Multimodal Political Bias Identification and Neutralization

Cedric Bernard cedricb@vt.edu
Xavier Pleimling xavierp7@vt.edu
Amun Kharel akharel@vt.edu
Chase Vickery cdvickery@vt.edu

1. Problem Statement

In a study conducted in 2015, researchers analyzed approximately 150 million tweets from 3.8 million Twitter Users, each pertaining to political and nonpolitical issues. In this study, researchers observed that people who share the same ideologies when it comes to political topics exchange information with each other much more than those who share different ideologies [1]. Through empirical observation, the Internet and social media are the primary news distributors today. Since we distribute news on a public platform, social media either acts as a public sphere that can host a variety of opinions and information or an echo chamber that strengthens views that it is built upon [6]. We postulate that the most significant contributor to the presence of political echo chambers in social media is the presence of subjective bias or emotionally charged language in the texts. In addition, propaganda is widespread in the news through cleverly crafted political images. For a healthy exchange of ideas between people of different political persuasions in social media, political images, and text should adhere to Dahlberg's six essential qualities of the Public Sphere. The six qualities are namely: Reasoned Exchange of Problematic Validity Claims, Reflexivity, Ideal Role Taking, Sincerity, Formal Inclusion and Discursive Equality, and Autonomy from State and Corporate Power [7, 21]. We need to cultivate these essential qualities in the domain of news and news distribution by debiasing images and texts in the news to make them objective and neutral regardless of their political standing. In previous work on text debiasing in the context of news, researchers aimed to debias text by finding methods to make the text more neutral [18]. Text debiasing was done by removing subjective bias using the method provided by [20]: using Wikipedia's Neutral Point of View (NPOV) policy. However, for debiasing political images, the work performed on it is very limited and not well explored. We will perform cross-modal alignment of textual images and news to solve the problem of debiasing political images. Then we will select the image with the least bias using the method in [24]. In summary, our work attempts to learn an algorithm to minimize and neutralize political bias in news articles.

2. Related Works

2.1. Debiasing

Machine Learning and Deep Learning models are trained in a large set of texts and images from the real world. These images and texts may contain gender, cultural, religious, political, and other social biases [14]. Therefore, several works have been proposed to address the problem of biases in real-world models. Some previous work was centered around debiasing the sentence level representations, which removes the religious, gender, racial, and cultural biases [14, 15]. [14] in particular analyzes the performance of debiasing on sentence-level downstream tasks such as sentiment analysis, linguistic acceptability, and natural language understanding. Although numerous works have been done to analyze biases in pre-trained language models and vision models individually, only a few works have been done on a multi-modal setting [23]. In terms of these works, [23] does research to demonstrate gender bias in VL-BERT, which works in a multi-modal setting. After thoroughly searching the existing literature, we found that only a few prior works have been done to debias images and texts in a multi-modal setting. However, performing work on debiasing images and texts has several implications. For example, debiasing in a multi-modal setting would lead us to make political websites more unbiased in both images and texts. For political bias detection, [4] uses over 6900 news articles with labels derived from a website to develop a neural model for bias assessment, and [18] both analyzes subjective bias in text and neutralizes the subjective bias. There are two core problems with Political News, which are namely subjective or biased language and news, and selective news reporting. [18] is used to address the first issue of modifying subjective bias in a paper, giving an example where the paper converts the sentence "John McCain exposed as an unprincipled politician" to "John McCain described as an unprincipled politician," which changes the subjective tone to a more objective one. Similar to detecting bias in political bias in texts, [24] collects a dataset of over one million unique images and associated news articles from left- and right-leaning news sources to develop a method to predict the image's political leaning. [25] models a real-world sce-

nario where image-text pairs convey complementary information with little overlap. The approach used in [25] helps preserve the semantic relationships between paired images and paired text. We can generate a series of semantically aligned images from [25] given a text or news headline. The objective of our project is to remove the subjectivity of the news and replace the news images with semantically aligned images that are politically less biased.

2.2. Political Bias Identification (Images)

Regarding work related to political image debiasing, some existing papers focused on finding and predicting political bias within images. For example, [24] utilizes a two-step process where a model first learns relevant visual concepts of an image to enable bias prediction. Then, a visual classifier is trained upon that model. In terms of political image debiasing, [13] aims to evaluate and proposes metrics for measuring bias and bias augmentation in image text pairs, as well as how current models and datasets perpetuate and increase bias, but does not provide work on how to remove bias from image and text caption pairs. Additionally, [2] studies non-verbal bias indicators in facial features of political images and statistical analysis of news articles published around the 2016 election. Other works, such as [17], use facial feature extraction methods to identify and correlate certain visual elements with various kinds of bias. This work shows that different political figures are displayed quantifiably differently depending on political orientation and outlet. Finally, several approaches to other more general image bias-related tasks could be extended to political bias identification. For identifying bias in images, some work focuses on reducing bias to attribute/label data provided with or extracted from the data [11]. Other common approaches focus on learning from lower-dimensional representations or determining bias by cross-analyzing multiple datasets [11]. A more recent and uncommon approach utilizes an ensemble classification system by training a low-capacity network and a high-capacity network to reduce bias [5, 11]. Overall, while several methods allow for identifying bias in images, most of these methods are, to our knowledge, not applied or extended to the political domain. Moreover, some approaches are not viable to perform regardless since annotating and performing cross-analysis on multiple datasets can be far too computationally expensive.

2.3. Political Bias Identification (Text)

Multiple techniques have been used in previous works to identify and reduce bias in the text. [18] and [22] use pre-trained BERT transformers to locate and correct biased words. [27] and [16] use adversarial training to create a classifier to detect and correct hate speech. [12] uses an attention-based mechanism on the article headline network to detect bias in the article body to mirror the order in which humans read articles. [3] uses a Gaussian mixture model to observe probability distributions of the frequency, positions, and order of information to detect article-level bias. Regarding text dataset annotation, [9] proposes a framework to decompose gender biases across multiple dimensions, including the gender of the speaker, the person being spoken to, and the person being talked about. However, these previous approaches don't leverage or interact with available visual data or only focus on correcting specific types of bias, i.e., gender, politics, and hate speech. By also processing images available in the article, our approach can use the images to inform the text bias and vice versa, as well as substitute the biased images with semantically similar ones but evaluated with a lower amount of bias.

2.4. Evaluations/Metrics/Losses in Similar Approaches

Due to the varied approaches in bias detection and debiasing in related works, several different methods are also used for evaluations. Frequently, individual datasets are scraped and developed by separate groups for specific projects. This approach tends to result in datasets where articles, usually just represented through the text modality, have been manually labeled with one particular kind of bias [3, 12], lending the ability to use these labels in a supervised classification task which identifies an article or sentence as biased. Bias detection here can vary between a binary classification (biased vs. neutral) or multi-class (biased towards one of a set of groups). Other works use existing corpora to identify and potentially correct biased language [18, 26]. In most cases concerning bias detection with these datasets, a popular evaluation metric is to check the model's accuracy in classifying it as biased in binary or multi-class settings. When evaluating text debiasing methods, basic natural language metrics such as BLEU can be used [18]. Human evaluations are also crucial for understanding the effectiveness of debiasing models, with graders manually deciding on how fluent an original and corresponding debiased text are and whether they have the same meaning [18].

2.5. Politics Dataset

We will use the dataset collected in [24]. This dataset is collected from biased news sources (from left/right) on 20 politically contentious issues such as Abortion, Black Lives Matter, LGBT, Welfare, etc. This dataset has around 1.8 million images/articles in total. This paper uses crowdsourced annotations to annotate over 14,000 sets that contain bias labels for images and image-text pairs alongside other metadata. Various curation mechanisms have been used to clean up the data from news sources. Also, the crowdsourced annotations are curated with quality control. Although [24] identifies the bias as left/right-leaning, this

paper does not have labels for neutral news sources. Another dataset that we will use is the Wiki Neutrality Corpus (WNC). WNC contains 180,000 sentence pairs with biased and neutralized information, metadata, and additional contextual sentences, which were all scraped from Wikipedia edits that ensured texts were as unbiased as possible [18]. This dataset will neutralize the biases in the text of political news articles.

3. Proposed Approach

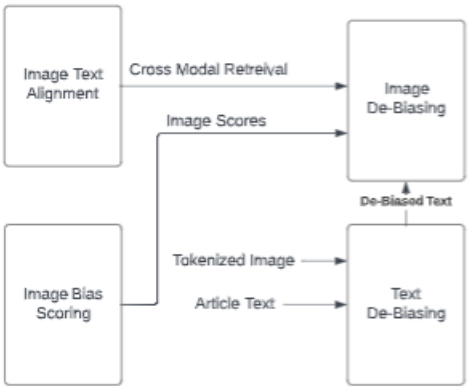

Figure 1. Overall Architecture

Figure 1 illustrates our overall architecture. Images and Texts that are used to train this model will be retrieved from various left-leaning, and right-leaning websites from the Politics Dataset [24]. 4 of us in the research team will label the websites from -1 to 1 to provide a bias score, where -1 refers to extremely left-leaning and 1 refers to extremely right-leaning. A score of 0 would be neutral. We will average the team members' scores to label the images and texts from the particular website from -1 to 1. Section 3.1 will identify biased words in the articles using the BERT [18] Biased Word Predictor. Then, we will train a BERT model with image and text inputs. Specifically, we will use Wikipedia images and text to generate an embedding space with unbiased images and text. During inference, we will mask the biased words and replace them with objective alternatives. Section 3.2 will use the CLIP [19] model to generate Cross-Modal Embedding Space with different biases. During inference, we will use the de-biased text from Section 3.1 to generate a de-biased image for our news article. Lastly, we will use the pre-trained Vision Transformer [10] used in Section 3.2 to predict the bias of each image in Section 3.3. We will fine-tune the Vision Transformer with more images and its respective labels.

3.1. Text Bias Identification and Neutralization

Our architecture for Text Bias Identification and Neutralization can be seen in Figure 2. First, we use the detection module from [18] to find biased words from news articles. Given an input sentence, we will generate an output sentence that is semantically similar to the input, but the bias will be neutralized. The source sentences will be passed through a BERT model [8] to determine the bias probability of each word of the source sequence using the technique presented in [18]. The source sequence will be from the Politics Dataset's news article.
We will fine-tune another BERT model with paired neutral image/text data (Wikipedia images and text). First, we will use the biased words detected from the detection module and mask them. After that, we will encode the article images into a sequence of tokens. Finally, the BERT Neutral Word Predictor will take in the encoded image sequence and the masked article text to predict the masked biased words and create a new sentence that has reduced bias.

3.2. Semantically Aligned Images

The second primary step of the approach is performing semantic alignment based on the bias score introduced in the description of the overall architecture. We would do this by creating semantic neighborhoods following the process of [25]. A modified triplet loss within a CLIP model is used as the learning process to develop semantic alignment between the images using existing semantics in the text. This loss enforces that the texts are semantically placed close to each other and using them to create a link between their respective images so that visually distinct images with similar meanings will be nearby in the semantic space. The ground-truth semantic similarity would be generated using a pre-trained Doc2Vec model. A new loss specific to this image debiasing task is also developed to account for the bias of images in this shared space. Aside from the modified triplet loss that draws distinct images of the same topic, the images should also be separated by bias. This process is done with another triplet-inspired objective in which we create positive pairs using an image with both the same semantics and the same bias and negative pairs using an image with the same semantics but different bias classification. Semantically aligned images within a bias score range of 10% of one another could be considered for a positive pair. This way, the original modified triplet loss will group semantically-aligned images, and these semantic groupings will be further subdivided based on bias classification (e.g., adjusted triplet loss provides a group of images about the topic "climate change" and the new triplet loss separates those images into left, right, and neutrally biased subgroupings). We can formulate a similar loss with semantically aligned text that differs in bias. With this, retrieval can be performed with debiased text, which will locate the nearest image that

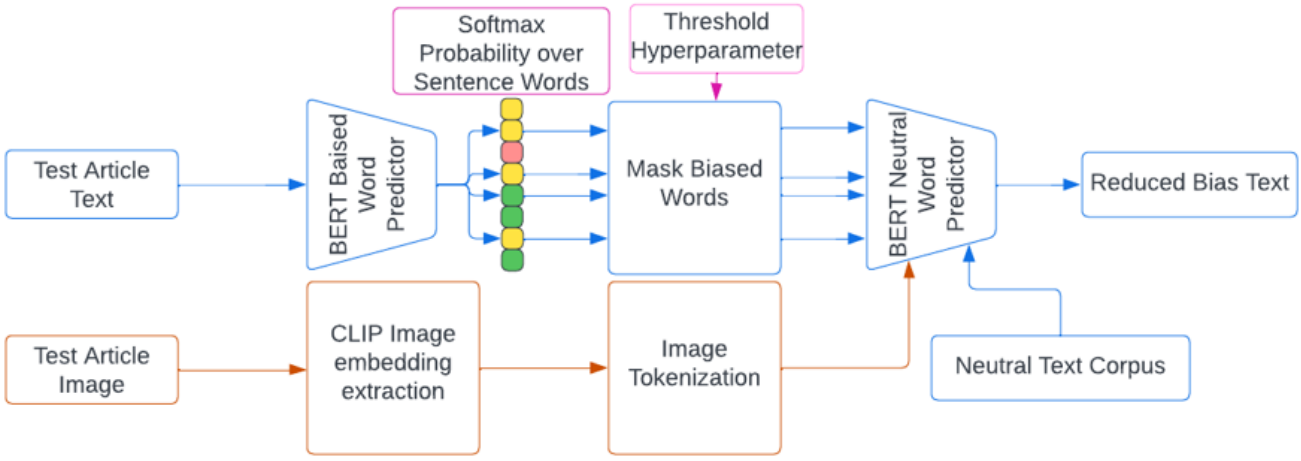

Figure 2. Text Bias Neutralization

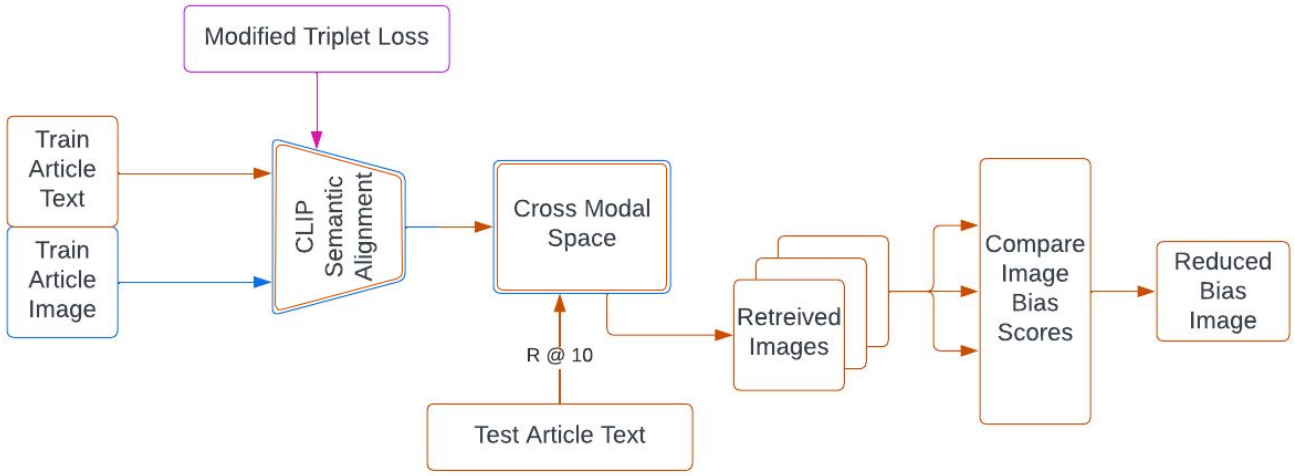

Figure 3. Semantic Alignment Model

is both semantically similar and neutrally biased. Here, the model would take the debiased text as the input and feed it through a text-to-image cross-modal retrieval process to retrieve an image as the output. The nearest semantic image would be the designated "replacement image," allowing the model to perform image debiasing by "replacing" the query image with a less biased alternative with similar semantics. We will use the Image Bias Prediction in Section 3.3 to get the least biased image.

3.3. Image Bias Prediction

The ViT model in Figure 4 is trained in Section 3.2. We will fine-tune this model further with more images and its respective bias score. Finally, we will pass an image through this model at inference, giving us a score from -1 to 1.

4. Experimental Evaluation Protocol and Expected Results

Because the proposed approach is multi-faceted, the evaluation protocol will assess each model aspect individually. First, we want to be able to classify an image using a bias score before being able to find a suitable replacement. A subset of the Politics dataset [24] will be withheld for testing the scoring ability of our model, and scoring accuracy will be measured based on the average distance between the bias prediction generated for each image and their true bias scores, measured in the range of -1 to 1 for extreme left to extreme right. Further qualitative and quantitative as-

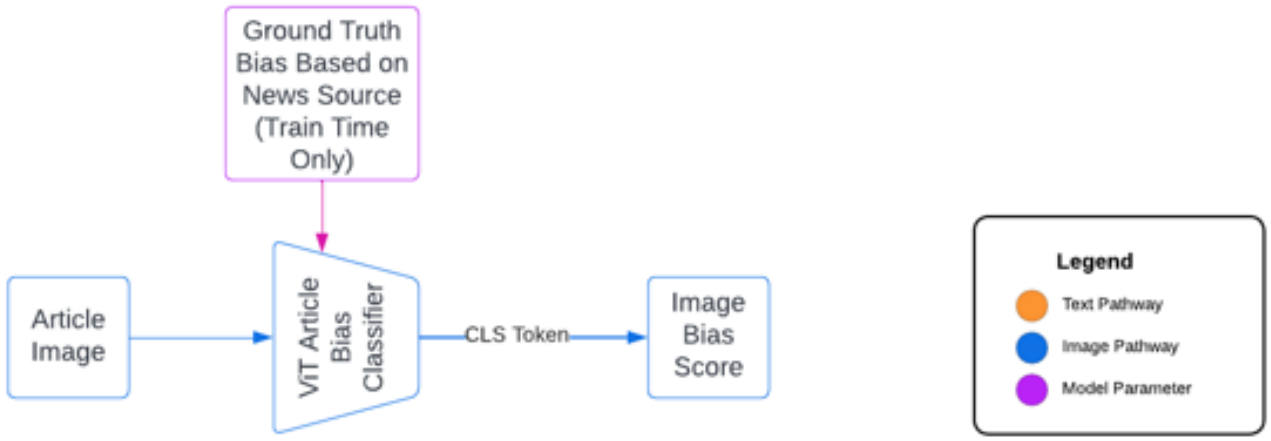

Figure 4. Image Bias Prediction

sessments of the difference distribution will further aid in identifying the strengths and weaknesses of the image bias scorer. An average bias difference metric of 0 means that every predicted bias score was correct. In contrast, an average bias difference metric greater than or equal to 0.25 means the scorer consistently gives images with an incorrect score to an extent where they would enter a different category, such as an "extreme left" image being scored as "left" or a "right" image being scored as "neutral" or "extreme right." Our classifier will be considered successful if 80% or more of the images fall within +/- 0.25 metric difference of their ground truth score. Due to political image debiasing being a somewhat novel task, qualitative studies will also occur. Understanding how biased (or unbiased) input images are relative to the scorer predictions and comparing them to the alternative produced by the system will provide additional intuition on other unmeasured proclivities of the model itself.

Additionally, a pre-trained CLIP model is used and fine-tuned using a modified triplet loss to improve semantic alignment between text and images in a shared embedding space [19, 25]. Performing this semantic alignment using text to guide the images provides the backbone for finding potential replacement images of the same object and possible images of concepts that relate to the image in context, i.e., connect to the same topic as that of the image. Image debiasing is a retrieval task, with heavily biased images being used to find images of similar objects and topics with less bias. To test its retrieval ability, the bias from the lowest-bias image from the input image's neighborhood (including the input image) will be used in calculating an average divergence away from neutral of the potential replacements.

$$\frac{1}{|X|}\sum_{x\in X}|\min_{x'\in N(x)\cup x} b(x')| \quad (1)$$

Where X is the set of test images from the Politics dataset, b is the bias function of an image that returns the ground truth bias for known images in the embedding space and estimated bias for newly scored input images, and N is the set of 10 nearest image neighbors of the input image. A final test metric of less than 0.25 will tell us that, on average, the retrieved images stayed in the neutral range. We also qualitatively evaluate image replacements to see better whether replacement images maintain similar meaning and relevance as the original images.

Because a unimodal text-debiasing model is required for the proposed approach, a bias text detector and neutralizer similar to [18] will automatically identify biased words in context and replace them with appropriate unbiased alternatives or deletions. We use BERT as the text encoder and word bias predictor. A multimodal BERT-based model taking image and text information as input is then used as the decoder to replace biased text. As done in [18], we will evaluate our text debiasing model using accuracy, which measures the percentage of words that were correctly edited in a way that matches the ground-truth debiasing edits of the WNC. The text neutralizer will be considered successful if a 93.5% or greater accuracy is achieved, as was accomplished in [18].

5. Current Progress

5.1. Results

First, using PyTorch from scratch, we wrote code for a Vision Transformer (ViT) [10]. We modified the code from the original paper by reducing the batch size and number of transformer layers because of the sheer number of parameters in those layers. Also, our code has linear value output prediction instead of multi-classification prediction. We replace the loss function from its original paper with Mean Squared Loss, which is convenient for regression problems.

Our model will predict a score from -1 to 1 once it is trained on the Politics Dataset with a labeled score.

Secondly, we used a pre-trained BERT encoder, and fine-tuned it on the Wiki-Neutral dataset with the objective to predict masked tokens, by freezing all but the last layer. We mask 15 percent of tokens in the input set. We can then run this pre-trained BERT model on our article dataset, for now manually masking the high bias words, to re-predict neutral words in place. One qualitative example, (Mitt Romney raving on the senate floor) becomes masked, (Mitt Romney [MASK] on the senate floor), then re-predicts to (Mitt Romney speaking on the senate floor).

Next, we created a spreadsheet with all the websites from the Politics Dataset [24] and their respective political leaning (right or left). Each team member will label these websites with a score of 0 to 1 for right-leaning websites and -1 to 0 for left-leaning websites. We will use several factors in our upcoming meetings to accurately label these websites. This newly labeled data will be used to mark each image-text pair with a political bias score.

Finally, we are writing code for the other parts of the main architecture, including the BERT-based Text Bias Neutralization and the CLIP Semantic Alignment models. This code will soon be finalized, tested, and integrated together. A GitHub repository was created to maintain the final code for this project.

5.2. Challenges and Bottlenecks

There are several challenges and bottlenecks that were encountered throughout our progress. One primary challenge for implementing this algorithm is properly integrating the different components. By the time of the proposal paper, this was one component of our architecture that we have not expanded upon and it was difficult to determine exactly what type and how many inputs would be passed from one part of the model to the next. Another challenge was designing the loss functions for the architecture, since having carefully defined loss functions allows the overall model learn as properly intended and not doing so would hinder our results significantly. Some other challenges include the difficulty of labeling the websites as right or left, and whether to make the algorithm multi-class or not.

In terms of unforseen bottlenecks, we neglected to verify whether or not the Politics Dataset contained a text modality or not. This became a bottleneck because our architecture is reliant on both image and text modalities and thus it would not be suitable to integrate a dataset that fails to contain one of the modalities in some form.

5.3. Ethical Considerations/Limitations

While the this approach is intended to reduce political bias in news that is supposed to cover objective truths, it is important to recognize that there are potential situations in which the technology could itself be biased or misused. Because bias scores must be manually assigned to each source in the dataset, there is potential for human bias to affect the grading of the data. In turn, this could negatively impact the downstream debiaser, for example if scorers tended to provide slightly more left-leaning scores, then a retrieved replacement image may be slightly more left-leaning than normal.

There are additional ethical considerations to examine as well. When would debiasing an article's images or text be considered censorship? In other words, how does one determine reasonable situations in which to use this debiaser or accept its results? This work also only considers bias as a value on a single spectrum. Additionally, this work takes an America-centric view of political media which is likely different from the political climate of other countries. Issues could arise both from this simplistic single-spectrum view as well as from the differing ranges of political opinions across different regions. Depending on these political and societal differences between countries, certain media could seem more biased or neutral than it would somewhere else.

Another ethical question that arises is: if the negative or positive aspects being covered are truthful, is it appropriate to replace an image that reflects that emotional element with one that does not? For example, in an article about human rights abuse, would it be appropriate to replace an image of distressed people with one of people in a similar situation who do not seem distressed? This is more difficult to separate because the truthful information being conveyed could be inherently evocative. Addressing this with a debiasing model could help remove a political slant to an extent, but would not help correct for other contextual information that was left out that creates bias in the piece.

An additional consideration is one related to peoples' perceptions of the model itself. If vulnerabilities or adversarial exploitations are found, we do not want the model to be used as a veneer for neutrality even when the underlying content is not unbiased nor objective. In a similar vein, the model could be used to find more biased image replacements, as there are not explicit restrictions placed on the bias of retrieved images.

References

[1] Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10):1531–1542, Oct. 2015. Publisher: SAGE Publications Inc. 1

[2] Levi Boxell. Slanted Images: Measuring Nonverbal Media Bias During the 2016 Election, Apr. 2021. 2

[3] Wei-Fan Chen, Khalid Al-Khatib, Benno Stein, and Henning Wachsmuth. Detecting Media Bias in News Articles using Gaussian Bias Distributions, Oct. 2020. arXiv:2010.10649 [cs]. 2

[4] Wei-Fan Chen, Khalid Al-Khatib, Henning Wachsmuth, and Benno Stein. Analyzing Political Bias and Unfairness in News Articles at Different Levels of Granularity, Oct. 2020. arXiv:2010.10652 [cs]. 1

[5] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Learning to Model and Ignore Dataset Bias with Mixed Capacity Ensembles, Nov. 2020. arXiv:2011.03856 [cs]. 2

[6] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data. *Journal of Communication*, 64(2):317–332, Apr. 2014. 1

[7] Lincoln Dahlberg. The Habermasian public sphere: A specification of the idealized conditions of democratic communication. *Studies in Social and Political Thought*, Jan. 2004. 1

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. arXiv:1810.04805 [cs]. 3

[9] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-Dimensional Gender Bias Classification, May 2020. arXiv:2005.00614 [cs]. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs]. 3, 5

[11] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A Survey on Bias in Visual Datasets, June 2022. arXiv:2107.07919 [cs]. 2

[12] Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84, Florence, Italy, Aug. 2019. Association for Computational Linguistics. 2

[13] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying Societal Bias Amplification in Image Captioning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13440–13449, New Orleans, LA, USA, June 2022. IEEE. 2

[14] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards Debiasing Sentence Representations, July 2020. arXiv:2007.08100 [cs]. 1

[15] Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W. Black. Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings, July 2019. arXiv:1904.04047 [cs, stat]. 1

[16] Odbal, Guanhong Zhang, and Sophia Ananiadou. Examining and mitigating gender bias in text emotion detection task. *Neurocomputing*, 493:422–434, July 2022. 2

[17] Yilang Peng. Same Candidates, Different Faces: Uncovering Media Bias in Visual Portrayals of Presidential Candidates with Computer Vision. *Journal of Communication*, 68(5):920–941, Oct. 2018. 2

[18] Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically Neutralizing Subjective Bias in Text, Dec. 2019. arXiv:1911.09709 [cs]. 1, 2, 3, 5

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, Feb. 2021. arXiv:2103.00020 [cs]. 3, 5

[20] Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. Linguistic Models for Analyzing and Detecting Biased Language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1659, Sofia, Bulgaria, Aug. 2013. Association for Computational Linguistics. 1

[21] Scott P. Robertson. *Social Media and Civic Engagement: History, Theory, and Practice*. Synthesis Lectures on Human-Centered Informatics. Springer International Publishing, Cham, 2018. 1

[22] Manjira Sinha and Tirthankar Dasgupta. Determining Subjective Bias in Text through Linguistically Informed Transformer based Multi-Task Network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, pages 3418–3422, New York, NY, USA, Oct. 2021. Association for Computing Machinery. 2

[23] Tejas Srinivasan and Yonatan Bisk. Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models, May 2022. arXiv:2104.08666 [cs]. 1

[24] Christopher Thomas and Adriana Kovashka. Predicting the Politics of an Image Using Webly Supervised Data. 1, 2, 3, 4, 6

[25] Christopher Thomas and Adriana Kovashka. Preserving Semantic Neighborhoods for Robust Cross-modal Retrieval, July 2020. arXiv:2007.08617 [cs]. 1, 2, 3, 5

[26] Jinglin Wang, Fang Ma, Yazhou Zhang, and Dawei Song. A Multibias-mitigated and Sentiment Knowledge Enriched Transformer for Debiasing in Multimodal Conversational Emotion Recognition, July 2022. arXiv:2207.08104 [cs]. 2

[27] Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. Demoting Racial Bias in Hate Speech Detection, May 2020. arXiv:2005.12246 [cs]. 2