViT-B/32 model.

E Common Bigrams

The FB and Pin datasets have distinctively different bigrams after data cleaning and the removal of stop words.

The most common bigrams for hateful FB memes are: [‘black people’, ‘white people’, ‘white trash’, ‘black guy’, ‘sh*t brains’, ‘look like’]. The most common bigrams for non-hateful FB memes are: [‘strip club’, ‘isis strip’, ‘meanwhile isis’, ‘white people’, ‘look like’, ‘ilhan omar’]

The most common bigrams for hateful Pin memes are: ‘im saying’, ‘favorite color’, ‘single white’, ‘black panthers’, ‘saying wh*res’, and ‘saying sl*t’. The most common bigrams for non-hateful Pin memes are: ‘best friend’, ‘dad jokes’, ‘teacher new’, ‘black lives’, ‘lives matter’, and ‘let dog’.

F T-SNE Text Embeddings

The meme-level embeddings are calculated by (i) extracting a 300-dimensional embedding for each word in the meme, using fastText embeddings trained on Wikipedia and Common Crawl; (ii) averaging all the embeddings along each dimension. A T-SNE transformation is then applied to the full dataset, reducing it to two-dimensional space. After this reduction, 1000 text-embeddings from each category—FB and Pin — are extracted and visualized. The default perplexity parameter of 50 is used. Fig.3 presents the t-SNE plot (Van Der Maaten and Hinton, 2008), which indicates a concentration of multiple embeddings of the Pin memes within a region at the bottom of the figure. These memes represent those that have nonsensical word tokens from OCR errors.

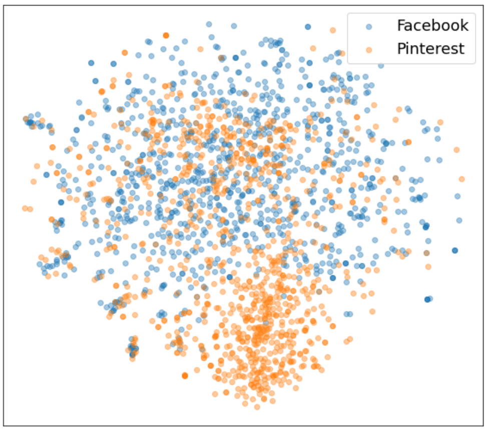


Figure 3: t-SNE of Facebook and Pinterest memes’ text-embeddings for a random sample of 1000 each.

G Face Recognition

G.1 Multi-Faces Detection Method

To evaluate memes with multiple faces, we develop a self-adaptive algorithm to separate faces. For each meme, we enumerate the position of a cutting line (either horizontal or vertical) with fixed granularity, and run facial detection models on both parts separately. If both parts have a high probability of containing faces, we decide that each part has at least one face. Hence, we cut the meme along the line, and run this algorithm iteratively on both parts. If no enumerated cutting line satisfies the condition above, then we decide there’s only one face in the meme and terminate the algorithm.

G.2 Additional Results on Facial Analysis

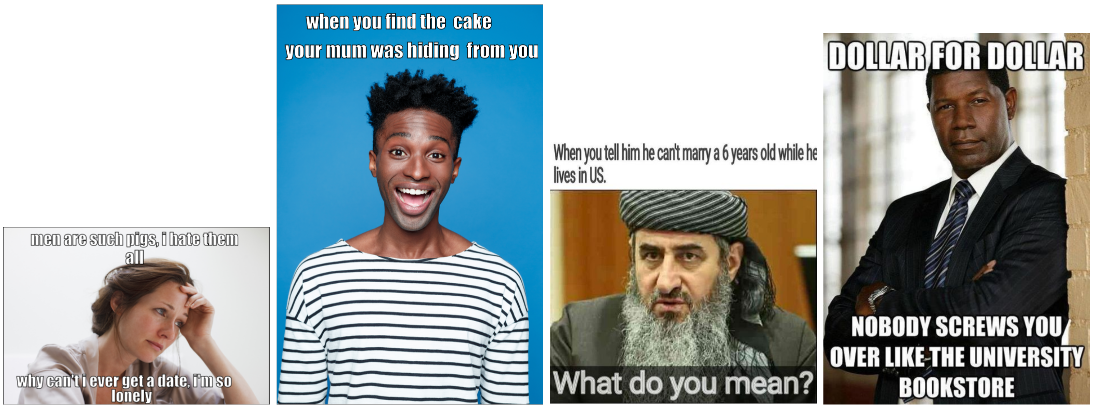
Table 7: Predicted ratio of emotion categories on faces from different datasets from pre-trained DEX model.

	FB		Pin	
categories	Hate	Non-Hate	Hate	Non-Hate
angry	10.6%	10.1%	9.0%	13.7%
disgust	0.3%	0.2%	0.7%	0.6%
fear	9.5%	10.2%	10.6%	13.0%
happy	35.1%	36.3%	34.2%	30.1%
neutral	23.1%	22.7%	23.4%	21.5%
sad	18.8%	18.7%	20.4%	18.6%
surprise	2.2%	1.7%	1.7%	1.8%

Table 8: Predicted ratio of racial categories of faces from different datasets from pre-trained DEX model.

	FB		Pin	
categories	Hate	Non-Hate	Hate	Non-Hate
asian	10.6%	10.8%	9.7%	13.9%
black	15.0%	15.3%	6.5%	11.0%
indian	5.9%	6.1%	3.2%	5.1%
latino hispanic	14.3%	14.5%	10.2%	11.7%
middle eastern	12.7%	11.2%	9.5%	10.1%
white	41.5%	42.1%	60.9%	48.1%

G.3 Examples of Faces in Memes


men are such pigs, i hate them all
why can't i ever get a date, i'm so lonely
when you find the cake
your mum was hiding from you
When you tell him he can't marry a 6 years old while he lives in US.
What do you mean?
DOLLAR FOR DOLLAR
NOBODY SCREWS YOU OVER LIKE THE UNIVERSITY BOOKSTORE

(a) FB Hate (b) FB Non-hate (c) Pin Hate (d) Pin Non-hate

Figure 4: Samples of faces in FB Hate, FB Non-hate, Pin Hate, and Pin Non-hate datasets, and their demographic characteristic predicted by the DEX model:
(a) Woman, 37, white, sad (72.0%); (b) Man, 27, black, happy (99.9%);
(c) Man, 36, middle eastern, angry (52.2%); (d) Man, 29, black, neutral (68.0%)