VL-BERT+: Detecting Protected Groups in Hateful Multimodal Memes

Piush Aggarwal, Michelle Espranita Liman, Darina Gold, and Torsten Zesch[*]
Language Technology Lab
University of Duisburg-Essen

Abstract

This paper describes our submission (winning solution for Task A) to the Shared Task on Hateful Meme Detection at WOAH 2021. We build our system on top of a state-of-the-art system for binary hateful meme classification that already uses image tags such as race, gender, and web entities. We add further metadata such as emotions and experiment with data augmentation techniques, as hateful instances are underrepresented in the data set.

1 Introduction

In this work, we present our submission to the Shared Task on Hateful Memes at WOAH 2021: Workshop on Online Abuse and Harms.[1] Detecting hateful memes that combine visual and textual elements is a relatively new task (Kiela et al., 2020). However, research can build on earlier work on the classification of hateful, abusive, or offending textual statements targeting individuals or groups based on gender, nationality, or sexual orientation (Basile et al., 2019; Burnap and Williams, 2014).

Shared Task Description We only tackle Task A, which is predicting fine-grained labels for protected categories that are attacked in the memes, namely Race, Disability, Religion, Nationality, and Sex. The memes are provided in a multi-label setting. Table 1 shows the label distribution of the provided data set.[2]

Labels	Train	Dev	%
None	5495	394	64.4
Religion	888	78	10.6
Race	801	59	9.4
Sex	552	44	6.5
Nationality	191	19	2.3
Disability	184	16	2.2
Race+Sex	66	4	0.8
Religion+Sex	52	2	0.6
Race+Religion	53	10	0.7
Nationality+Religion	38	3	0.4
Disability+Sex	36	4	0.4
Nationality+Race	52	2	0.6
Nationality+Religion	20	1	0.2
Disability+Race	16	1	0.2
Other	56	3	0.5
Total	8,500	640	100

Table 1: Overview of categories in WOAH 2021 data set. 'Other' refers to the remaining (very infrequent) instances annotated with different combinations of protected group labels.

Our System Our system is built on top of the winning system (Zhu, 2020) of the Hateful Memes Challenge (Kiela et al., 2020), which was a binary hateful meme detection task. Zhu (2020) fine-tuned a visual-linguistic transformer-based pre-trained model called *VL-BERT*$_{LARGE}$ and showed that metadata information of meme images such as race, gender, and web entity tags (recommended textual tags for the image based on data collected from the web) improved the performance of the hateful meme classification system. We replicate this system for a more fine-grained categorization of hateful memes, as proposed by the current shared task. Considering the data scarcity in this novel task, we also propose several data augmentation strategies and examine the effects on our classification problem. The evaluation metric used by the shared task is the (micro-averaged) area under the receiver operating characteristic curve *AUROC*.

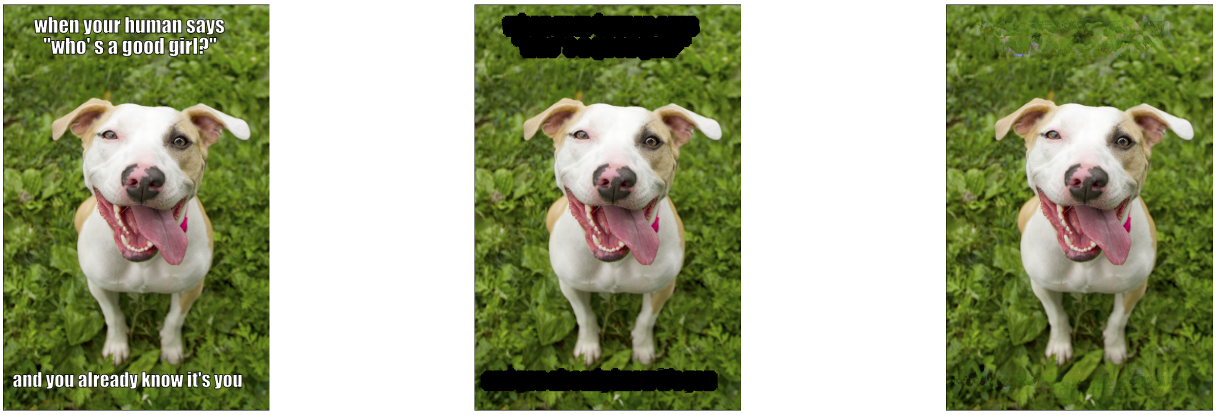
In addition, we consider **emotion tags** which are extracted from facial expressions available in the

[*] Equal contribution of the first two authors

[1] https://www.workshoponlineabuse.com/cfp/shared-task-on-hateful-memes

[2] In the data set, memes are labeled as pc_empty if they are not hateful and none of the protected categories can be applied. In this paper, we use None instead of pc_empty for better intuition.

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 207–214
August 6, 2021. ©2021 Association for Computational Linguistics



(a) (b) (c)

Figure 1: Image pre-processing: Recovering the original image of the meme (a) Original meme image (b) EasyOCR masking (c) Image inpainting

meme images. Based on experimental results and the shared task leaderboard scores, the inclusion of emotion tags along with VL-BERT$_{\text{LARGE}}$ model equipped with race, gender, and web entity tags exhibits the best performance for Task A. We make our source code publicly available.[3]

2 Related Work

Multi-modal hateful meme detection is the task of identifying hate in the combination of textual and visual information.

Textual Information In most previous works, hate speech detection has been performed solely in textual form. Despite many challenges (Vidgen et al., 2019), there have been several automatic detection systems developed to filter hateful statements (Waseem et al., 2017; Benikova et al., 2017; Wiegand et al., 2018; Kumar et al., 2018; Nobata et al., 2016; Aggarwal et al., 2019). One state-of-the-art model is BERT (Devlin et al., 2019). BERT is a contextualized transformer (Vaswani et al., 2017) based on a pre-trained language model which can be further fine-tuned for downstream applications such as hate speech classification.

Visual Information For hateful meme classification, the Facebook challenge team[4] proposed a unimodal training where a ResNet (He et al., 2015) encoder is used for image feature extraction. Apart from this, there has been a plenitude of work on extracting information from images, which is potentially useful for hateful meme detection. Image processing systems such as Faster R-CNN or Inception V3 models (Ren et al., 2016; Szegedy et al., 2015) are useful for detecting available objects in images. Smith (2007) and EasyOCR[5] can optically recognize the text embedded in an image.

Visual-linguistic Information There have been several ML-based approaches to solve the task of hateful meme detection. Blandfort et al. (2018) extracted textual features such as n-grams, affine dictionary along with local (Faster R-CNN) and global (Inception V3) visual features to train the SVM-based classification model. Sabat et al. (2019) proposed the fusion of vgg16 Convolutional Neural Network (Simonyan and Zisserman, 2015) based image features with BERT (Devlin et al., 2019) based contextualized text features to train a Multi-Layer Perceptron (MLP) based model. Earlier work (Liu et al., 2018; Gomez et al., 2019) proposed either early or late fusion strategies for the integration of textual and visual feature vectors. However, Chen et al. (2020); Li et al. (2020); Su et al. (2020); van Aken et al. (2020) and Yu et al. (2021) extracted visual-linguistic relationships by introducing cross-attention networks between textual transformers and transformers trained on visual features. Such networks deliver promising results on a variety of visual-linguistic tasks such as Image Captioning, Visual Question Reasoning (VQR), and Visual Commonsense Reasoning (VCR). Zhu (2020) and Lippe et al. (2020) exploited these networks for the binary classification of memes as hateful or non-hateful. The incorporation of additional metadata information as race, gender, and

[3] https://github.com/aggarwalpiush/HateMemeDetection

[4] https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/

[5] https://github.com/JaidedAI/EasyOCR

web entity tags, which are extracted from meme images, increased performance significantly in hateful meme classification (Zhu, 2020).

Hitherto, meme classification, having been introduced only recently, has been a binary task. Except for the VisualBERT (Li et al., 2019) based baseline[6] provided by the WOAH 2021 Shared Task, to our knowledge, there has been no work on detecting protected groups in hateful memes.

3 System Description

In this paper, we exploit the analysis proposed by Zhu (2020) for the fine-grained categorization of hateful memes.

3.1 Pre-processing

Both the visual and the textual parts of the memes are pre-processed. The data provided by the shared task consist of memes with their corresponding meme text. In this paper, we follow the steps proposed by Zhu (2020) to pre-process the provided input memes.

Text Pre-processing For text pre-processing, a BERT-based tokenizer (Devlin et al., 2019) is applied. This is also an integral part of the VL-BERT$_{\text{LARGE}}$ system (Su et al., 2020) (see Section 3.3).

Image Pre-processing The image part of the memes poses several challenges. First, meme images may consist of multiple sub-images, so-called *patches*. In this case, we segregate these patches using an image processing toolkit (Chen et al., 2019). Second, the text embedded in the images may add noise to the image features. Therefore, we aim to recover the original meme image before the text was added. To do so, we first apply EasyOCR-based Optical Character Recognition, which results in an image with black masked regions corresponding to the meme text as shown in Figure 1b. Then, *inpainting*, a process where damaged, deteriorating, or missing parts are filled in to present a complete image, is applied to these regions using the MMediting Tool (Contributors, 2020) (see Figure 1c).

3.2 Metadata

Understanding memes often requires implicit knowledge (e.g. cultural prejudice, clichés, historical knowledge) that human readers must have to understand the content. Such knowledge might be a big help for the classifier if explicitly provided. Zhu (2020) used meme image metadata, such as race, gender, and web entity tags to enhance binary classification performance on hateful memes. We utilized the same metadata and, in addition to that, emotion tags for the fine-grained categorization into protected groups.

Race and Gender We apply the pre-trained FairFace (Karkkainen and Joo, 2021) model to the provided meme images to extract the bounding boxes of detected faces with their corresponding race and gender metadata.

Web Entities Web entities are web-recommended textual tags associated with an image. They add contextual information to the images, making it easier for the model to establish the relationship between the meme text and image. We use Google's Web Entity Detection service[7] to extract these web entities.

Emotion Emotions are promising features for hate speech detection (Martins et al., 2018). Awal et al. (2021) investigated the positive impact of emotions in textual hate speech detection where emotion features are shared using a multi-task learning network. We exploit this in our system by extracting emotions based on facial expressions available in the meme image together with their corresponding bounding boxes. For this purpose, we use the Python-based emotion detection API[8] which classifies a face into the seven universal emotions described by Ekman (1992)—ANGER, FEAR, DISGUST, HAPPINESS, SADNESS, SURPRISE, and CONTEMPT.

3.3 VL-BERT$_{\text{LARGE}}$

VL-BERT$_{\text{LARGE}}$ (Su et al., 2020) demonstrates state-of-the-art performance on binary hateful meme classification Zhu (2020). Therefore, we investigate it for the detection of protected groups in hateful memes. VL-BERT$_{\text{LARGE}}$ is a transformer (Vaswani et al., 2017) back-boned visual-linguistic model pre-trained on the Conceptual Captions data set (Sharma et al., 2018) and some other text corpora (Zhu et al., 2015). It provides generic representations for visual-linguistic downstream tasks.

[6] https://github.com/facebookresearch/mmf/tree/master/projects/hateful_memes/fine_grained

[7] https://cloud.google.com/vision/docs/detecting-web

[8] https://pypi.org/project/facial-emotion-recognition

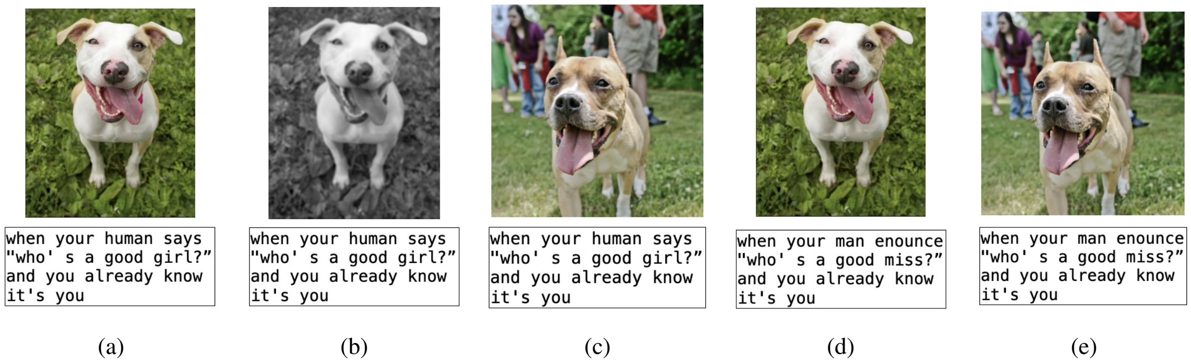

when your human says "who' s a good girl?" and you already know it's you
when your human says "who' s a good girl?" and you already know it's you
when your human says "who' s a good girl?" and you already know it's you
when your man enounce "who' s a good miss?" and you already know it's you
when your man enounce "who' s a good miss?" and you already know it's you
(a)
(b)
(c)
(d)
(e)

Figure 2: Data augmentation: (a) Original meme (b) Image augmentation with effects (c) Image augmentation with a visually similar image (d) Text augmentation (e) Image and text augmentation

One of the model training requirements is to identify objects and their location in the image. To do that, we use Google's Inception V2 Object Detection model.[9]

We extract features from both modalities (image and text) in the provided data set to fine-tune the pre-trained VL-BERT$_{\text{LARGE}}$ representation. Afterward, these features are used to train a multi-layer feedforward network (also called a downstream network) to generate the final classifier. We train the model for a maximum of 10 epochs with the other default hyperparameters provided by Su et al. (2020).

3.4 Data Augmentation

Data scarcity often leads to model overfitting. As shown in the training set distribution in Table 1, non-hateful memes comprise the majority of the data set. The non-uniform distribution of labels makes this data set quite small for model training. Therefore, we artificially augment the samples labeled with the protected groups. For image augmentation, we use the image augmentation toolkit by Jung et al. (2020) which alters images by adding effects like blur, noise, hue/saturation changes, etc. Additionally, we use Google's Web Entity Detection service to obtain visually similar images. For text augmentation, we generate semantically related statements using *nlpaug* (Ma, 2019). Furthermore, since we have original and augmented versions of images and texts, we combine them in three different ways: i) the original image with augmented text, ii) augmented image with the original text, and iii) augmented image with augmented text (see Figure 2).

3.5 Ensemble

The predictions of a single system may not be generalized enough to be used on unseen data due to high variance, bias, etc. However, relying on multiple systems can overcome these technical challenges. Therefore, we choose our best three systems based on their *AUROC* scores. We apply the majority voting scheme on the prediction labels provided by each system. The label with the highest number of votes will be selected as the final prediction for the ensemble system. In cases when all systems disagree, we choose the label with the highest prediction probability.

4 Results and Discussion

Table 2 shows the results for Task A on the provided development data set. We also compare our results with the VisualBERT (Li et al., 2019) based baseline as provided by the shared task organizers. Among the different configurations of our system, VL-BERT$_{\text{LARGE}}$ model with race, gender, emotion, and web entity tags (called +W,RG,E in the table) achieves the best *AUROC* score. We find that the inclusion of emotion tags has a positive effect on the overall performance when compared to other systems. To analyze the statistical significance among the approaches, we apply the Bowker test (Bowker, 1948) on the contingency matrices created on the number of agreements and disagreements between the systems. To compensate for the chance significance, we apply the Bonferroni correction (Abdi, 2007) on p value. We find that approaches marked with * are statistically significant compared to the best-performing solution.

When the model is trained on the train set along with augmented data, hardly any significant performance improvement is encountered. This is

[9] https://tfhub.dev/google/faster_rcnn/openimages_v4/inception_resnet_v2/1

		Protected Groups						Overall		Leader Board
Approach	sign.	RACE	SEX	REL.	DIS.	NAT.	NONE	F1	AUROC	AUROC
Baseline	*	.71	.84	.75	.84	.70	.78	.62	.85	
+W		.79	.86	.87	.90	.92	.71	.64	.91	
+W,RG	–	.81	.87	.91	.91	.85	.80	.70	.92	.912
+W,E		.77	.85	.90	.89	.77	.75	.68	.91	
+W,RG,E		.76	.89	.91	.94	.81	.79	.70	.92	**.914**
U | +W	*	.81	.87	.90	.90	.91	.71	.60	.87	
U | +W,RG	*	.83	.88	.90	.91	.87	.74	.62	.90	
I | +W		.79	.86	.89	.93	.91	.74	.67	.91	
I | +W,RG		.81	.86	.91	.88	.88	.77	.68	.92	
T | +W		.75	.82	.90	.84	.83	.76	.70	.91	
T | +W,RG		.75	.86	.86	.91	.83	.78	.70	.90	
IT | +W	*	.72	.80	.89	.81	.87	.75	.70	.88	
IT | +W,RG	*	.77	.88	.83	.79	.84	.77	.68	.90	
Ensemble		.75	.89	.92	.93	.79	.80	.71	.92	

Table 2: Classification results of hateful memes target (protected groups) classes on provided development data set. Abbreviations are as follows: RG: Race and Gender, W: Web Entities, E: Emotion, T: Text Augmentation, I: Image Augmentation, IT: Image and Text Augmentation, and U: Undersampling. * denotes that the approach is significantly different from the best performing system (+W,RG,E)) using the Bowker significance test, considering $p < 0.004$ after Bonferroni correction.

contrary to our expectations. We analyze the approaches with image and text augmentation (IT| +W and IT| +W,RG) (statistically significant from the best-performing system) and found a notable increase in False Negative errors, especially for RELIGION.

During post-experiment analysis, we find that the predictions for DISABILITY and RELIGION labels are better compared to others when the model is at a low False Positive rate. However, NATIONALITY performs relatively well at a high False Positive rate (see Figure 3). From the confusion matrices (Table 3), we find that the number of False Negatives is dominant in all classes. We believe that class imbalance is responsible for this behavior. To verify this, we train models on the undersampled training data set and found significant improvement on labels with low sample size. However, we also find a huge performance drop on the NONE label.

For the final submission, we generate predictions on the test set using our two best-performing models based on their *AUROC* score — VL-BERT$_{\text{LARGE}}$ +W,RG,E (**winning solution**) and +W,RG (2^{nd} rank) (see Table 2 for Shared Task leaderboard scores).

5 Summary

In this paper, we presented our approach to identify and categorize attacked protected groups in hateful memes. We performed experiments using a visual-linguistic pre-trained model called VL-BERT$_{\text{LARGE}}$ along with metadata information extracted from the meme image and text. Results show that the inclusion of metadata helps to improve system performance. However, the final system still lacks a robust understanding of hateful memes targeting protected groups.

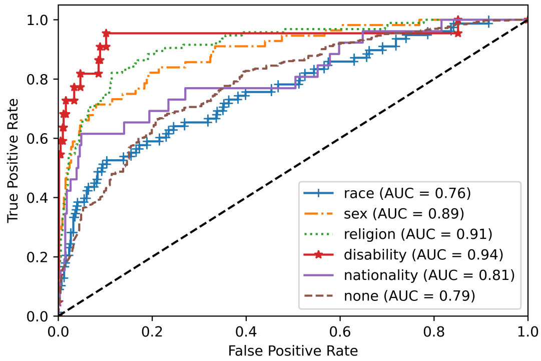


Figure 3: AUROC analysis for individual protected groups for configuration VL-BERT$_{\text{LARGE}}$ (+W,RG,E).

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

		Predictions		
		False	True	Total
Gold Values	False	531	14	545
	True	43	52	95
	Total	574	66	640

(a) RELIGION

		Predictions		
		False	True	Total
Gold Values	False	546	16	562
	True	56	22	78
	Total	602	38	640

(b) RACE

		Predictions		
		False	True	Total
Gold Values	False	608	6	614
	True	22	4	26
	Total	630	10	640

(c) NATIONALITY

		Predictions		
		False	True	Total
Gold Values	False	579	5	584
	True	37	19	56
	Total	616	24	640

(d) SEX

		Predictions		
		False	True	Total
Gold Values	False	617	1	618
	True	13	9	22
	Total	630	10	640

(e) DISABILITY

		Predictions		
		False	True	Total
Gold Values	False	108	138	246
	True	40	354	394
	Total	148	492	640

(f) NONE

Table 3: Confusion matrices for configuration VL-BERT$_{\text{LARGE}}$ (+W,RG,E).

References

Hervé Abdi. 2007. The bonferonni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3.

Piush Aggarwal, Tobias Horsmann, Michael Wojatzki, and Torsten Zesch. 2019. LTL-UDE at SemEval-2019 task 6: BERT and two-vote classification for categorizing offensiveness. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 678–682, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A. Gers. 2020. Visbert: Hidden-state visualizations for transformers.

Md Rabiul Awal, Rui Cao, Roy Ka-Wei Lee, and Sandra Mitrovic. 2021. Angrybert: Joint learning target and emotion for hate speech detection.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, 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 *13th International Workshop on Semantic Evaluation*, pages 54–63.

Darina Benikova, Michael Wojatzki, and Torsten Zesch. 2017. What does this imply? Examining the Impact of Implicitness on the Perception of Hate Speech. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 171 – 179, Berlin, Germany.

Philipp Blandfort, Desmond Patton, William R. Frey, Svebor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B. Gaskell, Rossano Schifanella, Kathleen McKeown, and Shih-Fu Chang. 2018. Multimodal social media analysis for gang violence prevention.

Albert H. Bowker. 1948. A test for symmetry in contingency tables. *Journal of the American Statistical Association*, 43(244):572–574. PMID: 18123073.

P. Burnap and M. Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. 2019. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

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.

MMEditing Contributors. 2020. Openmmlab editing estimation toolbox and benchmark. https://github.com/open-mmlab/mmediting.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2019. Exploring hate speech detection in multimodal publications.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

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. In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624, Virtual. Curran Associates, Inc.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11. Association for Computational Linguistics.

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.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes.

Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to combine modalities in multimodal deep learning.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Ricardo Martins, Marco Gomes, João Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. pages 61–66.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster r-cnn: Towards real-time object detection with region proposal networks.

Benet Oriol Sabat, Cristian Canton Ferrer, and Xavier Giro i Nieto. 2019. Hate speech in pixels: Detection of offensive memes towards automatic moderation.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

R. Smith. 2007. An overview of the tesseract ocr engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pretraining of generic visual-linguistic representations.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graph.

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

Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books.