Bias Detection Using Textual Representation of Multimedia Contents

Karthik L Nagar, Aditya Mohan Singh, Sowmya Rasipuram,
Roshni R Ramnani, **Milind Savagoankar**, **Anutosh Maitra**
{karthik.nagar, aditya.mohan.singh, sowmya.rasipuram,
roshni.r.ramnani, milind.savagaonkar, anutosh.maitra}@accenture.com

Abstract

The presence of biased and prejudicial content in social media has become a pressing concern, given its potential to inflict severe societal damage. Detecting and addressing such bias is imperative, as the rapid dissemination of skewed content has the capacity to disrupt social harmony. Advanced deep learning models are now paving the way for the automatic detection of bias in multimedia content with human-like accuracy. This paper focuses on identifying social bias in social media images. Toward this, we curated a Social Bias Image Dataset (SBID), consisting of 300 bias/no-bias images. The images contain both textual and visual information. We scientifically annotated the dataset for four different categories of bias. Our methodology involves generating a textual representation of the image content leveraging state-of-the-art models of optical character recognition (OCR), image captioning, and character attribute extraction. Initially, we performed fine-tuning on a Bidirectional Encoder Representations from Transformers (BERT) network to classify bias and no-bias, as well as on a Bidirectional Auto-Regressive Transformer (BART) network for bias categorization, utilizing an extensive textual corpus. Further, these networks were fine-tuned on the image dataset built by us - SBID. The experimental findings presented herein underscore the effectiveness of these models in identifying various forms of bias in social media images. We will also demonstrate their capacity to discern both explicit and implicit bias.

1 Introduction

Approximately 14 billion images are uploaded every day across social media platforms like Facebook. Instagram, Whatsapp, Snapchat[1]. Some of them contain harmful content that introduces or reinforces racial, gender, or cultural biases targeting certain groups of people. This can perpetuate stereotypes and normalize discrimination. Additionally, they cause a decline in the user experience and encourage fair-minded users to use less of or leave the particular social media platform. In recent times, memes have become popular across the internet in various forms, including images, videos, text, and other media. The analysis of memes has surged in popularity due to their emergence as a significant mode of online communication and cultural expression. Image-based memes frequently feature superimposed text and are characterized by humor, satire, or the conveyance of distinct cultural and social messages. Memes often follow a common format or template that can be readily adapted to diverse contexts, giving rise to a multitude of variations on a particular theme.

Diverse aspects of meme analysis, including the identification of hate (Kiela et al., 2020), emotion detection (Sharma et al., 2020), sarcasm (Kumar and Garg, 2019), and misogyny (Fersini et al., 2022) have been well studied. Several datasets are available as a part of SemEval - Semantic Evaluation workshop[2]. Extensive research papers have been published, encompassing both unimodal and multimodal approaches. Cutting-edge visual transformer models, such as VisualBERT, ViLBERT, MMBT have been employed[3]. These visual models performed well on visual linguistic tasks such as image captioning, visual question answering (VQA), etc., but failed when applied to multiple tasks on meme analysis (Pramanick et al., 2021). This might be because meme analysis involves associating text and image together for the prediction task. Moreover, memes are context-dependent, and thus focusing on text and image content alone is not sufficient. Secondly, unlike other tasks such as image captioning and VQA, text and image contents are most often uncorrelated.

[1] https://photutorial.com/photos-statistics/

[2] https://semeval.github.io/SemEval2023/

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

Figure 1: Four sample images from SBID dataset - *Social Bias Image Dataset* with their corresponding class labels. Images (a), (c) and (d) show the presence of implicit bias where the text and image are not directly related while their association changes meaning.

This study primarily aims to detect bias in images, notably memes. Several recent research studies have targeted the detection of bias within textual data in the context of hateful speech, social media posts, and the likes (Kiritchenko and Mohammad, 2018). Detection of bias in movie scripts has been studied by (Singh et al., 2022) where the authors introduced a novel dataset of movie scripts annotated for bias and the type of bias. Their annotation mechanism considered the context of the dialogue and also the previous utterance. Madaan et al. analyzed gender bias in Bollywood movies based on movie plots and movie posters (Madaan et al., 2017). The authors performed a detailed plot analysis by extracting male and female character mentions and images from Wikipedia movie pages. Jha et al. also analyzed Bollywood movie posters for gender bias from poster images (Jha, 2021). In some cases, gender stereotypes in a sentence are spotted by considering the presence of gender-specific stereotypical words (Sun et al., 2019), (Chaloner and Maldonado, 2019), however, they do not address contextuality or the implicit connotations for identifying potential prejudices.

While there has been substantial attention on analyzing memes and detecting bias in textual data, prior to this, the exploration of bias detection in multimedia contents has been overlooked. We aim to ascertain bias by combining the textual information with the visual elements in the image by generating a contextual text representation. A few sample images from the dataset indicating bias labels are given in Figure 1. Memes in Figure 1 require contextual knowledge to comprehend bias labels. Our focus in this work is on bias prediction and not on hate, where hate is considered a direct attack on a person or a group based on their characteristics. While memes demonstrating hate display the target community and opinions very clearly in a brutal manner as shown in Figure 2, biased memes are often difficult to interpret directly.

With the above motivation in mind, we aim to explore the detection of bias and the category of bias in multimodal memes. In particular, we make the following contributions.

- We propose a contextual text representation from multimodal content that utilizes appro-

Figure 2: Examples of hateful memes.

priate character attributes and captions.

- We fine-tuned transformer-based models for bias and category of bias prediction.
- We evaluated the proposed methodology on a small dataset consisting of 300 images demonstrating the presence of implicit bias.

The rest of the paper is structured as follows: Section 2 describes details about the curated dataset at length in terms of the annotation process and challenges faced. We also present details of textual datasets we used for fine-tuning in the same section. Section 3 presents our methodology to identify bias and the category of bias. Section 4 presents experimental results with discussion. We present a detailed analysis with example images in Section 5 and finally, we conclude in Section 6.

2 Dataset and Annotation

Bias, distinct from hatred or misogyny, possesses a subtle nature. In this context, we define bias as follows: a multimodal entity consisting of an image, optionally accompanied by embedded text, demonstrating partiality or prejudice towards or against an individual or group, often perceived as unfair. An annotated image dataset identifying bias is hard to get and harder to create because of the socio-cultural nuances associated with labeling multimodal content. It is relatively less difficult to label a textual dataset for bias. Various text datasets were curated in the past under different contexts, while an image-based dataset is not yet available to the best of our knowledge. Consequently, as part of our research, we have meticulously assembled a compact dataset comprising 300 images, known as the Social Bias Image Dataset (SBID). In this article, we present detailed annotations for SBID. We utilized a textually annotated dataset for training and fine-tuning classifier models.

2.1 Social Bias Image Dataset

We focus more on identifying bias from social media images. Images can be construed to contain conscious or unconscious bias depending on the socio-cultural context in which they are presented. The dataset included images with/without text overlaid and a single person appearing, thus signifying the presence of both textual and visual content. We considered images with a single person since associating bias in an image containing multiple people calls for a different modeling approach altogether. It will require additional computation to associate bias with the respective person.

Images are selected from two publicly available datasets namely, 1) Multimedia Automatic Misogyny Identification (MAMI) dataset (Fersini et al., 2022), and 2) Facebook Hateful meme dataset (Kiela et al., 2020). We also included a few images from Google Image Search with suitable keywords. Though both MAMI and Hateful meme datasets contain a vast number of images, not many were suitable for our bias prediction task. It must be noted that hate does not necessarily represent bias (and vice versa).

2.1.1 SBID Annotation

Two specialized annotators were brought in for the task of labeling biased images in SBID. The annotators came from two distinct social backgrounds with different life experiences. They were provided with a few basic guidelines before the annotation process. The Inter-Annotator Agreement was also verified with the kappa coefficient (McHugh, 2012). An average kappa score of 0.85 was obtained, indicating good agreement between the annotators. We considered images where both annotators agreed on bias and category of bias.

Table 1: Distribution of bias classes

Bias Category	No. of images
Gender	58
Race	47
Occupational	12
Other	33
No-bias	150

Annotators labeled images in two steps. In the first step, images were labeled as biased or no-bias class. In the second step, the annotators marked the sub-class of bias (gender, race, occupational, and other) in cases where a given image is biased. Table

1 gives the distribution of images under various bias categories. Images in categories other than race, gender, and occupation were fairly small and hence we merged them into the "Other" category.

2.1.2 Annotation process and challenges

Bias labeling is an intrinsically complex task and has to be done holistically i.e., annotators have to consider the implications of binding the text and image together. To be consistent with the approach, annotators were asked to observe for the presence of tagline or text on the image, and a single person appearing in the image. As expected, it was found that bias is not always present in an explicit manner. Many times, it is implicit and is well understood only if enough context is provided. For example, friends addressing each other with racial terms often are not considered to be biased but in a social context, the terms can make the image racially sensitive. The annotators handled these images with their own expertise, understanding of the context, and knowledge about world and social behaviors. Dataset has been split into 80% for training and 20% for testing respectively.

2.2 Textual Dataset Used for Model Fine-tuning

We used Hollywood Identity Bias Dataset (HIBD) (Singh et al., 2022) for fine-tuning the base model. HIBD provides annotated biases and stereotypes in the entertainment domain. This dataset consists of 35 movie scripts annotated for multiple biases like gender, race/ethnicity, occupation, and others that include religion, ageism, LGBTQ, personality, and body shaming. 27,558 labeled dialogues are made available, of which 976 dialogues were marked as biased. Each dialogue turn consists of a **Scene Description** that describes the scene for the given dialogue, **pre-context** derived from the previous 2 dialogues and **Dialogue** that describes the present utterance.

StereoSet (Nadeem et al., 2020) is a large-scale natural dataset created to measure stereotypical biases in four classes - gender, profession, race, and religion. This dataset was used to understand the presence of bias in popular language models like BERT, GPT2, ROBERTA. Our SBID dataset is more related to HIBD dataset with implicit examples, unlike StereoSet. Hence, we employed HIDB dataset for our fine-tuning. HIBD dataset is larger than StereoSet.

3 Methodology

Identification of bias is complicated since bias can cause harm even without using any explicit hateful content. Bias can be emotionally neutral and can inadvertently get introduced in the content for reasons those are personal and inoffensive. Often the contents posted on social media are sarcastic, and not all sarcasm necessarily are biased or stereotyped. It is also possible that the image posted is seemingly innocuous, but a tagline on the image or some other comments added to the image may make it offensive and harmful. The methodology we propose considers these specific tenets of bias and works on the principle of generating as holistic a description as possible of the content. We aim to create a contextual text representation of the image content by extracting text inscribed on the image, generating an image caption, and extracting character attributes of people in the image. We then use this representation in a two-stage framework; first a Bias Identification Model (BIM)) and then a Topic Classification Model (TCM) to determine the category of bias. The functional flow of our approach is shown in Figure 3. Both BIM and TCM are trained following a sequential adaptation strategy where the models are first fine-tuned on a large domain dataset before fine-tuning on SBID. For TCM to identify the topic of the textual representations, we trained a multi-class classification model that can identify gender, race, occupational, and other (age, religion, body shaming, LGBTQ, personality) bias. The individual functional blocks are explained below.

3.1 Generation of Contextual Text Representation from SBID

In this section, we describe different components that are used for capturing appropriate features from images.

3.1.1 OCR Extraction

We perform Optical Character Recognition (OCR) to extract the text overlaid on the image and we used this as *main-text*, often interchangeably referred to as punch line in our contextual representation. A well-known framework called PaddleOCR[4] has been used for this task. PaddleOCR by default uses the PP-OCRv3 model (Li et al., 2022) and works well even for text aligned in different directions. As the images were collected from many in-

[4]https://github.com/PaddlePaddle/PaddleOCR

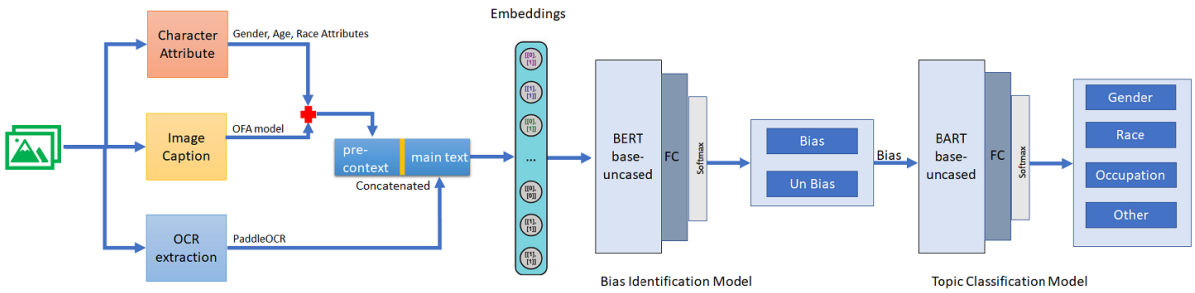

Figure 3: Functional Flow of the Proposed Model

the-wild datasets, the spacing between words was not consistent and resulted in the generation of incoherent text. The quality of the OCR is one of the very important components for the classifier model and hence additional scripting was performed to eliminate syntactic errors in OCR output.

3.1.2 Image Captioning

Image captioning is a process of recognizing the context of an image and creating an appropriate textual description of it. We employed OFA (Wang et al., 2022) framework in our experiments. OFA not only supports image captioning but also provides other abilities such as visual grounding, grounded captioning, image-text matching, and visual question answering. OFA uses a simple sequence-to-sequence learning approach and is pre-trained on 20M publicly available image-text pairs. We have used the OFA Large model which has 470M network parameters. Image captioning also can affect the classifier models if the captions do not encapsulate the right intent or actions.

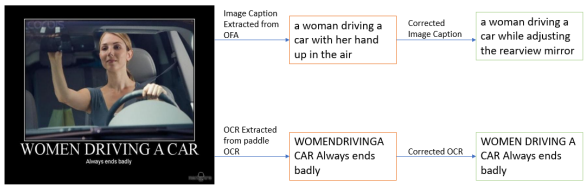

Figure 4: Example of Image caption and OCR correction

Figure 4 shows an example of incorrect outputs from OCR and image captioning where there is no space between words and the image caption generated was not capturing the correct intent. Additional scripting and a language model can correct most of these errors. Experimental results in the succeeding sections present the positive effect of caption correction on classifier accuracy.

3.1.3 Character Attribute Extraction

The attributes of the person appearing in the image are important in identifying prejudices. We extracted attributes such as age, gender, and race of the person in the image. We considered FaceNet[5] to detect the face. pre-trained SSR-Net[6] to detect gender and age and deepface[7] to extract race attribute.

Image caption along with the character attributes form the *pre-context* and is concatenated with *main-text* to create a textual representation of the image as shown in Figure 3. This representation is the input feature vector for the classifiers.

3.2 Bias Detection

Bias detection is performed in two stages - the first stage is the Bias Identification Model (BIM) for bias vs. no-bias prediction task and the second stage is a Topic Classification Model (TCM) to identify the category of bias for the cases found to be biased in the first stage.

3.2.1 Training BIM

BIM for bias vs. no-bias prediction is trained using a sequential adaption strategy. First, we fine-tuned the BERT model on HIBD text database and then fine-tuned on SBID. Images in SBID dataset are converted to contextual text representations as described in the previous section. The SBID dataset is split into train and test partitions with 240 and 60 images for training and testing respectively. We compute the performance of 1) HIBD model, and 2) SBID model on the SBID test split. As the HIBD dataset is a textual database labeled for bias classes, we used the model directly to test on 60 images in the test split.

[5]https://ieeexplore.ieee.org/document/7298682
[6]https://www.ijcai.org/proceedings/2018/150
[7]https://github.com/serengil/deepface

Table 2: Bias Identification and Topic Classification Model Abbreviations

BIM Models	Description	TCM Models	Description
Without Caption Correction			
BIM-HIBD-WC	BERT finetuned on HIBD	TCM-HIBD-WC	BART finetuned on HIBD
BIM-HIBD-SBID-WC	BIM-HIBD finetuned on SBID	TCM-HIBD-SBID-WC	TCM-HIBD finetuned on SBID-WC
With Caption Correction			
BIM-HIBD-C	BERT finetuned on HIBD	TCM-HIBD-C	BART finetuned on HIBD
BIM-HIBD-SBID-C	BIM-HIBD finetuned on SBID	TCM-HIBD-SBID-C	TCM-HIBD finetuned on SBID-C

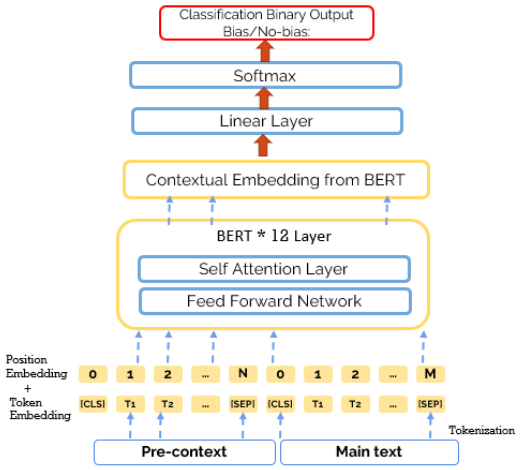

Figure 5: Fine-tuning BERT for Bias Identification

3.2.2 Training TCM

In the TCM stage, images falling into the bias class from BIM are further classified to identify the topic. TCM is modeled as a multi-class classification task to identify the topic belonging to gender, race, occupation, and others. As in the BIM, we fine-tune first on HIBD before fine-tuning on SBID. We employed BART (Lewis et al., 2019) for this task. As in BIM task, we evaluate the performance of both HIBD model and SBID model on SBID test data.

Fine-tuning BERT and BART models was done by adding additional dense layers on top as shown in Figure 5. As BERT and BART are language models trained on huge amounts of data, these models are expected to learn domain knowledge even from small amounts of training data. In our framework, OCR output is used as *main-text*, and image captions along with attributes is used as *pre-context*. For all our experiments, we used a simple language model to correct OCR output to take care of minimum errors like spacing.

4 Experimental Results and Discussion

An experimental evaluation of the proposed approach was performed on SBID test data. As mentioned previously, BIM and TCM were trained starting with HIBD. The approach presented an opportunity to understand the effectiveness of different sets of training data on model performance. We have listed the different experimental setups in Table 2. Each setup follows a notation in the form model-dataset(s) for training with or without caption correction C or WC. Captions play a very important role in fetching the right intent for the current task. Hence, we repeat experiments with captions generated by OFA model and captions corrected with the right intent.

4.1 Experimental results without caption correction

In this section, we demonstrate all results obtained with the captions generated by OFA model. Firstly, we perform fine-tuning of BERT and BART models on HIBD text database for BIM and TCM tasks respectively. The fine-tuned models are later used to test on SBID test set. As mentioned in Section II, HIBD dataset is annotated for bias labels for each dialogue. The dialogue serves as the $main-text$ for the model and is preceded by the $pre-context$. Table 3 presents fine-tuned model performance on SBID evaluation set for BIM and TCM tasks. The F1 score has been calculated for the task and similar has been done in Table 4. Evaluation results are presented using $pre-context$, $main-text$, and by merging both to form $full-text$ to understand the significance of each of the data streams.

It can be observed that $pre-context$, and $main-text$ together are important to understand bias. Meaningful usage of embedded text on the image alongside the captions and character attributes helps detect implicit references of the content. For both prediction tasks, improved performance is achieved when base models are fine-tuned on

image-based datasets.

Table 3: Performance of BIM and TCM tasks on SBID test set using text-based and image-based models. BIM - Bias Identification Model, TCM - Topic Classification Model, HIBD - Hollywood Identity Bias Dataset, SBID - Social Bias Image Dataset, WC - Without Caption Correction

Model	pre-context	main-text	Full text
BIM-HIBD-WC	0.42	0.73	**0.74**
BIM-HIBD-SBID-WC	0.53	0.80	**0.79**
TCM-HIBD-WC	0.63	0.80	**0.85**
TCM-HIBD-SBID-WC	0.70	0.81	**0.85**

4.2 Experimental results with caption correction

Capturing the right intent and context in caption generation systems is one of the major challenges (Ghandi et al., 2023). Several attention-based, graph-based networks were proposed in the past to handle the relations between objects and attributes correctly. The employed framework OFA is a task-agnostic and modality-agnostic framework that uses a transformer as the backbone architecture for an encoder-decoder network. OFA model outperformed image captioning tasks on well-established datasets. Image captions generated by the OFA model were rephrased using a language model. The language model has analyzed both the image's characteristics and the accompanying text to generate alternative captions.

Table 4 displays the fine-tuned model performance when image captions are corrected. It can be observed from Table 3 and 4 that the incorporated caption correction has consistently performed well as compared to models that used image captions without correction. As in the previous case, $main - text$ and $pre - context$ both play an important role in the prediction task.

Table 4: Performance of BIM and TCM tasks on SBID test set using text-based and image-based models. BIM - Bias Identification Model, TCM - Topic Classification Model, HIBD - Hollywood Identity Bias Dataset, SBID - Social Bias Image Dataset, C - With Caption Correction

Model	pre-context	main-text	Full text
BIM-HIBD-C	0.48	0.74	**0.76**
BIM-HIBD-SBID-C	0.59	0.80	**0.87**
TCM-HIBD-C	0.79	0.85	**0.87**
TCM-HIBD-SBID-C	0.74	0.81	**0.89**

5 Analysis

The caption generated by our OFA model is not the best fit for the current task. In most of the cases, the captions are general. Character attributes play a major role in associating the character with the image content. To get a deeper insight into the contextual text representation, we examine the classification results on the following images. The caption produced for an illustrative image depicted in Figure 6 is "a man standing in the grass holding a frisbee". The caption along with OCR "Fat can't hide" is an example of a biased image while the BIM model predicted as "no-bias". The annotators originally labeled the image as a biased image belonging to the body-shaming (other bias) category. OFA model failed to capture the attributes or characteristics of the person correctly. While the model made an accurate prediction with the caption obtained by the language model - "An obese man wearing a tank top and a fanny pack is standing in a field holding a frisbee".

Figure 6: Example of a biased image

When there is no text present in the image, the contextual representation relies solely on the image caption and character attributes for prediction. In the case of the image example depicted in Figure 1c, if the text on the image is removed, the image is unbiased. However, with the presence of text on the image, the image's meaning undergoes a complete transformation, rendering it a biased image, which our model is correctly capturing.

5.1 Error Analysis

While our model exhibited strong overall performance, it failed to capture the context in a few

(a) (b)

Figure 7: Instances of misclassifications from SBID dataset

instances. Figure 7a shows an example of an unbiased image case where our model inaccurately labeled as a biased image belonging to the "race" category. This might be due to the reason that the word "killing" in the text is inappropriately associated with the word "black man" in the caption. Figure 7b shows an example of a biased image while our model labeled it unbiased. This could be attributed to the implicit nature of the content, where the tech support roles are primarily managed by the South Asian community, who are considered not very fluent in English.

6 Conclusion & Future Work

Detection of bias is an inherently challenging problem since bias is often implicit and is associated with contextual and socio-cultural nuances. Seemingly innocuous visuals can portray a very different meaning when connected with a textual statement and vice versa. Towards this end, we attempted a transformer model-based approach that creates a fully textual representation of images with a description of contexts and then tries to classify the image. Further, a topic is also associated with the textual narration of the image to identify the category of bias. This approach had to be adopted as it was hard to obtain an acceptable bias-labeled image dataset. We also handcrafted a Social Bias Image Dataset (SBID) with 300 annotated images to fine-tune the transformer models which were essentially trained on textual data. Our results demonstrated that the models fine-tuned on the image dataset performed better as compared to using pre-trained text-based models in detecting bias in social media images. We expect the models to get stronger with SBID growing in size. The model performance today heavily depends on the effectiveness of the OCR system and the ability to generate succinct image captions. In future work, it is interesting to include images with multiple people and images with objects that resemble human form in the dataset. Images with multiple people in it will demand establishing additional relationships between them and further research can aim to address this issue. The OCR model output can improve with the incorporation of modern language models. A larger suitably annotated image dataset will also enable direct leveraging of multi-modal transformer models.

References

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. 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)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review.

Alok Jha. 2021. The representation of gender in bollywood film posters: A semiotic analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, India. Global Media Journal: Indian Edition.

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.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems.

Akshi Kumar and Geetanjali Garg. 2019. Sarc-m: sarcasm detection in typo-graphic memes. In *International Conference on Advances in Engineering Science Management & Technology (ICAESMT)-2019, Uttaranchal University, Dehradun, India*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system.

Nishtha Madaan, Sameep Mehta, Taneea S Agrawaal, Vrinda Malhotra, Aditi Aggarwal, and Mayank Saxena. 2017. Analyzing gender stereotyping in bollywood movies.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, pages 15–18.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Momenta: A multimodal framework for detecting harmful memes and their targets. *arXiv preprint arXiv:2109.05184*.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas Pykl, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis–the visuolingual metaphor! *arXiv preprint arXiv:2008.03781*.

Sandhya Singh, Prapti Roy, Nihar Sahoo, Niteesh Mallela, Himanshu Gupta, Pushpak Bhattacharyya, Milind Savagaonkar, Nidhi, Roshni Ramnani, Anutosh Maitra, and Shubhashis Sengupta. 2022. Hollywood identity bias dataset: A context oriented bias analysis of movie dialogues.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340, USA. PMLR, PMLR.