Feels Bad Man: Dissecting Automated Hateful Meme Detection Through the Lens of Facebook's Challenge

Catherine Jennifer[1], Fatemeh Tahmasbi[2], Jeremy Blackburn[2], Gianluca Stringhini[3], Savvas Zannettou[4], and Emiliano De Cristofaro[1]

[1]UCL [2]Binghamton University [3]Boston University [4]TU Delft

Abstract

Internet memes have become a dominant method of communication; at the same time, however, they are also increasingly being used to advocate extremism and foster derogatory beliefs. Nonetheless, we do not have a firm understanding as to which perceptual aspects of memes cause this phenomenon. In this work, we assess the efficacy of current state-of-the-art multimodal machine learning models toward hateful meme detection, and in particular with respect to their generalizability across platforms. We use two benchmark datasets comprising 12,140 and 10,567 images from 4chan's "Politically Incorrect" board (/pol/) and Facebook's Hateful Memes Challenge dataset to train the competition's top-ranking machine learning models for the discovery of the most prominent features that distinguish viral hateful memes from benign ones. We conduct three experiments to determine the importance of multimodality on classification performance, the influential capacity of fringe Web communities on mainstream social platforms and vice versa, and the models' learning transferability on 4chan memes.

Our experiments show that memes' image characteristics provide a greater wealth of information than its textual content. We also find that current systems developed for online detection of hate speech in memes necessitate further concentration on its visual elements to improve their interpretation of underlying cultural connotations, implying that multimodal models fail to adequately grasp the intricacies of hate speech in memes and generalize across social media platforms.

1 Introduction

Social networking sites have facilitated communication among users worldwide, connecting like-minded individuals who share similar values and enabling the formation of online communities. As methods of discourse on the Web undergo continuous change to enhance the comprehensibility of personal opinions, so does the potential to advocate hateful beliefs. Most notably, Internet memes are used as a way of communicating such concepts in an engaging manner, with the most viral of memes acquiring the most attention and becoming near impossible to moderate (Roose 2021). There is a rising interest in developing approaches to better manage this problem; for instance, in 2020, the Hateful Memes Challenge was launched by Facebook AI (FacebookAI 2020), aiming to improve the detection of hate speech in multimodal memes.

In spite of the efforts made in multimodal Artificial Intelligence (AI), this problem extends beyond the modalities of a meme; rather, there is very little focus on the visual characteristics which make such content attractive enough to obtain a large quantity of resubmissions on social networks. The human brain can interpret an image in a mere 13 milliseconds (Trafton 2014), and image memes have made it convenient for users to quickly comprehend its connotative message. Moreover, the circulation of viral hateful memes on the Internet occurs by the influence of different social platforms on each other (Zannettou et al. 2018). However, the Hateful Memes Challenge introduce a dataset with memes that have been generated *artificially*, whereby, such samples do not accurately capture the characteristics of hateful memes that originate and spread on other platforms, limiting the generalizability of these results.

In this paper, we focus on predicting the dissemination of toxic image memes by running experiments on memes from 4chan's Politically Incorrect Board (/pol/) and Facebook's Hateful Memes Challenge using Vision and Language (V&L) machine learning models to evaluate the potency of multimodal machine learning classification for viral hateful memes. Overall, we identify and tackle the following research questions:

1. How significant is the influence of multimodality in image memes?
2. How portable are models trained on Facebook's challenge memes on other social platforms?
3. What are the characteristics of hateful viral memes?

Methodology. We start with performing three experiments involving four V&L classifiers and using Kiela et al.'s challenge dataset for multimodal classification (Kiela et al. 2021), and a set of hateful memes collected from /pol/ by Zannettou et al. (Zannettou et al. 2018). More specifically:

- We use methods by Velioglu and Rose (Velioglu and Rose 2020) to train a VisualBERT model on Zannettou et al.'s dataset to assess the importance of text in hateful meme images.
- We focus on the portability of Kiela et al.'s samples on other social networks by evaluating the prediction performance of a UNITER model (with the settings from Muennighoff (Muennighoff 2020)) on 4chan memes.
- We use three models – UNITER, OSCAR, and an ensemble classifier – each of which are trained, optimized,

	#Memes	4chan Hateful	4chan Non-Hateful	Facebook Hateful	Facebook Non-Hateful
1A	8,923	3,442	0	0	5,481
1B	8,259	2,778	0	0	5,481
2	10,251	750	1,001	3,019	5,481
3	2,596	1,297	1,299	0	0

Table 1: Summary of datasets.

and tested only on samples from 4chan to evaluate the generalizability of the Hateful Memes Challenge's best learning algorithms.

- We conduct a feature analysis to inspect the visual attributes with the most influential impact on the classification accuracy of classifiers from the first and third experiments to discover indicators of virality.

Findings. Our main findings can be summarized as follows:

1. The visual characteristics of memes offer a plethora of information to effectively communicate the image's intended meaning without the inclusion of text. This is evident from the model's ability to correctly identify hateful memes 80% of the time in both unimodal and multimodal representations.
2. The Hateful Memes Challenge dataset is not adequately representative of multimodal hate speech to support the creation of detection algorithms, as demonstrated by the second experiment, when the classifier is evaluated on samples from /pol/.
3. We find four principal characteristics associated with virality in hateful memes: subject matter, facial expressions, gestures, and proportion. In general, hateful viral memes incorporate two or more of these attributes, which is evident from the capacity of the best classifier across all three experiments to correctly classify 84% of viral memes from 4chan as hateful.

Remarks. In this paper, we use the following definition of *hate* for our investigation: "speech or expression that denigrates a person or persons on the basis of (alleged) membership in a social group identified by attributes such as race, ethnicity, gender, sexual orientation, religion, age, physical, or mental disability, and others," as per (M.Curtis 2016). Also, please be advised that this paper contains uncensored hateful images, which might be disturbing to some readers.

2 Background

2.1 Facebook Hateful Memes Challenge

The Hateful Memes Challenge was launched by Facebook AI to support the development of autonomous systems for the recognition of harmful multimodal content (FacebookAI 2020). For this purpose, Kiela et al. proposed a challenge set comprised of multimodal memes conveying hateful or non-hateful messages, constructed in a fashion which makes it difficult for unimodal classifiers to effectively discriminate between the two classes. In particular, samples containing contradictory meanings through their modalities, i.e., "benign confounders" are included in the dataset such that only multimodal models are able to accurately interpret their communicative signals for better predictability.

2.2 4chan

4chan is an anonymous image-sharing board widely recognized for its user's radical opinions and influence on other social media sites. Particularly members from the /pol/ board that majorly harbor content promoting far-right, misogynistic, and transphobic views which has impacted the information ecosystem and sparked widespread controversy, e.g., the 2016 US presidential elections that flourished the creation of antisemitic memes on /pol/ to advance an agenda of white supremacy (Hine et al. 2017; Zannettou et al. 2020). In fact, many hateful viral memes trace back to 4chan as its source of origin, with small fringe communities such as /pol/ having the potential to spread such content on larger, more mainstream platforms (e.g., Twitter) (Zannettou et al. 2018).

3 Datasets

In this section, we introduce the two benchmark datasets used throughout the experimentation, of which is summarized in Table 1. We focus on 4chan's /pol/ community and Facebook in this study; thus, we use 10,567 images collected by Ling et al. (Zannettou et al. 2018) as a baseline for the analysis of toxic meme virality. This dataset consists of both multimodal and unimodal samples which are either hateful or non-hateful.

Additionally, we use the Hateful Memes Challenge dataset created by Kiela et al.(Kiela et al. 2021) considering that Facebook is the most commonly used networking platform to date (DataReportal 2021) and has the potential to exert social influences on the Web ecosystem at mass, making the spread of memes more prominent on the service. Moreover, Facebook's challenge set comprises 12,140 examples of multimodal hate speech that expresses sociocultural information through is visual modes which makes it suitable for evaluative purposes. In the rest of the paper, the two datasets are divided into four subsets and used for the three experiments as follows:

1. A set of 5,481 multimodal non-hateful images from Facebook merged with 3,442 multimodal hateful images from 4chan, and another set of 5,481 multimodal non-hateful Facebook images merged with 2,778 unimodal hateful 4chan images.
2. A set of 1,001 non-hateful and 750 hateful 4chan images with text.
3. A set of 1,299 non-hateful and 1,297 hateful 4chan images with text.

Pre-Processing. Facebook's hateful memes competition provided their participants with meme images' extracted text in separate JSON Lines (.jsonl) (Crockford 2022) formatted files for training, validation, and testing procedures named as 'train.jsonl', 'dev_unseen.jsonl' and 'test_unseen.jsonl', respectively. Each image file name in Kiela et al.'s dataset

serves as a reference to match the textual and visual modalities of a sample prior to its classification (DrivenData Inc. and Facebook AI 2020), leading to the creation of train, validation, and test .jsonl files adapted specifically for each 4chan sample set. We use the Optical Character Recognition (OCR) Python package known as EasyOCR (LeCun 2022) to extract text from 4chan image memes and include the textual content next to the corresponding file's ID in the .jsonl files.

To ascertain that the extracted text was precisely as depicted in its original meme image, and avoid providing the experimental models with distorted input that would affect its prediction performance, every JSON line in the newly created files was carefully inspected for the manual correction of text arrangement errors or slang words that were not entirely captured by EasyOCR.

4 Experimental Setup

In this section, we describe the procedures undertaken by three classification experiments, and the metrics used to evaluate the performance of each learning model.

4.1 The Effects of Multimodality in Hateful Memes

We consider two cases in the first experiment to obtain a greater insight into the impact of multimodality on model predictability: the importance of text in image memes, and the effect of unimodality in image memes for accurate classification. The focal point of this experiment is to test whether toxic viral 4chan memes that comprise multiple visual modes have enhanced influential potential on Facebook memes as opposed to those comprising a single communication mode, and allow us to further improve our understanding of how moderately sized fringe Web communities influence mainstream platforms. The first case will henceforth be referred as *Multimodal Memes*, and the second case will be referred as *Unimodal Memes*.

Train, Validate, and Test Splits. For *Multimodal Memes*, we benchmark the second-best ranking model of the challenge – a VisualBERT CC majority-vote ensemble (Velioglu and Rose 2020) – on the dataset consisting of 8,923 image memes with text, culminating a balanced data distribution of 3,442 hateful 4chan memes and 5,481 non-hateful Facebook memes to form the training set. For better comparative analysis, we maintain a similar class label distribution to the Hateful Memes Challenge dataset of 85%, 5%, and 10% for training, validation, and testing, respectively. Likewise, *Unimodal Memes* uses the dataset comprising 2,778 hateful 4chan image memes without text and 5,481 non-hateful Facebook images memes with text for training and tuning the VisualBERT classifier with close distribution to Kiela et al.'s train set.

Model implementation. We perform feature extraction using the object detection algorithm known as *Detectron* (Girshick et al. 2018) to capture important patterns in image memes and enhance the learning ability and generalizability of the classifier. Specifically, we use the *Mask RCNN* deep neural network (He et al. 2018) based on the *ResNet-152* architecture (He et al. 2015) to extract features from 100 bounding boxes per image (refer to Figure 1a for an example illustration of a processed sample). We then perform a Hyperparameter Search on multiple VisualBERT CC derivatives to discover a combination of the most optimal parameters for training and select 27 classifiers with the highest ROC-AUC scores on the validation set from (Kiela et al. 2021). Finally, a majority-vote approach is taken to combine each model's estimations on Kiela et al.'s test set and form a single ensemble classifier. Figure 1b depicts an overall visual interpretation of this procedure (note that both sub-experiments undergo the same process).

4.2 Determining the Generalizability of Facebook's Meme Samples

In this experiment, we study the influential potential of hateful memes from mainstream social media platforms on comparatively small Web communities. Specifically, we assess a UNITER model (Muennighoff 2020) on the 4chan test set after being trained on Kiela et al.'s dataset. We chose UNITER given its adequacy for Natural Language Processing (NLP) tasks and taking into account that all samples used for this second experiment contain embedded text.

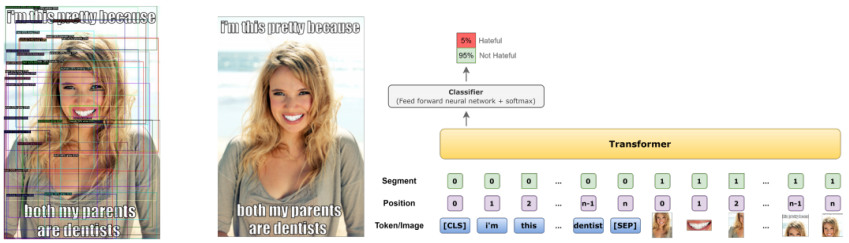
Train, Validate, and Test Splits. All image memes in the 4chan test set for this study are multimodal, with 750 being hateful and 1,001 non-hateful; however, we do not alter the train and validation sets from (Kiela et al. 2021). Considering that the majority of memes distributed on social media are benign, a larger portion of samples in the test set are non-hateful to simulate a realistic depiction of how well the model would perform upon deployment on Web platforms and thus end up with an imbalanced class distribution for testing.

Model implementation. Like in the first experiment, we use Detectron to extract image features from memes, but consider 36 bounded boxes per image instead of 100, since UNITER outperforms VisualBERT (Muennighoff 2020) with fewer parameters (Chen et al. 2020). We fine-tune the classifier on the validation set using a binary cross-entropy loss function to compare its probability predictions against true class labels, and use the Adam optimizer (Kingma and Ba 2014) with the same hyperparameter settings defined in (Ma and Yarats 2021) to train the model for five epochs.

4.3 Evaluating Vision-Language Models on Toxic Viral 4chan Memes

Finally, we evaluate the classification performance of three models, namely UNITER, OSCAR (Muennighoff 2020), and an average-prediction ensemble formed by computing the weighted mean of the model's combined predictions on multimodal 4chan samples to contrast their capacity to distinguish hateful memes from benign ones.

Train, Validate, and Test Splits. We split the dataset for this study into three subsets of which two will be used for training and optimization, and another for testing. To account for the moderate quantity of samples used in this experiment,

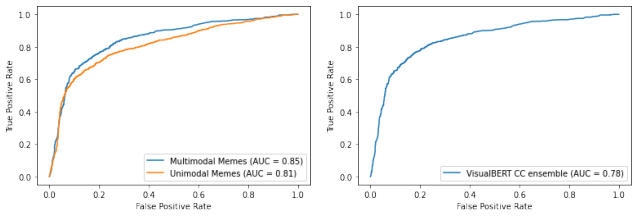


(a) **(b)**

Figure 1: VisualBERT CC ensemble model implementation: (a) an example of Detectron image feature extraction using 36 bounding boxes, (b) an example image meme (left), and a demonstration of the model's architecture and classification process (right) from Velioglu and Rose (Velioglu and Rose 2020).

Classifier	Multimodal Memes				Unimodal Memes			
	Precision	Recall	F1	AUC	Precision	Recall	F1	AUC
VisualBERT CC ensemble	0.66	0.80	0.73	0.85	0.70	0.69	0.69	0.81

Table 2: Results for The Effects of Multimodality in Hateful Memes.



(a) **(b)**

Figure 2: ROC curves for: (a) Multimodal Memes and Unimodal Memes, (b) VisualBERTCC ensemble classifier trained only on Facebook samples.

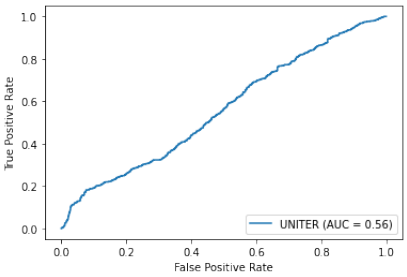


Figure 3: ROC curve for UNITER.

Classifier	Precision	Recall	F1	AUC
UNITER	0.54	0.24	0.33	0.56

Table 3: Results for Determining the Generalizability of Facebook's Meme Samples.

Classifier	Precision	Recall	F1	AUC
UNITER	0.98	0.93	0.95	0.99
OSCAR	0.95	0.95	0.95	0.99
Average-Prediction Ensemble	0.96	0.96	0.96	0.99

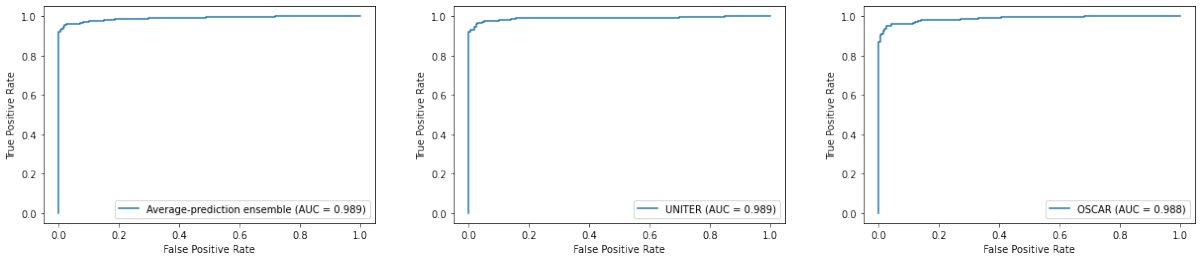
Table 4: Results for Evaluating Vision-Language Models on Toxic Viral 4chan Memes.

we follow a 70:10:20 data split such that precisely 1,997, 199, and 400 image memes with text are used for training, validation, and testing, respectively. Unlike the first and second experiments, this dataset excludes samples containing long text due to the constraint of maximum 512 tokens imposed by transformer models (Devlin et al. 2019).

Model implementation. To introduce diversity in the models' predictions and reduce feature redundancy, we extract feature vectors from various Regions of Interests (RoIs) in image memes by defining different quantities of bounding boxes. Then, the UNITER and OSCAR models undergo the same procedure for training and optimization as UNITER in *The Effects of Multimodality in Hateful Memes* experiment. A third, ensemble classifier is created from the two aforementioned by averaging their individual predictions upon completion of their training.

4.4 Metrics

The four following measures are used to assess the performance of the models: Precision, Recall, F1-Score, and the Area Under the Curve of the Receiver Operating Characteristic (AUC-ROC). Precision and Recall are widely recognized as effective approaches to determine classification performance on imbalanced datasets, with the F1-Score providing a balance between the two measures to dictate an overall outcome of each classifier's estimation quality on unseen data. We also chose the AUC-ROC metric to compare how well the classifiers are able to discriminate between the classes under different test scenarios.



(a) Average-prediction ensemble **(b)** UNITER **(c)** OSCAR

Figure 4: ROC curves for third experiment. AUC values for average-prediction ensemble, UNITER, and OSCAR classifiers.

5 Results

5.1 The Effects of Multimodality in Hateful Memes

Performance results for this experiment are shown in Table 2. We see that the VisualBERT CC classifier has a higher recall than precision after being tested under the conditions of *Multimodal Memes* based on the fact that the training dataset comprised solely of multimodal samples. Moreover, this type of model has been developed specifically for V&L tasks, which has enabled it to generate more truthful estimations on the test set upon learning associations between memes' text and image features. Unsurprisingly, the precision score is greater in *Unimodal Memes* since all samples in Kiela et al.'s dataset contains embedded text causing the classifier to identify a larger quantity of benign memes than hateful ones. A 69% recall rate nevertheless suggests the predictive capacity of VisualBERT CC is above that of a mediocre one.

In Figure 2a, we plot the ROC curves for both sub-experiments. An AUC score of approximately 0.80 across the experiments means the VisualBERT classifier can correctly discriminate between hateful and non-hateful memes 80% of the time, regardless of the modalities involved. *Unimodal Memes* notably reveals that memes do not necessitate text to inflict extremist ideology and increase its potential for online dissemination demonstrating that image characteristics of memes are just as meaningful as those incorporating text. We also plot the ROC curve of the model's classification performance when trained only on Kiela et al.'s training set in Figure 2b, which projects a very close AUC result of 0.78 to that of the sub-experiments, further supporting our findings.

5.2 Determining the Generalizability of Facebook's Meme Samples

We report the results of UNITER's performance in Table 3, which shows poor classification performance given its near-chance AUC score (0.56). The ROC curve in Figure 3 provides a better interpretation of this outcome. Furthermore, the model has a recall rate of 0.24 on the 4chan test set after being trained on Facebook's dataset, indicating low discriminatory ability between the classes. This implies that Kiela et al.'s dataset may not adequately simulate memes shared on social media considering one of the best performing models cannot generalize well to memes from other social platforms.

Figure 3 further shows that UNITER incorrectly labels many hateful memes as non-hateful, suggesting that it is unable to capture the visual features in test samples at a rate sufficient enough to make truthful classifications, resulting in an inadequate true positive rate (TPR) for addressing the challenge of hate speech recognition in multimodal memes.

Evaluating Vision-Language Models on Toxic Viral 4chan Memes. Table 4 shows the results of each classifier, i.e., UNITER, OSCAR, and average-prediction, attain AUCs of 0.989, 0.988, and 0.989, respectively. We observe that the average-prediction ensemble achieves the greatest recall rate compared with the other two. However, UNITER obtains a greater precision (0.979) than the ensemble model meaning it can correctly label memes as hateful approximately 98% of the time. Nevertheless, a higher recall is favorable for this classification task and although the average-prediction classifier has the same AUC (0.989) as UNITER, it delivers the best overall performance. We also see that OSCAR is 0.02% more likely to accurately identify hateful memes than UNITER given its recall of 0.950, but is nonetheless the weakest performing model as this score is impeded by its inferior AUC.

Figures 4a, 4b, and 4c demonstrate ROC curve plots corresponding to UNITER, OSCAR, and the average-prediction model, respectively. Ultimately, UNITER and the average-prediction ensemble have exceptional discriminative ability between both classes with identical potential (considering their AUCs); however, we can discern in Figure 4a that the ensemble classifier achieves a higher TPR making it preferable for this problem.

6 Related Work

Meme propagation. Previous work has focused on measuring and tracing meme dissemination on the Web. Zannettou et al. (Zannettou et al. 2018) introduced a custom metric to measure the visual similarity between image memes to track variants of meme families from polarized Web communities such as 4chan's /pol/, Gab, and The_donald, in an effort to study their impact on meme propagation and analyze the influential correlations between the social networking platforms.

Indicators of viral image memes. Arturo Deza and Devi Parikh (Deza and Parikh 2015) conducted a semantic evaluation of the perceptual cues in viral memes, identifying 5 key attributes that link to virality: 'Animal', 'Synthgen', 'Beautiful', 'Explicit', and 'Sexual' - Each of which elicit different emotional reactions from its viewers and potentially affect their decision to share a post.

Detection of hateful and offensive memes. Kiela et al. (Kiela et al. 2021) introduced a challenge dataset of 10,000 artificially generated multimodal memes representative of real ones publicized on social platforms, and annotated as hateful or non-hateful. Various approaches to this competition have been tried, including the use of early fusion strategy with transformer models to combine the visual elements and textual content of memes prior to their classification (Zhu 2020).

Novelty: The detection of hateful speech in multimodal memes (Kiela et al. 2021) is the most similar work to ours thus far. However, this work is the first to consider the anticipation of such content prior to its publication - viral hateful memes in particular, which become extremely challenging to moderate once posted. We also identify limitations in Kiela et al.'s dataset and the approaches used by winning contestants of the Hateful Memes Challenge.

7 Conclusion

7.1 Limitations

As previously mentioned, hateful viral memes from fringe Web communities such as 4chan's /pol/ also appear in comic strip format (e.g., Tyrone (Know Your Meme 2022)). The training set from our *Evaluating Model Generalization on Toxic Viral 4chan Memes* experiment contains 119 of 1,997 images which comprise of multiple panels depicting a popular meme subject, with almost half of the TPs produced by the average-prediction classifier showing this attribute.

Although seemingly minor, the study of memes composed of more than a single panel is worthy of consideration to try to comprehend how memes such as the Tyrone comic series was successful in gaining mass resubmissions and imitations on social media. Unfortunately, we could not examine this particular element closely enough to deem it as another possible indicator of virality given none of Kiela et al.'s (Kiela et al. 2021) data samples are viral due to their nature of construction. Thus, further investigation in the context of spatial vicinity is necessary to establish whether the presence of viral meme subjects enhance the virality potential of multiple-panel image memes.

Another aspect of this study is the examination of how many panels in a comic meme strip would be too many, provided each panel displays text to illustrate a story, and an abundance of text reduces a meme's online influence (as shown from our experimental outcomes). However, there is currently limited availability of meme datasets for such studies and thus we hope that future work in this area will contribute to their development.

Moreover, careful creation and scrutiny of train, validation, and test datasets is very time consuming, consequently limiting the number of available samples for experimentation and excluding images depicting long multipaned story illustrations reducing sampling diversity.

We also encountered GPU compatibility issues due to the fact that learning algorithms from the Hateful Memes Challenge necessitate appropriate CUDA versions to operate. Hence, we could not use the first ranked classifier for our study given its high CUDA version requirement for our experimentation environment. Needless to say, it is worthy to continue this exploratory research using methods by (Zhu 2020) to observe their model's learnability.

Finally, it is important to note that our sociocultural identity has a strong influence on our understanding of online content. For instance, the ethnic background of one individual may cause a hateful perception toward an image meme, but perhaps not by another individual. This differentiation in points of view has arguably been the greatest challenge in our work thus far and is evident in the experimental classifier's biased tendency to label samples containing certain terms (e.g., Jew) as hateful when in reality such terms are also used in non-hateful contexts. Nevertheless, we hope that future work will study the visual characteristics of meme images to better interpret the true intentions of their creators.

7.2 Main Take-Aways

This paper presented a multimodal deep learning approach to determine whether advancements made toward the detection of hateful memes by the Hateful Memes Challenge and the solutions thereof generalize to 4chan and other fringe communities. Our experiments showed that the inclusion of text in image memes does not significantly impede the spread of extremist views, given the very close classification scores obtained by the models when evaluated on unimodal memes.

We found that Kiela et al.'s challenge dataset (Kiela et al. 2021) does not realistically depict actual memes shared on social media, which has resulted in the development of learning algorithms that are incapable of adequately recognising hateful memes from other social networking platforms.

Our results also attested to the effectiveness of ensemble V&L classifiers for enhancing detection performance.

Overall, our work provides a first step toward assessing the viability of state-of-the-art multimodal machine learning methods, in an effort to improve the creation and deployment of autonomous systems for hate speech detection in memes posted on the Web.

References

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Crockford, D. 2022. JSON Lines. https://jsonlines.org/.

DataReportal. 2021. Global Social Media Stats. https://datareportal.com/social-media-users.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*, 5.

Deza, A.; and Parikh, D. 2015. Understanding image virality. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

DrivenData Inc. and Facebook AI. 2020. Hateful Memes: Phase 1. https://www.drivendata.org/competitions/64/hateful-memes/page/206/.

FacebookAI. 2020. Hateful Memes Challenge and dataset for research on harmful multimodal content. https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/.

Girshick, R.; Radosavovic, I.; Gkioxari, G.; Dollár, P.; and He, K. 2018. Detectron. https://github.com/facebookresearch/detectron.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2018. Mask R-CNN. *arXiv:1703.06870*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385*.

Hine, G.; Onaolapo, J.; De Cristofaro, E.; Kourtellis, N.; Leontiadis, I.; Samaras, R.; Stringhini, G.; and Blackburn, J. 2017. Kek, cucks, and god emperor trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. In *ICWSM*.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2021. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv:2005.04790*.

Kingma, D. P.; and Ba, J. 2014. ADAM: A method for stochastic optimization. *arXiv:1412.6980*.

Know Your Meme. 2022. Sheeeit Meme. https://knowyourmeme.com/memes/sheeeit.

LeCun, Y. 2022. JaidedAI/EasyOCR: Ready-to-use OCR with 80 supported languages and all popular writing scripts including Latin, Chinese, Arabic, Devanagari, Cyrillic and etc. https://github.com/JaidedAI/EasyOCR.

Ma, J.; and Yarats, D. 2021. On the adequacy of untuned warmup for adaptive optimization. *arXiv:1910.04209*.

M.Curtis, W. 2016. Hate speech. Encyclopedia Britannica – https://www.britannica.com/topic/hate-speech.

Muennighoff, N. 2020. Vilio: State-of-the-art Visio-Linguistic Models applied to Hateful Memes. *arXiv:2012.07788*.

Roose, K. 2021. The Truth About 'The Truth About' Memes. https://www.nytimes.com/2020/09/24/technology/truth-about-breonna-taylor-meme.html.

Trafton, A. 2014. In the blink of an eye. https://news.mit.edu/2014/in-the-blink-of-an-eye-0116.

Velioglu, R.; and Rose, J. 2020. Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge. *arXiv:2012.12975*.

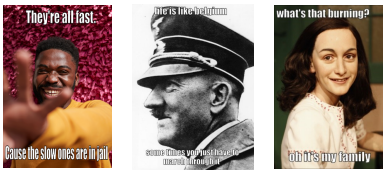
Zannettou, S.; Caulfield, T.; Blackburn, J.; Cristofaro, E. D.; Sirivianos, M.; Stringhini, G.; and Suarez-Tangil, G. 2018. On the Origins of Memes by Means of Fringe Web Communities. In *ACM IMC*.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 786–797.

Zhu, R. 2020. Enhance Multimodal Transformer With External Label And In-Domain Pretrain: Hateful Meme Challenge Winning Solution. *arXiv:2012.08290*.

(a) **(b)** **(c)** **(d)**

Figure 5: National stereotypes and famous figures: (a) Mexican, (b) American, (c) Donald Trump caricature, and (d) Barack Obama from Zannettou et. al.'s dataset (Zannettou et al. 2018).



(a) **(b)** **(c)**

Figure 6: Subject of hateful memes (a) an African American stereotype, (b) Adolf Hitler, and (c) Anne Frank from the HM dataset (Kiela et al. 2021). ©Getty Images

.

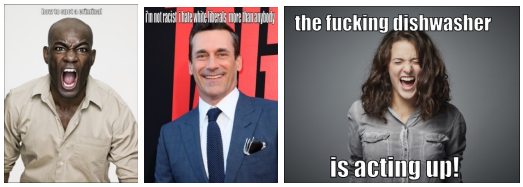
A Feature Importance & Virality

In this section, we conduct a formal analysis of the characteristics of hateful memes that contribute to its virality potential with respect to the most prominent features recognized by the two best performing classifiers from the first and third experiments (the VisualBERT CC and average prediction ensemble models). Hereby, we discuss the top four features learned by the models leveraging Know Your Meme – the largest encyclopedia of Internet memes – as a guideline for the affirmation of memes' virality.

1. **Subject matter:** 67% of viral memes in the training set and 55% of true positive (TP) classifications made by the average-prediction ensemble depict a character, stereotype representation, caricature, or famous individual. This suggests that images containing a region of primary focus (or emphasis) have a greater likelihood of becoming viral. We find this to be the case as the subject of a hateful meme image is indicative of its target audience (see example Figure 5). Likewise, the VisualBERT CC classifier correctly predicts 599 and 520 hateful memes from sub-experiment test sets 1 and 2, respectively. Although Kiela et al.'s dataset does not consist of viral memes, we still see that the model's performance was influenced by racial stereotype portrayals (e.g., Figure 6a) and impactful historical figures (as shown in Figures 6b and 6c), further supporting this finding.
2. **Facial expressions:** Image subjects that portray emotions through their facial expressions strongly impacted classification decisions made by the VisualBERT CC classifier (93.7% of viral hateful memes in the training set, and 84% of TPs displayed this feature), showing that meme virality is influenced by the expression of sentiment to advocate beliefs. The majority of viral hateful 4chan memes used in the first experiment depict subjects which convey emotions through facial expressions (e.g., Figure 7), and 87% of hateful test samples were classified correctly by the VisualBERT CC model while demonstrating this attribute (see Figure 8). Peculiarly, the two hateful memes shown in Figure 9 were falsely predicted as non-hateful when tested under the condition of *Unimodal Memes*, implying that

(a) (b) (c) (d)

Figure 7: Emotions portrayed through facial expressions: (a) surprise and malevolence, (b) fear, (c) sadness, and (d) anger from Zannettou et. al.'s dataset (Zannettou et al. 2018).



(a) (b) (c)

Figure 8: TP test samples portraying: (a) enragement, (b) happiness, and (c) frustration from the HM dataset (Kiela et al. 2021). ©Getty Images.

image features can be as informative as text for the anticipation of a viral meme.

3. **Gestures:** Gestures of meme subjects are on par with their facial expressions with regard to feature importance. This attribute also indicates underlying connotations of an image to change the entirety of its meaning (e.g., the text in Figure 11a alone is not hateful yet demonstrates antisemitic undertones when interpreted with the stereotypical illustration of a Jewish man malevolently rubbing his hands together). Subsequently, subjects' gestural behavior is considered by the average-prediction model before producing a final prediction solely based on its textual meaning (93.7% and 84% of viral hateful memes in the train and test sets possess this trait together with the top 1 feature). Similarly, the VisualBERT CC classifier uses this characteristic to assess each test sample in its entirety prior to classification (refer to Figure 10b for an example of a correctly labelled sample indicating this attribute).
4. **Proportion:** The majority of hateful viral memes possessing two or more of the above-mentioned features use a close-up shot such as those shown in Figure 11 (70% and 84% of viral hateful memes in the training sets used for the first and third experiments tightly frame their subjects of focus – the same is the case for 29% and 55% of TPs made by the VisualBERT CC and average-predictions models). We argue that meme authors depict the full form of the image's figure to convey their message more clearly through facial emotional expressions and gestures. Results from *The Effects of Multimodality in Hateful Memes* show that the VisualBERT CC model picks up on this feature to distinguish between the two classes for Kiela et al.'s test samples (see Figure 12 for TP prediction examples).

Figure 9: Misclassified samples from Multimodal Memes, from the HM dataset (Kiela et al. 2021). ©Getty Images.

(a) (b)

Figure 10: Example of: (a) textual meaning influenced by gestural behavior, (b) correctly predicted sample containing contradictory modalities, from Zannettou et. al's dataset and the HM dataset (Zannettou et al. 2018; Kiela et al. 2021). ©Getty Images.

Figure 11: Example close-up viral hateful meme images from Zannettou et. al.s dataset (Zannettou et al. 2018). ©Getty Images.

Figure 12: TPs made by VisualBERT ensemble classifier, from the HM dataset (Kiela et al. 2021). ©Getty Images.