Text or Image? What is More Important in Cross-Domain Generalization Capabilities of Hate Meme Detection Models?

Piush Aggarwal[1], Jawar Mehrabanian[2], Weigang Huang[3], Özge Alacam[4], and Torsten Zesch[1]

[1]CATALPA, FernUniversität in Hagen {*firstname.lastname@fernuni-hagen.de*}
[2]FernUniversität in Hagen {*jawar.mehrabanian@studium.fernuni-hagen.de*}
[3]Universität Duisburg-Essen {*weigang.huang@stud.uni-due.de*}
[4]LMU Munich and Universität Bielefeld {*oezge.alacam@uni-bielefeld.de*}

Abstract

This paper delves into the formidable challenge of cross-domain generalization in multimodal hate meme detection, presenting compelling findings. We provide enough pieces of evidence[1] supporting the hypothesis that only the textual component of hateful memes enables the existing multimodal classifier to generalize across different domains, while the image component proves highly sensitive to a specific training dataset. The evidence includes demonstrations showing that hate-text classifiers perform similarly to hate-meme classifiers in a zero-shot setting. Simultaneously, the introduction of captions generated from images of memes to the hate-meme classifier worsens performance by an average F1 of 0.02. Through blackbox explanations, we identify a substantial contribution of the text modality (average of 83%), which diminishes with the introduction of meme's image captions (52%). Additionally, our evaluation on a newly created confounder dataset reveals higher performance on text confounders as compared to image confounders with an average ΔF1 of 0.18.

1 Introduction

Recently many hate-meme detection multimodal (MM) systems have been proposed, see (Sharma et al., 2022) for a survey and (Kougia and Pavlopoulos, 2021; Aggarwal et al., 2021; Gold et al., 2021; Zhu, 2020; Muennighoff, 2020; Li et al., 2019; Chen et al., 2020) for individual contributions, but it is an ongoing concern that they do not generalize well in a cross-domain setting. Possible causes are (i) the implicit knowledge captured by multimodal hate messages (memes) (Ma et al., 2022; Gomez et al., 2020; Zhong et al., 2016; Hosseinmardi et al., 2015), (ii) additional annotation noise in multi-modal settings (Oriol Sàbat, 2019), and (iii) more complex network architectures.
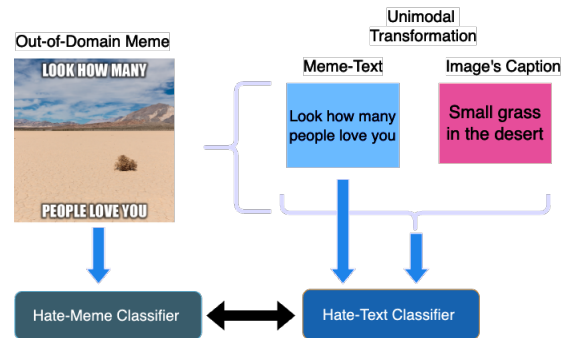

Figure 1: Illustration of our experimental arrangement for assessing the hate meme model's performance compared to unimodal text-based hate classifiers. The evaluation involves a test meme from a domain not included in the model's training data.

In this study, we explore the generalization capabilities of MM models for detecting hate memes. While previous studies (Wang et al., 2020; Ma et al., 2021) support the significant role of image modality in other multimodal-based downstream tasks, however, in the case of meme classification, the meaning can only be correctly inferred from also looking at the image, so we find the analysis to be of special importance and worth replicating. Consequently, we initiate the evaluation of these models in settings outside their domain. We observe a significant decline in performance, with an average macro F1 score of 0.28.

We aim to tackle this issue by utilizing a text-only (unimodal) hate classifier, specifically crafted for the detection of hateful memes. Previous research (Nozza, 2021; Alshalan and Al-Khalifa, 2020; Talat et al., 2018) demonstrates relatively higher generalization capabilities in the context of unimodal text-only hate. Our approach involves applying an unimodal transformation to memes by concatenating the text within the meme with a caption generated from the meme's image. Subsequently, we train a text-based classifier using a combination of nine diverse hate speech datasets

---

[1]Our code and dataset are released at `https://github.com/aggarwalpiush/HateDetection-TextVsVL`

Name	Reference	# Train/Dev/Test	tokens	% hate	Domain
HARMEME	(Pramanick et al., 2021)	3.5k	160k	26.21	Covid-19/US Election
MAMI	(Fersini et al., 2022)	10k	590k	50	Misogynistic
FB	(Kiela et al., 2020)	10k	370k	37.56	*mixed*

Table 1: Properties of hate-meme datasets used in our study.

and assess its performance on a transformed meme test set. We observe performance levels from our unimodal classifier that are comparable to those of MM models. In certain instances, our unimodal classifier even exhibits an improvement in performance, with an average F1 increase of 0.05 compared to late-fusion-based MM models. The results make us infer that the text modality demonstrates superior generalizability compared to the image modality in detecting hateful memes. Figure 1 gives an overview of our experimental setup.

Additionally, we find that MM models behave differently than textual-based models. We retrain the MM models on hateful meme datasets which also include captions generated from the images available in the memes. Surprisingly, in comparison with existing models, in general, we find small performance drops (average ΔF1 of 0.02) in both in-domain as well as out-of-domain settings regardless of the presence of captions in the test sets.

We explain the behavior of MM models by computing the contribution of text and image modality individually toward the prediction. We apply Shapley values (Parcalabescu and Frank, 2023) to the features used in the models and average the final score for each modality (Section 3). Our results indicate a substantial contribution (83%) of textual modality by the models evaluated on all the datasets we have used in our study. Nevertheless, incorporating the image caption of the meme into the input data during the MM model training results in a decreased textual contribution of 52%. We believe that images in hateful memes are more like facilitators and provide context to the MM models.

To validate this, we compose a confounder dataset where we subset from the HARMEME and FB dataset (Pramanick et al., 2021; Kiela et al., 2020), selecting 100 memes featuring celebrities or known figures such as *Donald Trump, Nelson Mandela and Adolf Hitler*. We observe that MM models are sensitive to text confounders, while the prediction labels remain unchanged when the model is triggered with image confounders. (An average Δ F1 of 0.18 is observed when the MM model is evaluated on Text and Image confounder sets).

Although, prior studies such as (Wang et al., 2020; Ma et al., 2021) have represented similar hypotheses. However, we find such studies for explicit types of downstream tasks.

In this paper, we present compelling evidence substantiating the hypothesis that the generalization of multimodal classifiers across diverse domains is primarily attributable to the textual component of hateful memes. Remarkably, our findings reveal a heightened sensitivity of the image part to the nuances of a specific training dataset. We believe we are the first to provide a thorough analysis supporting this idea, making our work unique in contributing to the field.

2   Related Work

Kirk et al. (2021) demonstrate the high generalization behaviour of CLIP models (Radford et al., 2019) when it is fine-tuned on the Hateful meme FB dataset (Kiela et al., 2020) and tested on in-house hate meme test set collected from pinterest[2]. However, their model is evaluated without using the meme's text which we believe provides significantly greater valid information for hateful meme detection. Cuo et al. (2022) attempts to investigate the poor generalizability behavior of VL-models towards COVID-19-specific hate meme detection task. The application of the gradient-based explanation method demonstrates the significance of image modality is twice of textual one during predictions. Not specific to hate meme classification task, Ma et al. (2022) evaluate the robustness of Visual-Linguistic transformers on missing modality datasets and found even poorer performance than uni-modal models and proposed a method that performs an optimal fusion of modalities which end up with better results. Error analysis of visio-linguistic models also indicates model bias (Hee et al., 2022). While prior studies recommend investigating the contributions of each modality to model predictions to uncover the root cause of their lim-

---

[2] https://www.pinterest.com/

ited generalization, these analyses tend to be overly specific, focusing solely on the in-house COVID-19 test set. Additionally, suggested methods like gradient-based explanation (Selvaraju et al., 2019) are susceptible to deception through small input changes, as demonstrated in adversarial attacks (Parcalabescu and Frank, 2023).

3 Modality Contribution with Shapley Values

Applying the method proposed by Parcalabescu and Frank, 2023, we attempt to investigate the modality contribution of existing hate meme detection models. There are multiple existing methods that can be used to estimate the importance of the model's features in the prediction process. Shapley values provide important ingredients for sample-based explanations that can be aggregated in a straightforward way into dataset-level explanations for machine learning methods (Covert et al., 2020). We calculate Shapley values for meme text tokens and image patches utilized in MM models during prediction. Each entity (token or patch) through its shapely value gauges its impact on the model prediction, such as the likelihood of image-sentence alignment. It can be positive (enhancing the model prediction), negative (diminishing it), or zero (no discernible effect).

4 Datasets

4.1 Hateful Meme Datasets

In order to analyze the generalizability of available hate meme classifiers and modality contribution, we have used three benchmark datasets (see Table 1).

Kiela et al. (Kiela et al., 2020) (FB) comprises 10,000 memes sourced from Getty images, semi-artificially annotated with benign confounders. It includes (i) *multimodal hate* where both modalities possess benign confounders, (ii) *unimodal hate* where at least one of the modalities is already hateful, (iii) *benign image*, (iv)*benign text* confounders and (v) *random not-hateful* examples. The first four are labeled as *hateful*, while the last is labeled as *non-hateful*. The dataset is divided into 85% training, 5% development, and 10% test sets, with balanced proportions for each meme variety in the development and test sets.

Pramanick et al. (Pramanick et al., 2021) (HARMEME) consists of COVID-related memes from US social media, identified using keywords like *Wuhan virus*, *US election*, *COVID vaccine*, *work from home*, and *Trump not wearing mask*. Unlike (Kiela et al., 2020), these memes are original, shared across social media, and their textual content is extracted using Google Vision API. The dataset is categorized into *hateful* (including *harmful* and *partially harmful*) and *non-hateful*, totaling 3,544 data points. The split for training, validation, and test sets is 85%, 5%, and 10%, respectively.

Fersini et al. (Fersini et al., 2022) (MAMI) focuses on SUBTASK-A, with memes labeled as *misogynist* or *non misogynist*. These are relabeled as *hateful* and *non-hateful* for consistency. The memes are collected from social media threads featuring women personalities such as Scarlett Johansson, Emilia Clarke, etc. as well as hashtags such as `#girl`, `#girlfriend`, `#women`, `#feminist`. Google Vision API is used for meme text extraction. With a balanced set of 10,000 instances, 10% are used for both development and test sets, randomly stratified.

4.2 Confounder Dataset

In order to validate the generalization capabilities of multimodal (MM) models for a specific modality, we create a tailored dataset for validation. We conducted an exhaustive search on the FB (Kiela et al., 2020) dataset. A meticulous filtration process was implemented to exclude any instances featuring recognized celebrities or known figures such as *Donald Trump, Nelson Mandela, and Adolf Hitler*. Subsequently, attention was directed towards memes labeled as *hateful*. The selection was judiciously limited to a total of 100 figures, ensuring controlled and representative samples. The final stage of the methodology involved leveraging the identified set of hateful memes to construct a total of 100 benign images and text confounders. For image confounders, manual replacement of the celebrity figure with an analogous counterpart such as *Anne Frank* with *Adolf Hitler* (See Appendix A for complete list of the figures that were taken into account for the confounder dataset). Furthermore, to maintain simplicity and coherence for text confounders, the Polyjuice framework was incorporated (Wu et al., 2021). It is a counterfactual generator, that is instrumental in facilitating control over the nature and positioning of perturbations in the textual content, enhancing the precision and consistency of the devised framework. Figure 2
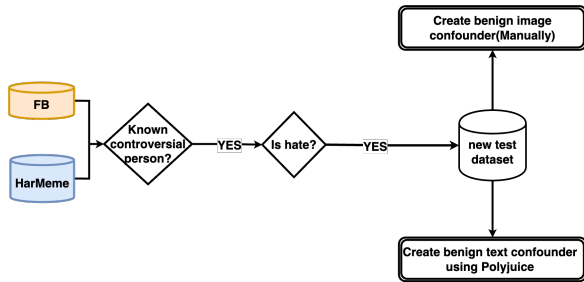
Figure 2: A schematic showing the data collection process of our proposed dataset.

illustrates the data collection process for our proposed dataset.

Annotation process We recruited 12 annotators (university graduated volunteers and regular social media users) to read the introduction, where the objective of the annotations along with the task is explained in detail (See Appendix B). From each of the collected memes, we solicit various aspects for analysis. First and foremost, we inquire about the *Image-text Relation*, seeking insights into the nuanced connection between the textual and visual components within a meme. Another crucial facet is the *Modality towards Hate*, which serves as an evaluative measure for the modality of a meme that may convey hate or offensive content. For a more granular understanding, we introduce *Decision Parts*, allowing annotators to pinpoint specific tokens or elements in the meme that contribute to its characterization as either hateful or non-hateful. To quantify the degree of hateful content, we employ a *Hatefulness Score*, utilizing a scale that ranges from 0 to 5. A score of 0 denotes non-hateful content, while a score of 5 signifies highly hateful material. Additionally, annotators are prompted to provide a *Confidence Score* reflecting their certainty regarding the accuracy of their judgments. This score operates on a scale from 0 (indicating a lack of confidence) to 5 (reflecting a high level of confidence). To maintain the integrity of the annotation process, we afford annotators the option to discard a sample, ensuring that only pertinent and valid memes[3] are included in the analysis. We ended up with very good inter-rater agreement among the annotation with Krippendorff alpha (Krippendorff, 2011) as 0.8. Furthermore, it is noteworthy that the average *Confidence scores*

is 4.38 out of 5 which shows very high confidence among the annotators.

5 Experimental Models

Unimodal Hate Recognition We use an online hate speech detection system called Perspective API[4] which consists of multilingual BERT-based models trained on millions of comments from a variety of sources, including comments from online forums such as Wikipedia and The New York Times. These models are further distilled into single-language Convolutional Neural Networks (CNNs) for different languages. We also fine-tune BERT (Devlin et al., 2019) and SVM-based hate detection models on nine hate speech datasets which will be discussed in Section 6.2.

Multimodal Hate Recognition Most of the promising studies on hate speech detection employ multi-modal based visual-linguistic pre-trained models (Chen et al., 2020; Li et al., 2019, 2020; Su et al., 2020; Tan and Bansal, 2019) which are originally designed to tackle basic visual-linguistic problems such as visual-question answering (VQA). These models caries semantic understanding between text and visual objects which makes them highly efficient for many downstream tasks. To analyze the vulnerability of the hate meme detection models, we investigate two early fusion and one late fusion-based multimodal (MM) models. *VisualBert* (Li et al., 2019), an early fusion visual-linguistic transformer-based model, pre-trained on image caption as well as VQA datasets. We also investigate *Uniter* model (Chen et al., 2020) stands for UNiversal Image-TExt Representation which is also an early fusion visual-linguistic transformer-based model with additional pre-training with Visual Genome, Conceptual Captions, and SBU Captions. As the third MM model, we train a late-fusion ensemble model where we employ distinct extraction pipelines for image and text features. For image feature extraction, we utilize Resnet (He et al., 2016), a highly deep residual learning framework designed for generating image features. To derive the text representation, we employ the widely-used RoBERTa model (Liu et al., 2019). Subsequently, we concatenate the features from both modalities and feed them through a 128-layer feed-forward network with ReLU activation and a dropout rate of 0.2 to produce predictions. The

---

[3]We offer annotators the option to exclude a meme if they lack sufficient knowledge to comprehend its content. Ultimately, we include only those memes that none of the annotators choose to discard.

model is trained for 30 epochs using the Adam optimizer (Kingma and Ba, 2014), with a learning rate of 10^{-5} and weight decay set to 0.1. This classifier is referred to as *Rob+Resnet* for the purpose of illustration.

Image Caption Generation We use *ClipCap* (Mokady et al., 2021), which is based on Contrastive Language Image Pretraining (CLIP) (Radford et al., 2021) model to encode the image and pre-trained language model GPT-2 (Radford et al., 2019) to decode a caption. We also use *BLIP* (Li et al., 2022) which is a multimodal mixture of encoder-decoders optimized on three objectives during the pre-training process which include image-text contrastive loss, image-text matching loss, and language modeling loss. Unlike other models, it also performs caption bootstrapping in order to deal with noisy input data. We use both of these models in their default settings[5].

6 Experiments & Results

We conduct multiple sets of experiments in this study. Initially, we assess the cross-domain performance of hate-meme classifiers to gauge their generalization capabilities. Subsequently, we compare the performance of text-only hate classifiers on the textual component of memes with that of the hate-meme classifiers. We also assess the impact of captions generated from the image component of memes on text-only hate classifiers and hate meme detection models. Additionally, we compare the modality contribution from the blackbox explanations of the models with and without the introduction of captions. Finally, we apply the models to a confounder dataset to evaluate their sensitivity to a particular modality confounder set.

6.1 Generalization of Hate-meme Classifiers

To test the generalization capabilities of hate-meme classifiers, we fine-tune three state-of-the-art pre-trained models (VisualBert, Uniter and Rob+Resnet) on one datasets (train split) and test on the test splits of all three resulting in 9 train-test scenarios per model as can be seen in Table 2. Overall, we find huge performance drops across all the datasets for cross-domain testing. Since domains of HARMEME and MAMI are exclusive, we encounter symmetry among each other (F1 of .398 and .393 for VisualBert and .453 and .467 for

		Test		
	Train	HARMEME	MAMI	FB
VisualBert	HARMEME	.80	.40	.48
	MAMI	.39	.85	.51
	FB	.44	.60	.66
Uniter	HARMEME	.79	.45	.48
	MAMI	.47	.85	.53
	FB	.57	.54	.64
Rob+Resnet	HARMEME	.79	.40	.47
	MAMI	.39	.83	.45
	FB	.41	.49	.62

Table 2: F1(Macro) score of Hate-meme classifiers in cross-domain settings. Grey highlighted values represent in-domain baselines.

Reference	# Posts	tokens	% hate
(Davidson et al., 2017)	25K	245K	6
(Mollas et al., 2022)	1K	14K	43
(Kennedy et al., 2022)	28K	411K	15
(de Gibert et al., 2018)	10K	169K	11
(Mandl et al., 2019)	7K	174K	36
(Basile et al., 2019)	13K	254K	4
(Samoshyn, 2020)	2K	38K	48
(Waseem and Hovy, 2016)	17K	131K	32
(Waseem, 2016)	4K	31K	16
Total	107K	1467K	23

Table 3: Hatespeech datasets used to train the hate-text classifiers. For all datasets, the collection is based on hate slurs matching, therefore all of them consist *mixed* domains.

Uniter). On the hand, for FB, as there is no specific domain, we find relatively less decrement (however it is still huge) in the F1 scores. The results clearly infer a lack of generalization capabilities among these models.

6.2 Zero-shot Text-only Classifiers

We now compare the multimodal hate-meme classifiers to unimodal text-only classifiers. For that purpose, we train two text-only classifiers (SVM and BERT) on a large collection of hate speech datasets (see Table 3). Overall, we use around 0.1 Million posts having 1.4 Million tokens out of which 23% posts are hateful. In the case of SVM, for tokenization and feature extraction, we use ArkTokenizer and fastext embeddings respectively. In the case of BERT, we follow the uncased-large model[6] for fine-tuning. We also use the hate speech classifier

		Testset			
Model		HARMEME	MAMI	FB	
Image + Text	VisualBert	.44	.60	.51	1
	Uniter	.57	.54	.53	2
	Rob+Resnet	.47	.45	.49	3
Text	Perspective API	.45	.52	.49	4
	BERT	.48	.53	.52	5
	SVM	.45	.45	.41	6
Text + caption (ClipCap)	Perspective API	.50	.52	.53	7
	BERT	.50	.54	.52	8
	SVM	.47	.45	.43	9
Text + caption (BLIP)	Perspective API	.50	.53	.53	10
	BERT	.51	.53	.53	11
	SVM	.46	.44	.43	12

Table 4: Hate-meme vs. Hate-text Classifiers F1 Performance on cross-domain data. For Hate-meme classifiers, we indicate the best F1 value among the two training sets. A color gradient ranging from red to green is employed to emphasize the transition from lower to higher F1 values, respectively.

as provided by the Perspective API[7] which outputs a toxicity score for a given text. A toxicity score greater than 50% is considered hate otherwise non-hate.

Table 4 compares the zero-shot domain transfer results of hate-meme and hate-text classifiers. We encounter a close resemblance between them in their performances. Among cases where the text-only classifier is applied only on meme text, BERT model performance is superior to the rest of the two with an average F1 score of .51 followed by Perspective API (F1 of .59) (depicted in Table's line 4 and 5). We observe a similar performance by multimodal hateful meme classifiers (average F1 of .52 for *VisualBert* and .55 for *Uniter* and .47 for *Rob+Resnet*) (see line 1, 2 and 3).

With the quite good performance of the text-only classifiers, it might be worthwhile trying to extract the semantics of the image as text. For this purpose, we append captions generated by caption models (see Section 5) along with meme text and input to the hate-text classifier that we have trained on multiple corpora (as described in Section 6.2). Table 4 illustrates the performance of *ClipCap* and *BLIP* models. Compared with hate-meme and hate-text classifiers, we find a slight improvement in

BERT with an average F1 of .52 which is 1 point higher than BERT tested only on meme text (depicted in line 8). However, it is 3 points lower than the Uniter model. Notably, Perspective API exhibits an improvement in performance, with an average F1 increase of 0.05 compared to *Rob+Resnet* model (depicted in line 10). This outcome suggests that classifier generalization is predominantly influenced by the textual modality. This pattern further implies a potential bias towards textual elements in meme data, leading to limitations in the ability of the multimodal model to integrate image meaning for this particular task. Mann-Whitney U Test shows that the results are statistically significant with $p < 0.05$.

6.3 Impact of Captions

In this section, we illustrate the effect of incorporating the captions in the training. During the training process for each of the hateful meme classifiers, we incorporate image captions generated using a *BLIP* model into the *Rob+Resnet*. We then assess the performance of the resulting model that includes captions in comparison to its original counterpart. Our evaluation is conducted both on (i) the original test set and (ii) plus with captions. In Table 5, it is evident that when the models trained with including captions perform poorly in both in-domain and out-of-domain testing scenarios, regardless of the presence or absence of captions in the test sets. A plausible explanation for this phenomenon could be the neutralization of contextual nuances introduced by the supplementary captions in the meme's text. However, we also see a performance increase (average Δ F1 of 0.09) in the case of the model trained on HARMEME dataset when tested on an out-of-domain test set with concatenated captions. One potential explanation for this behavior could be attributed to the high resolution of the original images in this dataset, marked by an average bit depth of 43.90, a notable contrast to other datasets, with bit depths of 9.54 for the FB and 4.30 for the MAMI dataset (Aggarwal et al., 2023).

6.4 Impact of Modality

Shapley Values Computation To calculate modality contribution, we determine Shapley values for feature maps, which are utilized by MM models for prediction. To achieve this, we generated patches of meme images such that each text token will be generally represented in a patch. From the existing set of image patches and text tokens

	Train	Test (Meme text)			Test (with Caption)		
		HARMEME	MAMI	FB	HARMEME	MAMI	FB
Meme text	HARMEME	.79	.40	.47	.65	.51	.55
	MAMI	.39	.83	.45	.40	.81	.47
	FB	.41	.49	.62	.35	.49	.56
With Caption	HARMEME	.77	.41	.46	.65	.52	.53
	MAMI	.39	.77	.42	.39	.78	.44
	FB	.41	.50	.49	.34	.49	.48

Table 5: F1(Macro) score of Roberta+Resnet based Hate-meme classifiers when trained with image caption.

	Train	Test (Meme text)		(with Cap+Celeb)	
		I	T	I^+	T^+
Meme text	HARMEME	.42	.45	.46	.44
	MAMI	.17	.39	.19	.39
	FB	.43	.75	.41	.72
With Caption	HARMEME	.27	.39	.34	.43
	MAMI	.10	.34	.17	.42
	FB	.33	.53	.35	.54

Table 6: F1(Macro) score of Roberta+Resnet based Hate-meme classifiers on Confounder datasets (T: Text Confounders, I: Image Confounders).

(a) FB
TS=0.69
F1=0.62

(b) Harmeme
TS=0.85
F1=0.79

(c) Mami
TS=0.95
F1=0.83

(d) FB (+C)
TS=0.12
F1=0.48

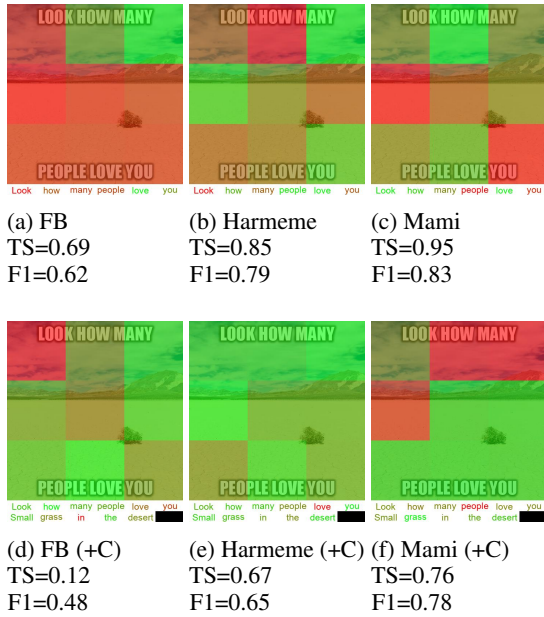(e) Harmeme (+C)
TS=0.67
F1=0.65

(f) Mami (+C)
TS=0.76
F1=0.78

Figure 3: Example of Modality Contrbution of Rob+Resnet based hate meme detection model when trained on different hateful meme datasets. Here notation (+C) refers additional caption used in model training. RED and GREEN colour illustrate low and high contribution respectively.

(entities), we selected a subset and masked the

remaining entities in the set. The determination of the number of subsets was influenced by the Monte Carlo approximation method. The Shapley value for each entity is computed by subtracting the model's output while it is present from that while it is absent. The resulting value was normalized considering the possible combinations of subsets. Ultimately, to compute the Shapley values for text contributions, the result outcomes of textual tokens are summed and normalized. The following algorithm delineates the process of generating modality scores.

Input: Meme image I, Meme text T, model f, random number P, Shapley Value ϕ, Text contribution score TS
image patches $I_p = \lceil \sqrt{\text{len}(T)} \rceil^2$

1: **for all** $t \ \epsilon \ t_1, ..., (I_p + T)$ **do**
2:     **for all** $i \ \epsilon \ 1, ..., 2 * P + 1$ **do**
3:         choose subset $S \subseteq (I_p + T)$ where $len(S) = i$ and $t \notin S$
4:         $\phi(t) = \sum_{S,t} \frac{f(S+t)-f(S)}{\gamma}$ where γ is normalizing factor
5: $\phi(T) = \sum_{n=1}^{len(T)} \phi(t_n)$
6: $\phi(I) = \sum_{n=1}^{len(I_p)} \phi(t_n)$

$$\text{\textbf{return} } TS = \frac{\phi(T)}{\phi(T) + \phi(I_p)}$$

Explanation of Hate Meme Detection Models
We calculate the contributions of modalities toward predictions of late fusion ensemble based *Rob+Resnet* model using Shapley values. The classifiers are trained on each of the datasets with and without captions concatenated with the meme's text. To illustrate, Figure 3 shows Shapely values on a meme example for different models. The colours RED and GREEN indicate low and high contributions, respectively. In addition, the Text Contribution score (TS) as well as the F1 score evaluated on the in-domain evaluation set is provided in the caption of each of the subfigures. Image modality contribution (IS) can be computed using $IS = 1 - TS$. We find that text contribution is quite higher ($TS >> IS$) for all the models (average TS of .83). When a caption is added to the text, the contribution score of the text modality decreases to .52. With this we infer that adding captions to memes strengthens the focus on the image modality. We observed that when we include the image caption along with the meme's text, the models establish a correlation between the caption and the meme's image. In such cases, the models tend to focus on the image's information related to the image caption, a behavior not exhibited when the caption is absent. It infers that the meme text inherently carries a more potent message of hatefulness, which is mitigated by the inclusion of image captions. Nevertheless, it's important to highlight that the F1 score also decreases when captions are introduced to meme text. This might also mean that to the existing models, images in hateful memes are more like facilitators and provide context to the models. As an example, in Figure 3 we see that the dominancy of image patches is much higher for models trained along with captions. Similarly, we also see less dominancy of important hate context tokens such as *LOVE* (Gröndahl et al., 2018; Aggarwal and Zesch, 2022) in this case.

6.5 Classifiers on Confounder Dataset

In Section 4.2, we elaborate on the composition of the confounder dataset. We divide it into two subsets. The first subset is termed the text confounder set (T), wherein meme instances are categorized based on images resembling those in hateful memes. Similarly, the second subset is designated as the image confounder set (I), where meme instances are categorized based on text resembling that found in hateful memes. In addition, we also concatenate the textual component of these sets with the image's caption and names of the celebrities available in the image and called them extended sets (T^+ and I^+). In this way, we have four evaluation sets to assess hate meme classifiers.

We evaluate *Rob+Resnet* classifier which is already trained on the original hateful meme datasets and also in concatenation with captions. Table 6 illustrates the classifier's performance in terms of F1 (macro) scores. Overall, the performance on T is notably higher than that on the I across all variants of models. However, this difference is quite small in the case of HARMEME dataset (the average Δ F1 is 0.26, 0.23, and 0.08 for FB, MAMI and HARMEME respectively). A similar trend is observed in the case of extended sets. Overall there is Δ F1 of 0.18 is observed which illustrates that the classifier is highly sensitive to memes undergoing changes in text while maintaining the same image, a sensitivity not observed in the other modality. Similar to the observations in Table 5, the addition of captions to the meme's text significantly reduces performance for both the image and text confounder sets. This further adds evidence of the importance of the textual component of memes for hate detection models.

7 Conclusion

Commencing from the observation that multimodal hate-meme classifiers exhibit poor generalization to other datasets, we demonstrate that comparable cross-domain performance can be achieved by disregarding the image segment and concentrating solely on the text. Furthermore, we reveal that text classifiers exhibit improved performance when incorporating image content into the text classifier through image captioning. Intriguingly, the introduction of captions generated from meme images to the hate meme classifier leads to a deterioration in performance. The insights obtained from the analysis of modality-specific contributions, along with the diminishing effect of including captions, indicate that current multimodal models are primarily focused on finding alignment between image and text tokens at a concrete level. The addition of captions generated by other multimodal models misdirects attention to those low-level alignments, whereas text-image alignment in hate text classifiers typically occurs at a more abstract (metaphor-

ical) level. It is evident that the meaning of the image could be extracted and incorporated at a higher level, where current models and training regimes fall short in addressing this issue. Additionally, our evaluation on a newly established confounder dataset underscores superior performance on text confounders as opposed to image confounders. These findings strongly support the assertion that the image component of multimodal hate meme classifiers exhibits limited transferability, with the generalization capabilities primarily dependent on the text component of the meme.

8 Limitations

Employing a proprietary API such as Perspective API introduces challenges to reproducibility. Nonetheless, we mitigate this limitation by training our own BERT classifier, offering a comparably high-performing and fully reproducible alternative. In our approach, we consciously restrict ourselves to a single multimodal classifier, chosen for its high efficiency in general, for both the explanation phase and confounder study. However, consistency across the results enhances the viability of our analysis. There is a lack of propositions about questions such as why MAMI models are different than others not affected at all from caption inclusion, or what makes HARMEME models easier than FB's (as shown in Table 4). In this study, our focus has been on employing existing models that have been utilized or proposed for the Hateful Meme Classification task. Nevertheless, it is also worthwhile to acknowledge the study like multimodal gate method introduced by (Arevalo et al., 2020). Such a method proposes a systematic control over the contributions of modalities through a multimodal gate mechanism. We also believe that adopting such an approach could offer insights into several aspects, including (i) potential enhancements in hate meme detection, (ii) investigating whether the challenges stem from insufficient attention to the visual modality, and (iii) understanding if, even with increased attention to the visual modality, models might still concentrate on less relevant aspects of inputs, proving counterproductive for meme comprehension. Arguably, considering the recent progress in pre-trained large language models (PLMs) with the ability to analyze multimodal data, exemplified by MiniGPT-4 (Zhu et al., 2023), they could be contemplated for inclusion in the study. Nevertheless, challenges such as

hallucination (Li et al., 2023), mainly stemming from their longer average response length, pose concerns that we believe may have implications for tasks like hate meme detection.

9 Ethics Statement

Predicting whether a meme is hateful or not might infringe on the fundamental right of free speech if the prediction is used by a government or service provider to remove the post or block the posting user. If viewed from this perspective, it might be good news that the technology –as we show in this paper– barely works. On the other hand, not addressing hate speech would give further rise to possible discrimination, making it a problem for equal participation in any society. In terms of carbon emission, we conducted experiments primarily on GPUs to assess their resilience and develop countermeasure models. Using a private infrastructure with a carbon efficiency of 0.432 $kgCO_2eq/kWh$, we performed 120 hours of computation on 24 GB memory size Quadro RTX 6000 GPU. The total estimated emissions were 15.55 $kgCO_2eq$, with no direct offset (Lacoste et al., 2019).

References

Piush Aggarwal, Pranit Chawla, Mithun Das, Punyajoy Saha, Binny Mathew, Torsten Zesch, and Animesh Mukherjee. 2023. Hateproof: Are hateful meme detection systems really robust? In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 3734–3743, New York, NY, USA. Association for Computing Machinery.

Piush Aggarwal, Michelle Espranita Liman, Darina Gold, and Torsten Zesch. 2021. VL-BERT+: Detecting protected groups in hateful multimodal memes. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 207–214, Online. Association for Computational Linguistics.

Piush Aggarwal and Torsten Zesch. 2022. Analyzing the real vulnerability of hate speech detection systems against targeted intentional noise. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 230–242, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Raghad Alshalan and Hend Al-Khalifa. 2020. A deep learning approach for automatic hate speech detection in the saudi twittersphere. *Applied Sciences*, 10(23).

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. 2020. Gated multimodal networks. *Neural Computing and Applications*, 32(14):10209–10228.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. *UNITER: UNiversal Image-TExt Representation Learning*, page 104–120. Springer International Publishing.

Ian C. Covert, Scott Lundberg, and Su-In Lee. 2020. Understanding global feature contributions with additive importance measures. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Keyan Cuo, Wentai Zhao, Mu Jaden, Vishant Vishwamitra, Ziming Zhao, and Hongxin Hu. 2022. Understanding the generalizability of hateful memes detection models against covid-19-related hateful memes. In *International Conference on Machine Learning and Applications*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022). Association for Computational Linguistics*.

Darina Gold, Piush Aggarwal, and Torsten Zesch. 2021. Germemehate: A parallel dataset of german hateful memes translated from english. In *Multimodal Hate Speech Workshop 2021*, pages 1–6.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, AISec '18, page 2–12, New York, NY, USA. Association for Computing Machinery.

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

Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. 2022. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3651–3655, New York, NY, USA. Association for Computing Machinery.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *Social Informatics*, pages 49–66, Cham. Springer International Publishing.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

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.

Advances in Neural Information Processing Systems, 33:2611–2624.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization.

Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 26–35, Online. Association for Computational Linguistics.

Vasiliki Kougia and John Pavlopoulos. 2021. Multimodal or text? retrieval or BERT? benchmarking classifiers for the shared task on hateful memes. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 220–225, Online. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *ICML*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

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

Chunpeng Ma, Aili Shen, Hiyori Yoshikawa, Tomoya Iwakura, Daniel Beck, and Timothy Baldwin. 2021. On the (in)effectiveness of images for text classification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 42–48, Online. Association for Computational Linguistics.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18177–18186.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Niklas Muennighoff. 2020. Vilio: state-of-the-art visio-linguistic models applied to hateful memes. *arXiv preprint arXiv:2012.07788*.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

Benet Oriol Sàbat. 2019. Multimodal hate speech detection in memes. B.S. thesis, Universitat Politècnica de Catalunya.

Letitia Parcalabescu and Anette Frank. 2023. MM-SHAP: A performance-agnostic metric for measuring multimodal contributions in vision and language models & tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4032–4059, Toronto, Canada. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Andrii Samoshyn. 2020. Hate speech and offensive language dataset.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 5597–5606. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. *Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection*, pages 29–55. Springer International Publishing, Cham.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yue Wang, Jing Li, Michael Lyu, and Irwin King. 2020. Cross-media keyphrase prediction: A unified framework with multi-modality multi-head attention and image wordings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3311–3324, Online. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

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, San Diego, California. Association for Computational Linguistics.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*

and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6707–6723, Online. Association for Computational Linguistics.

Haoti Zhong, Hao Li, Anna Squicciarini, Sarah Rajtmajer, Christopher Griffin, David Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 3952–3958. AAAI Press.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290*.

A   List of Controversial Figures/Celebrities

Table 7 presents the comprehensive list of controversial figures and celebrities under consideration for the compilation of the confounder dataset.

B   Confounder Dataset - Instructions

Figure 4 illustrate the instructions manual provided to each annotator before starting the annotation process.

Preview of Instructions (PDF)

Here is a preview of the first page of the instructions document:

Guidance - Meme Annotation Tool

The intention of the annotation tool is to help people analyse and classify images—specifically memes—based on their context and content.

The annotation tool's user-friendly interface enables users to give each meme different labels and scores, making it easier to see trends, moods, and potentially harmful content.

The tool's annotations can be used to perform research, train machine learning models, or learn more about the traits of memes in a dataset.

Understanding the content of memes, their relationship to text, their stance on hate, and a number of other characteristics are made easier through this method.

Each choice is significant for the analysis that follows, which aims to determine whether classifiers can actually be trained to identify problematic figures in memes with neutral context.

Overview of the categories and options for annotation:

I. **Image-Text Relation:**

This category explores the connection between the accompanying text and the image in a meme. Users can choose from the following options:
- **Neutral**: The image and text have no particular relation to each other.
- **Needs Context**: The meme requires additional context to understand the relation between the image and text.
- **Text Supports Image**: The text provides context or enhances the meaning of the image.
- **Image Supports Text**: The image provides context or enhances the meaning of the text.

II. **Modality Towards Hate:**

This category examines the degree to which the meme expresses hate or offensive content. Users can select from the following options:
- **None**: The meme contains no hate or offensive content.
- **Text Supports Hate**: The hate or offensive content is primarily conveyed through the text.
- **Image Supports Hate**: The hate or offensive content is primarily conveyed through the image.
- **Text & Image Supports Hate**: The hate or offensive content is conveyed through both the text and the image.

III. **Decision Parts:**

Users can describe the specific tokens or elements in the meme that contribute to its hateful or non-hateful nature.
This helps in identifying the key components responsible for the content's overall categorization.

IV. **Hatefulness Scale**:

This scale allows users to rate the level of hateful or non-hateful content in the meme on a scale

Figure 4: Pdf preview of instruction manual.

Controversial Figures/Celebrities
Adolf Hitler
Anne Frank
Joseph Goebbels
Donald Trump
Nana Addo Dankwa Akufo-Addo
Barack Obama
Abu Bakr Al-Baghdadi
Joe Biden
Osama Bin Laden
King Charles
Prince Harry
Bill Clinton
Bill Cosby
BillGates
Chris Evans
James Franco
Pauline Hanson
Hassan Rouhani
Kamala Harris
Kevin Hart
George W. Bush
Hillary Clinton
Hulk Hogan
Stephen Hawking
Martin Luther King Jr.
Vince McMahon
Colin Koepernick
Melania Trump
Michelle Obama
Nadeschda Andrejewna Tolokonnikowa
Wladimir Putin
Ilhan Omar
Mike Pence
Bridget Powers
Pope Francis
Will Smith
Greta Thunberg
Justin Trudeau
Stevie Wonder
Darryl Worley
Caitlyn Jenner
Conchita Wurst
Mark Zuckerberg

Table 7: Illustrated the list of controversial figures and celebrities used in confounder dataset.