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 with each tweet pertaining to different political and nonpolitical issues. In this study, researchers observed that people who share the same ideologies when it comes to political issues exchange information with each other much more than those who share different ideologies [1]. Through empirical observation, we see that the Internet and social media serve as the primary news distributor 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 opinions that it is built upon [5]. We postulate that the biggest contributor for 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 [6, 18]. 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 [15]. Text debiasing was done by removing subjective bias by using the method provided by [17]: using Wikipedia's Neutral Point of View (NPOV) policy. For debiasing political images, the work performed on it is very limited and not well explored. To solve the problem of debiasing political images, we will perform cross-modal alignment of textual images and news. Then we will select the image with the least bias using the method in [21]. In summary, our work attempts to learn an algorithm that would minimize and neutralize political bias from 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 [12]. Therefore, to address the problem of biases in the real world models several works have been proposed. Some previous work was centered around debiasing the sentence level representations, which removes the religious, gender, racial and cultural biases [12, 13]. [12] 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, few works have been done on a multi-modal setting [20]. In terms of these works, [20] does research to demonstrate gender bias in VL-BERT, which works in a multi-modal setting. After thorough search of the existing literature, we found that hardly any prior work has been done to debias both images and texts in a multimodal setting. Performing work on debiasing both images and texts does have 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, [3] uses over 6900 news articles with labels derived from a website to develop a neural model for bias assessment and [15] 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. [15] 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, [21] 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. [22] models a real world scenario where image-text pairs convey complementary information with little overlap. The approach used in [22] helps preserve the semantic relationships between paired images and paired text. From [22] given a text or news headline, we can generate a series of semantically aligned images. The objective of our project is to remove subjectivity of the news and also replace the news images with semantically aligned images which are politically less biased.

2.2. Political Bias Identification (Images)

In terms of work related to political image debiasing, some existing papers focused on finding and predicting political bias within images. [21] utilizes a two-step process where a model first learns relevant visual concepts of an image to enable bias prediction and then a visual classifier is trained upon that model. Otherwise, it appears that image debiasing in the context of politics does not seem to be well explored. There are several approaches to other more general image bias related tasks that could be extended to political bias identification. For identifying bias in images, some work focuses on reducing bias to attribute/label data that is provided with or extracted from the data [10]. Other common approaches focus on learning from lower-dimensional representations or determining bias by cross-analyzing multiple datasets [10]. A more recent and uncommon approach utilizes an ensemble classification system through training a low-capacity network and a high-capacity network in order to reduce bias [4, 10]. Overall, while there are several methods that allow for identifying bias in images, most of these methods are, to our knowledge, not applied or extended to the politics domain and 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 are used in previous works to identify and reduce bias in text. [15] and [19] use pre-trained BERT transformers to identify and correct biased words. [24] and [14] use adversarial training to create a classifier to detect and correct hate speech. [11] uses an attention based mechanism on the article headline network to detect bias in the article body, with the goal of mirroring the order in which humans read articles. [2] uses a Gaussian mixture model to observe probability distributions of the frequency, positions and order of information to detect article level bias. In terms of text dataset annotation, [8] proposes a framework to decompose gender biases across multiple dimensions, including the gender of the speaker, person being spoken to, and person being spoken about. However, these previous approaches don't leverage or interact with available visual data, or only focus on correcting specific types of bias, ie. gender, political, 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 ones which are semantically similar, 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, there are also several different methods used for evaluations. Frequently, individual datasets are scraped and developed by individual 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 a specific kind of bias [2, 11], 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 corpuses to identify and potentially correct biased language [15, 23]. In most cases concerning bias detection with these datasets, a popular evaluation metric is simply to check the accuracy of the model in classifying as biased in binary or multi-class settings. When it comes to evaluating text debiasing methods, basic natural language metrics such as BLEU can be used [15]. Human evaluations are also important 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 [15].

2.5. Politics Dataset

We will be using the dataset that is collected in [21]. 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 [21] 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 as well as metadata and additional contextual sentences which were all scraped from Wikipedia edits that ensured texts were as unbiased as possible [15]. This dataset will be used to neutralize the biases in the text in political news articles.

3. Proposed Approach

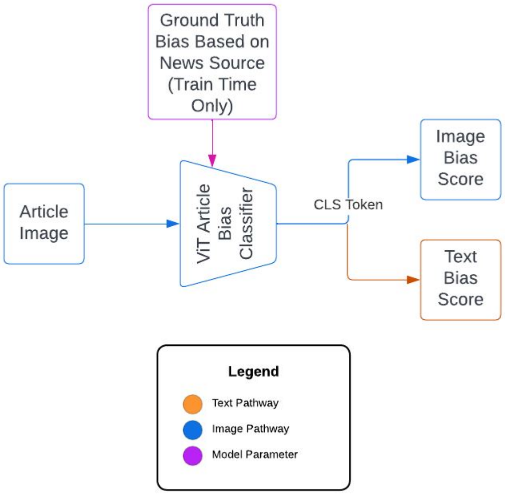

Figure 1. Image Bias Prediction

3.1. Image Bias Prediction

This approach consists of three main aspects: identifying the bias score of an input image, semantic alignment between article images, and performing text debiasing. First, a model must be trained that is able to identify image bias score at inference time because our article source is not assumed to be available during testing and debiasing. A Visual Transformer model (ViT) will be trained to accept an image as input and return a score representing the image's political bias on a scale from -1 to 1 which represents "extreme" left and right respectively, with a neutral image being represented as 0 [9]. During training, the article news source, and thus bias level is available. We assume that all images in an article inherit the bias level of its parent news source, and thus the target bias level for an image during training comes from the news source's score. The loss function used to train the scorer is a simple mean squared error which helps to penalize produced scores that lie outside the desired score range as well as help draw predictions that are opposite of their true values more heavily towards their actual side of bias. Training an image bias scorer enables further preprocessing of image bias scores when performing low-bias image retrieval in the second stage as well.

3.2. Semantically Aligned Images

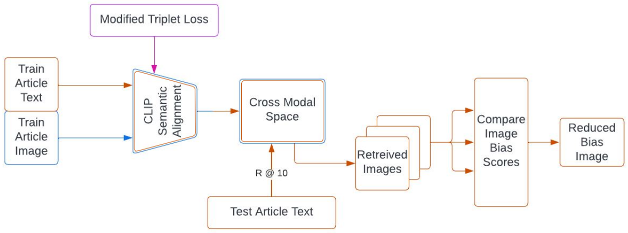

Figure 2. Semantic Alignment Model

The second main step of the approach is performing semantic alignment based on the bias score. This would be done through the creation of semantic neighborhoods following the process of [22]. A modified triplet loss within a CLIP model is used as the learning process to create semantic alignment between the images using existing semantics in the text. This loss enforces that the texts are semantically placed close to each other as well as using them to create a link between their respective images which will be smoothened by that same loss. The ground-truth semantic similarity would be generated using a pretrained Doc2Vec model. The model for this step would take a query image and its associated text as the input and feed it through a text-to-image cross-modal retrieval process to retrieve an image as the output. Within that retrieval process, the 10 most semantically related images would be retrieved using the text of the biased query image. The semantic image in the top 10 that has the lowest bias score would be the designated "replacement image", allowing the model to perform image debiasing by "replacing" the query image with a less biased alternative.

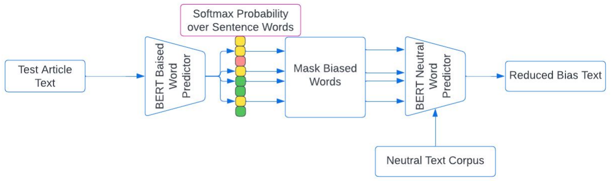
Softmax Probability over Sentence Words
Test Article Text
BERT Baised Word Predictor
Mask Biased Words
BERT Neutral Word Predictor
Reduced Bias Text
Neutral Text Corpus

Figure 3. Text Bias Neutralization

3.3. Text Bias Identification and Neutralization

First, we use the detection module from [15] to neutralize the text. Given a input sentence, we will generate an output sentence that has similar meaning to the input but the bias will be removed. We will have source sequences with subjectively biased text and target sequences with the neutralized version of source sequence from the Wiki Neutrality Corpus (WNC). Given the source and target, we will use the BERT [7] model to figure out bias in the words. We train this model by using the difference between the source and target text. We use the loss function by calculating the average negative log likelihood of the labels:

$$L = \frac{1}{n}\sum_{i=1}^{n}[p_i^* logp_i + (1 - p_i^*)log(1 - p_i)] \tag{1}$$

where p_i^* is 1 if the source text was deleted or modified during the neutralizing process. It is 0 if the associated word was unchanged during editing. p_i is the probability that each module input word is subjectively biased. This probability is calculated in [15] section 3.

We will train another BERT model with Neutral Text Corpus with several pre-training tasks. We will use the biased words detected from the detection module and mask them. After that, we will use the BERT Neutral Word Predictor to predict the masked biased words to create a new text that has Reduced Bias Text.

4. Experimental Evaluation Protocol and Expected Results

Because the proposed approach is multi-faceted, the evaluation protocol will assess each of the model aspects 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 [21] will be withheld for testing the scoring ability of our model, and scoring accuracy will be measured based on the provided classes of extreme left, left, right, and extreme right bias, representing -1, -0.5, 0.5, and 1 respectively. The metric used here will be the average distance between the bias prediction generated for each image and their true bias scores. Further qualitative and quantitative assessments 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 every predicted bias score was exactly correct, while an average bias difference metric greater than or equal to 0.25 means the scorer consistently misscores images 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 take place. 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 further fine-tuned using a modified triplet loss to improve semantic alignment between text and images in a shared embedding space [16, 22]. Performing this semantic alignment using text to guide the images provides the backbone for finding not only potential replacement images of the same object, but also potential images of concepts that relate to the image in context, i.e. relate to the same topic as that of the image. Image debiasing acts as 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.

$$\frac{1}{|X|}\sum_{x\in X}|\min_{x'\in N(x)\cup x} b(x')| \tag{2}$$

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 better see 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 [15] will be used for automatically identifying biased words in context and replacing them with appropriate unbiased alternatives or deletions. While we also use BERT as the text encoder and word bias predictor, we intend to use another transformer-based model for the decoder as well to better attend across the biased words for better replacements in this seq2seq task. As done in [15], 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 [15].

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] 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

[3] 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

[4] 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

[5] 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

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

[7] 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

[8] 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

[9] 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

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

[11] 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

[12] 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

[13] 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

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

[15] 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, 4

[16] 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]. 4

[17] 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

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

[19] 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

[20] 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

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

[22] Christopher Thomas and Adriana Kovashka. Preserving Semantic Neighborhoods for Robust Cross-modal Retrieval, July 2020. arXiv:2007.08617 [cs]. 1, 2, 3, 4

[23] 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

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