Understanding the Generalizability of Hateful Memes Detection Models Against COVID-19-related Hateful Memes

Keyan Guo
University at Buffalo
keyanguo@buffalo.edu

Wentai Zhao[§]
Northville High School
wentaizhao4@gmail.com

Jaden Mu[§]
East Chapel Hill High School
jaden.mu@gmail.com

Nishant Vishwamitra
University of Texas at San Antonio
nishant.vishwamitra@utsa.edu

Ziming Zhao
University at Buffalo
zimingzh@buffalo.edu

Hongxin Hu
University at Buffalo
hongxinh@buffalo.edu

***Abstract*—Multimodal vision and language (V&L) models have demonstrated promising results in the detection of hateful memes, a critical social problem. However, these models have two critical problems. First, these models are based on traditional datasets, so there is no guarantee that these models can generalize to also detect new types of hateful memes. Generalizability of these models is important since the topics of interest in online hate are rapidly changing. Since new categories of hateful memes are needed to study the generalizability of these models, this crucial aspect of these models is yet unexplored. Second, the impact of the relationship between the modalities on the detection of hateful memes has not been sufficiently studied. Hateful meme detection is a contextual task, and the relationship between the image and text inputs should be considered for their prediction. Studying this relationship is not straightforward since these models only provide prediction scores without an explanation. In this work, we propose three novel studies to understand V&L models' generalizability for new types of hateful memes and the relationship between visual and textual modalities during prediction. To study generalizability, we use COVID-19 as a case study and conduct a large-scale measurement of state-of-the-art models to investigate their generalizability on COVID-19-related hateful memes. Furthermore, we conduct two novel experiments to understand the relationship between visual and textual modalities. Using gradient-based explanation techniques, we explore the importance attributed to each modality. Then, using ablations of the modalities, we scrutinize the impacts of textual and visual inputs on the models' prediction. Our studies show that these models are not generalizable to new types of hateful memes (average F1-score of only 0.2), give significantly more importance to the visual modality during prediction (about 2X higher average importance than textual modality), and that the textual modality actually provides significantly greater valid information to detect new types of hateful memes.**

***Index Terms*—online hate, hateful memes, COVID-19 pandemic, vision and language models**

I. Introduction

Memes, often represented as images with embedded text descriptions, are widely used on social media for their humor and convenience [12], [20]. However, as memes have grown in popularity, hateful memes have emerged as a tool used by perpetrators to propagate online hate [20]. To call more attention to this societal and psychological issue, Facebook AI recently announced a public challenge on hateful memes and released a hateful meme dataset with some preliminary classification solutions [21]. As a result, many approaches have delivered promising results in hateful memes detection [23], [27], [37], [40]. We also find that multimodal models, primarily vision and language (V&L) models such as MMBT [19] and VisualBERT [24], have been majorly used to predict hateful memes.

Although the results of the hateful meme detection studies are uplifting, we note two potential problems remaining in these works. *First*, the topics of interest of online hate are not always set in stone [34]. As the world we live in is changing rapidly, new events are constantly emerging, creating an ever-changing cycle of online hate on the Internet. However, the existing works [23], [27], [37], [40], [41] that attempt to address the problem of hateful memes predominantly focus on the detection of existing hateful memes using traditional datasets [20], overlooking the generalizability of their methods when different types of hateful memes are brought forth. A key challenge to studying this problem is a lack of new datasets of hateful memes, and there is an urgent need for collecting such new datasets. *Second*, the impact of the relationship between visual and textual modalities in these pre-trained V&L models' on the detection of hateful memes have not been explored. The task of hateful memes detection is a contextual task, wherein the decision of a meme being hateful or not depends on both image and text modalities. However, the existing works [20] predominantly focus on improving detection efficiency, while ignoring the study of the impact of this relationship that is crucial for this contextual task. Since the pre-trained V&L models only provide hateful scores, and no explanation for this score is provided, understanding the visual and textual modalities' relationship during the classification is a crucial challenge.

The COVID-19 pandemic has recently emerged as one of society's most critical challenges. During the pandemic, large numbers of COVID-19-related online hate were engen-

[§]Work done during internship at University at Buffalo.

(a) Anti-Asian sample (b) Mask-related sample

Fig. 1: Examples of COVID-19-related hateful memes.

dered [14], [35], [36], [42] which caused great distress to the society. An important aspect of the online hate spread during COVID-19 was the significant role of hateful memes featured in the propagation of this hate. In this work, we exploit the opportunity presented by COVID-19 and utilize it as a case study to study the problem of generalizability of pre-trained V&L models, against new types of hateful memes found during the pandemic, and shed light on the relationship between the input modalities. We collect a novel COVID-19-related hateful memes dataset based on six diverse categories from Twitter based on current events during the pandemic. Our dataset consists of 5,000 memes overall, with 1,340 memes categorized as hateful and 3,660 memes categorized as non-hateful. A few samples from our dataset are depicted in Figure 1, which shows hateful memes related to anti-Asian hate (Figure 1 (a)) and mask-related hate (Figure 1 (b)). Then, we propose three research questions based on our dataset to address the problems of generalizability of detection models and the relationship between modalities in these models: **RQ1**: Can the pre-trained V&L models generalize to the task of detecting new types of hateful memes, such as COVID-19-related hateful memes? **RQ2**: Do these V&L models attribute similar importance to both modalities for predicting a meme as hateful? **RQ3**: Do both modalities provide similar valid information to the V&L models for predicting a meme as hateful? We conduct experiments involving state-of-the-art pre-trained V&L models such as MMBT [19], VisualBERT [24] and our COVID-19-related hateful memes dataset to address these research questions. For RQ1, we perform a large-scale measurement to investigate if the V&L models can yield promising results when detecting emerging hateful memes, and find that they cannot generalize to the COVID-19-related hateful memes, indicated by a low average F1-score of 0.2. For RQ2, we use gradient-based explanation techniques [33] to study the contribution of each modality to the model's decision by comparing the gradients from both visual and textual embeddings, and find that the visual modality is attributed much more importance (about 2X times) than the textual modality. For RQ3, we implement ablation studies to explore whether visual or textual inputs provide similar valid information for the model's decisions about hateful memes, and find that the textual inputs provide far greater valid information for the model's predictions than visual inputs.

II. Background and Related Work

Our work relates to hateful memes detection during the COVID-19 pandemic and the study of V&L models via AI explanation methods. In this section, we review related work in these areas.

A. Hateful Memes Detection

The prevalence of social media platforms facilitating image sharing has sparked emerging research on image-based memes. For example, a recent study [9] proposed a pipeline to detect the history of meme transformation. Another study [39] proposed a pipeline to use unsupervised methods to cluster images from social media and then use Know Your Meme [1] data to identify which of the clusters contained memes. They modeled disseminating memes between several online communities, showing that fringe communities like 4chan's Political Incorrect board (/pol/) are influential meme disseminators. Using the same pipeline to classify online political memes, a recent research study [8] used graph learning to create an evolutionary tree of memes. Another recent study [13] extracted features of memes using a pre-trained deep neural network and optical character recognition and demonstrated that the extracted features could help predict meme virality. Another work [12] uses both visual and textual features from memes to identify them and also link the text content to demographic information about users. Finally, the Facebook AI team created a hateful memes dataset to help build systems that better understand V&L hate speech [20]. However, these studies do not address the problem of the generalizability of hateful meme detection models. Furthermore, all of these existing works overlook the relationship between the image and text modalities on the models' prediction. A recent study [22] suggests that the hateful memes dataset shared by Facebook AI may lack practicality for "memes in the wild". In our work, we shed light on the core reason of this problem, *i.e.*, the lack of generalizability, and offer insights into the reasons behind this problem.

B. Explainable Vision and Language Models

Deep neural networks have empowered Artificial Intelligence techniques to achieve much success in many application domains. And with several vision and language fusion methods employed by deep learning models, it has been demonstrated that they can obtain impressive performance in multimodal tasks, especially in V&L tasks [17], [28], [30]. However, these black-box V&L models with the complex structure and processes of deep neural networks make them opaque and difficult to interpret or understand the internal state and decision rationales [7]. The topic of explainable V&L models has gained attention and study from some researchers [10], [15], [25], [29]. Cao et al. [10] designed a set of probing tasks and corresponding experiments to analyze how pre-trained V&L models drive success in vision and language tasks. Parcalabescu et al. [29] investigated the reasoning ability of pre-trained V&L models through a quantitative study. Similarly, Li et al. [25] performed another quantitative study with attention heads from pre-trained visual and linguistic models to

evaluate their semantic grounding capabilities. Frank et al. [15] proposed a diagnostic method based on cross-modal input ablation on image-text matching tasks to explore the effectiveness of pre-trained V&L models in integrating information across modalities. Hee et al. [16] first emphasized explaining V&L models under the hateful meme detection task. However, they focused on understanding the learning process when the models were training for the hateful meme classification task. In our paper, different from the previous works, we aim to investigate the generalizability of pre-trained V&L models against the new types of hateful memes and the relationship between input modalities using explanation techniques.

III. Experiments

In this section, we first present the methods of how we collect and annotate the COVID-19-related hateful meme dataset. Then we discuss the details of each implementation setup in our experiments. We next explain the experiment process and analyze the corresponding results for each proposed research question.

A. Dataset

To collect memes related to COVID-19, we first needed a set of hashtags to search on Twitter. To this end, we compiled a set of COVID-19-related hashtags, which we then used to collect Twitter memes. We started with an initial set of manually compiled 93 hashtags that we found prevalent during the COVID-19 pandemic [2]–[5]. Then, we used an open-source API [1] to collect tweets embedded with images shared during the month of August 2020 based on these initial hashtags. Next, we supplemented our initial set of hashtags with additional hashtags found in those collected tweets. We repeated this process for all months from February 2020 to April 2021 to finally compile a total of 195 hashtags. Our final list consisted of COVID-19-related hashtags such as *covidiots*, *ChinaVirus*, *americafirst*, *WearAMask*, *trump2020* and *COVID19Vaccine*.

Next, we used certain criteria to exclude memes that were invalid. First, we restricted our dataset to consist of only those memes that are in English. Since we focus on multimodal memes [20] (i.e., images with superimposed text), we first removed memes that did not have any text in them. We also removed memes with very long text ($>$ 30 words) to exclude screenshots and news articles, using an open-source tool Tesseract [18]. Then, we excluded those memes that did not have any image-based content (or Regions of Interest) in them (i.e., just plain background images) using the YOLO object detector [31]. Next, we removed duplicated memes. Finally, we were left with 114,064 valid memes in our dataset.

Since COVID-19-related hate speech is a new phenomenon, we could not use online annotators to annotate our dataset accurately. Therefore, we annotated a randomly selected subset of the dataset with 5,000 memes ourselves. We developed a rigorous annotation process to establish the ground truth of memes based on the meme's content. In our annotation scheme, we annotated any meme as hateful, that is: (1) directed towards an individual or a group of people, organization, or country, and (2) attacks victims using violent or dehumanizing speech, scandalization of personal appearance, propagates harmful stereotypes or misrepresentation, makes statements of inferiority, expressions of contempt, disgust or dismissal, cursing and calls for exclusion or segregation [6], [19].

We annotated the memes in two rounds. In the first round, the annotators independently labeled a set of randomly selected 200 memes. After independent labeling, the annotators resolved disagreements and updated the labeling guidelines based on the discussions. This resulted in the final annotation criteria presented above. In the second round, a subset of randomly selected 5,000 memes was annotated independently by all annotators. This round led to a near-perfect inter-rater agreement. Overall, 1,340 memes were annotated as hateful and 3,660 memes were annotated as non-hateful. And we use this annotated dataset for the experiments we designed.

B. Experiment Settings

To address the proposed research questions, we measure the capability of three state-of-the-art, pre-trained V&L models, Multimodal bitransformer (MMBT) [19], VisualBERT [24] and VisualBERT-COCO [26] (*i.e.*, VisualBERT trained on Microsoft's Common Objects in Context dataset) against hateful memes from our COVID-19-related memes dataset. These models represent some of the best performing V&L models used for diverse V&L tasks, such as Visual Question Answering (VQA), Referring Expressions (RE), and Image Captioning, in addition to hateful memes detection. Furthermore, these models use BERT [11] as the text encoder and use the features from the "fc6" layer of Faster-RCNN [32] with ResNeXt-152 [38] as the image encoder. For our experiments, we deploy the MMBT, VisualBERT, and VisualBERT-COCO models pre-trained by Facebook AI [21] to explore the three proposed research questions.

C. Measurement of State-of-the-Art Hateful Memes Detectors

In this experiment, we aim to understand the generalizability of the pre-trained V&L models against new types of hateful memes. Specifically, we investigate how well each pre-trained V&L model is at detecting COVID-19-related hateful memes.

With deploying the three aforementioned models, MMBT [19], ViusalBERT [24] and VisualBERT-CoCo [26], for each meme in the COVID-19-related hate memes dataset, we record the model's prediction of whether it was hateful or not and compare the results to the original annotated labels. To comprehensively measure and explain the performance of each model, we calculate the Precision, Recall, and F1 score and the total accuracy for the test dataset separately.

From our results depicted in TABLE I, we note that all three models *fail to satisfactorily generalize* to new types of hateful memes in the COVID-19-related memes dataset. Among the three models, VisualBERT-COCO has the highest Precision (0.48), followed by VisualBERT (0.35) and MMBT (0.33).

[1] https://github.com/tweepy/tweepy

Model	Precision	Recall	F1-score	Accuracy
MMBT	0.33	0.32	0.33	0.7
VisualBERT	0.35	0.15	0.21	0.7
VisualBERT-COCO	0.48	0.03	0.05	0.74

TABLE I: Prediction results of pre-trained V&L models for COVID-19-related hateful memes

On the other hand, we observed that for the Recall and F1-score, MMBT offers the highest performance, followed by VisualBERT and VisualBERT-COCO. For example, MMBT achieves the highest recall (0.32), followed by VisualBERT (0.15) and VisualBERT-COCO (0.03). However, the table shows that the F1-score is consistently low for all three models with, indicating that none of these models are capable of generalizing to new types of hateful memes. It should also be noted that although the accuracy for each modality is close to the results in Facebook AI's paper [21], this does not indicate that the models are able to generalize on new types of hateful memes, since the COVID-19-related hateful memes dataset is unbalanced in nature, which means it contains far more non-hateful meme samples than hateful meme samples. Thus, this finding answers the RQ1, *i.e.,* the pre-trained V&L models are not generalizable against new types of hateful memes.

D. Studying Modality Attribution Using Gradient-based Explanation

To study the impact of the relationship between visual and textual modalities, we conduct an experiment to investigate the importance attributed to each modality by pre-trained V&L models. Existing works [33] have indicated that gradients can be representative of the importance attributed to a feature by a classifier. We use the same gradient-based method to represent the impact of each modality on the model output. Since a gradient represents the importance of an input feature, the summation of the gradients of a modality multiplied by its inputs gives us a modality's total importance to the model's output (*i.e.*, prediction score) computed using the following equation.

$$Importance = \sum_{i} \sum_{j} \frac{\partial y}{\partial A_{ij}^{k}} \tag{1}$$

where y is the gradient of the score, with respect to feature map activations A^k of a convolutional layer k, *i.e.* ∂A_{ij}^{k}. These gradients during backpropagation over the width i and height j dimensions to obtain the neuron importance weights.

We computed the average and the standard deviation of the importance attributed to each modality for the 1,340 COVID-19-related memes annotated as hateful in our dataset. As hateful memes usually consist of short sentences, some special tokens generated by the text encoder [11] such as [*CLS*] and [*SEP*] are used to split different sentences. Additionally, some empty tokens such as [*PAD*] are included in the text input to represent all sentences uniformly, as the sentences are of varying lengths. However, in our experiment, we removed all the gradients produced from such tokens so that the analysis only considers the tokens in the original input. We have depicted the results of this experiment in TABLE II. In all the models that we tested in our experiment, *the importance attributed to the visual modality significantly outweighed the textual modality*. These results are consistent with the findings of Hee et al. [16], which used a similar gradient-based method to measure the contributions of visual and textual modalities for traditional memes in the hateful meme dataset released by Facebook AI [21]. However, although all models put more emphasis on the visual modality for both Covid-19 hateful memes and traditional memes, we found significantly higher standard deviations for the contributions of both modalities for Covid-19 memes than for traditional memes. This suggests that the contribution of image and text modalities vary significantly across samples for Covid-19 hateful memes, allowing for variations in the relative contribution of the modalities, in contrast to the relatively constant contributions of each modality for traditional memes regardless of the sample. This experiment presents the answer for the RQ2, *i.e.,* the V&L models do not attribute similar importance to both modalities for predicting a meme as hateful.

E. Studying Valid Information in Input Modalities Using Ablation

In this study, we further investigate the importance of textual and visual inputs to the model's prediction by conducting an ablation study. In this experiment, we understand the impact of each modality on the final prediction score given by a model. We achieve this by presenting only one modality to the model at a time while blocking the other modality. To obtain textual-only predictions, we replaced the original image of each hateful meme with an image of a completely blocked image (*i.e.*, the image tensor is zeroed out) and passed this new image along with the original text into the models. To obtain the visual-only predictions, we replaced the text input with [*PAD*] tokens, which indicate empty word tokens, and passed this input and the original image into the models. By doing so, the prediction is made solely based on the other unmodified modality.

Our findings from this experiment are depicted in TABLE III. In every category for all the models, the *F1-score from the textual modality is greater than the visual modality*. This indicates that all three models perform better when only given the text from a meme compared to only the image. Furthermore, the pre-trained models of MMBT and VisualBERT-COCO cease to function correctly when only given an image, which is reflected by a row of zeros in their corresponding Visual section in the table. From these results, we can conclude that the textual modality carries more valid information than the visual modality. As our finding from the previous experiment III-D, the pre-trained V&L models pay more attention to

Model	Text Input		Visual Input	
	Average	Standard Deviation	Average	Standard Deviation
MMBT	113.82	184.53	1063.80	1749.15
VisualBERT	1378.62	1866.37	2354.34	2903.74
VisualBERT-COCO	775.82	1238.56	1452.42	2314.74

TABLE II: Gradients from text and visual modalities in pre-trained V&L models.

Model	Modality	Precision	Recall	F1-Score
MMBT	Textual	0.31	0.16	0.21
	Visual	0	0	0
VisualBERT	Textual	0.31	0.05	0.09
	Visual	0.23	0.01	0.02
VisualBERT-COCO	Textual	0.57	0.01	0.03
	Visual	0	0	0

TABLE III: Results of ablation study for different modalities in pre-trained V&L models.

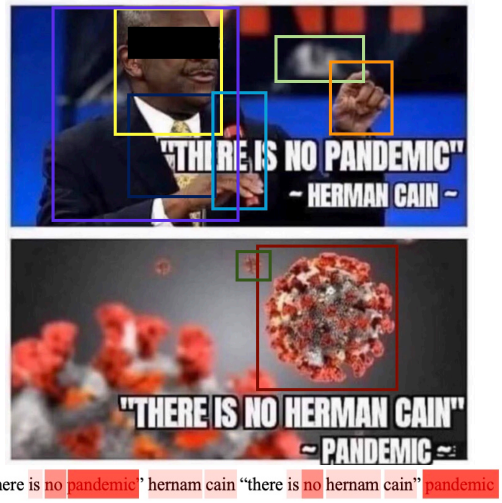

Fig. 2: Hateful meme sample depicting important regions and tokens in image and text modalities.

their visual modality. But from the ablation study, we note that the visual modality actually provides less helpful information for COVID-19 hate meme detection. *This observation introduces a distinct gap between the importance attributed by the model for each of the modalities and the actual contribution of valid information by modality.* For example, for the sample in Figure 2 annotated as hateful, our experiment shows that the pre-trained MMBT can predict it correctly with the textual-only input. But if we only give the image part, the model will not generate the ideal prediction. The gradients for textual and visual modalities are 7.7 and 66.84, respectively, which means the visual modality is attributed far more importance by the model for the prediction than the textual modality. Thus, we suspect that the excessive importance of visual modality could compromise the classification ability of the V&L models. However, that is not to say that the visual input is unimportant. When we compare our results from TABLE III to TABLE I, we notice that in general, all models would perform better when given both visual and textual modalities as opposed to only one of them. This supports the idea that both the textual and the visual inputs are necessary to enable optimal performance of V&L models, although the textual input contributes more to the prediction than the visual input for the new types of hateful memes. This experiment indicates the answer for the RQ3, *i.e.,* the visual and textual modalities do not provide similar valid information to the V&L models for predicting a meme as hateful.

IV. Discussion

A. Limitations

Our study has several limitations. First, we used a set of hashtags to collect our memes. Though the number of hashtags we used is quite large, the hashtags are by no means exhaustive and hence do not fully represent all the hashtags relevant to our categories. Second, due to the specificity of COVID-19-related hateful memes, we used a process of annotation that involved manually annotating a subset of memes. We note that larger studies involving more online participants would provide more insights into the annotation of new types of memes as hateful or non-hateful. In the future, we plan to train a large group of annotators for this task to collect more manually annotated memes. Third, we only focused on the memes shared in the English language and in a specific geographic location (North America). We acknowledge that the concept of online hate is varied among different geographies and languages, and our findings could be specific to a certain language and geography.

B. Future Work

For future work, we will conduct more experiments involving real-world users to shed more light and bring new insights into the annotation of new types of memes as hateful or not hateful. Furthermore, we will measure more V&L models against more "memes in the wild", such as the hateful memes recently witnessed during the 2022 Russian invasion of Ukraine to further confirm our studies and deepen our understanding. In addition, we will conduct further studies to investigate and narrow the gaps between the vision and language model's modality attribution and the amount of valid information from visual and textual modalities. We believe only if we fully understand these gaps can we improve the existing V&L models or build new V&L models with enough generalizability against new types of hateful memes.

V. Conclusion

In this work, we posed three research questions to understand the generalizability of V&L models and the impact of the relationship between the visual and textual modalities on these models for new types of hateful memes found during the COVID-19 pandemic. We collected and manually annotated a dataset of COVID-19-related memes from Twitter. Based on our dataset, we further analyzed the pre-trained V&L models with three experiments. Our experiments showed that the state-of-the-art pre-trained V&L models are significantly limited against the new types of hateful memes. Also, we found that the visual modality is attributed more importance during prediction than the textual modality by the V&L models. Our ablation study showed that the text input provides more valid information, which could help improve the models' prediction.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant No. 2129164, 2114982, 2228617, 2120369, 2226339, and 2037798.

REFERENCES

[1] Know Your Meme. https://knowyourmeme.com/. Accessed: 2022-09-19.

[2] 2020 Time Capsule #5: The 'Chinese Virus'. https://www.theatlantic.com/notes/2020/03/2020-time-capsule-5-the-chinese-virus/608260/, 2020.

[3] Coronavirus outbreak: What is "covidiots" trending on twitter? https://www.financialexpress.com/lifestyle/coronavirus-outbreak-what-is-covidiots-trending-on-twitter/1907432/, 2020.

[4] Trump plans to suspend immigration to u.s. https://www.nytimes.com/2020/04/20/us/politics/trump-immigration.html, 2020.

[5] Urban dictionary has a new word for coronavirus screw-ups. https://nypost.com/2020/03/24/urban-dictionary-has-a-new-word-for-coronavirus-screw-ups-covidiot/, 2020.

[6] Hate Speech. https://transparency.fb.com/policies/community-standards/hate-speech/?from=https%3A%2F%2Fm.facebook.com%2Fcommunitystandards%2Fhate_speech%2F&refsrc=deprecated, 2021.

[7] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[8] David M Beskow, Sumeet Kumar, and Kathleen M Carley. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management*, 57(2):102170, 2020.

[9] Aparna Bharati, Daniel Moreira, Joel Brogan, Patricia Hale, Kevin Bowyer, Patrick Flynn, Anderson Rocha, and Walter Scheirer. Beyond pixels: Image provenance analysis leveraging metadata. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1692–1702. IEEE, 2019.

[10] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164, 2020.

[13] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. Memesequencer: Sparse matching for embedding image macros. In *Proceedings of the 2018 World Wide Web Conference*, pages 1225–1235, 2018.

[14] Emilio Ferrara, Stefano Cresci, and Luca Luceri. Misinformation, manipulation, and abuse on social media in the era of covid-19. *Journal of Computational Social Science*, 3(2):271–277, 2020.

[15] Stella Frank, Emanuele Bugliarello, and Desmond Elliott. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. *arXiv preprint arXiv:2109.04448*, 2021.

[16] Ming Shan Hee, Roy Ka-Wei Lee, and Wen-Haw Chong. On explaining multimodal hateful meme detection models. In *Proceedings of the ACM Web Conference 2022*, pages 3651–3655, 2022.

[17] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021.

[18] Anthony Kay. Tesseract: an open-source optical character recognition engine. *Linux Journal*, 2007(159):2, 2007.

[19] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. Supervised multimodal bitransformers for classifying images and text. *arXiv preprint arXiv:1909.02950*, 2019.

[20] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

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

[22] Hannah Rose Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. Memes in the wild: Assessing the generalizability of the hateful memes challenge dataset. *arXiv preprint arXiv:2107.04313*, 2021.

[23] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5138–5147, 2021.

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

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, 2020.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871*, 2020.

[28] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, 71:1183–1317, 2021.

[29] Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. *arXiv preprint arXiv:2012.12352*, 2020.

[30] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788, 2018.

[31] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[34] Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179, 2021.

[35] Joshua Uyheng and Kathleen M Carley. Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines. *Journal of computational social science*, 3(2):445–468, 2020.

[36] Nicolás Velásquez, R Leahy, N Johnson Restrepo, Yonatan Lupu, R Sear, N Gabriel, Omkant Jha, B Goldberg, and NF Johnson. Hate multiverse spreads malicious covid-19 content online beyond individual platform control. *arXiv preprint arXiv:2004.00673*, 2020.

[37] Riza Velioglu and Jewgeni Rose. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975*, 2020.

[38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[39] Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*, pages 188–202, 2018.

[40] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2021.

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

[42] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. *arXiv preprint arXiv:2005.12423*, 2020.