Unsupervised Discovery of Implicit Gender Bias: A New Analysis of Reddit and Fitocracy

Arsh Banerjee
arshb@princeton.edu

Pierce Maloney
pmaloney@princeton.edu

Christian Ronda
cronda@princeton.edu

Abstract

Despite their potential harms, social biases are ubiquitous in our social media platforms, internet forums, and even our algorithms. Differing human judgments have made detecting bias difficult for these platforms and algorithms are unlikely to flag content unless it is explicitly extreme. We apply a causal framework to identify implicit gender bias at the comment level to two corpora: Reddit and Fitocracy. The framework established by Field et al. eliminates confounds and flags text likely to contain bias. Accordingly, we determine the method's effectiveness and then utilize sentiment analysis to explain any difference in results. Our analysis shows how sentiment strongly correlates to the ability to detect implicit gender bias. Our work offers insight into how implicit gender bias detection can differ across different platforms, and unsupervised approaches can become more effective.

1 Introduction

From court decisions to mortgage applications, bias has embedded itself into every aspect of our lives. Both implicit and explicit, these biases are engrained in our actions and speech and often lead to the perpetuation of systemic inequalities and discrimination. Accordingly, social media platforms have seen growing pressure to regulate hateful speech and due to the scale of the internet, they rely on automated systems to detect bias within posts and comments. These automated systems, however, are often trained on datasets that contain biases themselves or are subject to human judgment through manual annotations. As a result, these automated systems often perpetuate the very biases they are attempting to resolve. Identifying biases in text is inherently difficult, as human annotations are unreliable due to the subjective nature of implicit biases and the need for context. An unsolicited compliment may be indictive of gender bias, while the same comment on a body positivity forum could be perfectly acceptable. The problem then becomes an issue with defining bias as opposed to detecting it. This project attempts to replicate and evaluate a system to detect implicit gender biases within comments to a post. We utilize a causal framework established by Field et al. which allows for the unsupervised discovery of implicit gender bias Field and Tsvetkov (2020). To identify these implicit biases without explicitly defining the term, the following counterfactual is posed:

Would the addressee have received different text if their gender were different?

Field et al. utilized this framework for two Facebook corpora, with one related to comments on posts related to male and female public figures and the other related to comments on posts related to male and female politicians. Their work showed that not only could they detect implicit gender bias successfully but that they could do so without any human judgments. We apply Field et al.'s approach to two new corpora with posts and responses sourced from Fitocracy (a fitness forum) and Reddit. Furthermore, we then perform a textual analysis of the corpus to evaluate the differences among the corpus in terms of other factors like sentiment. Both corpora differ substantially from the Facebook corpora in the types and manner of discussion. By applying the framework to new corpora and then conducting sentiment analysis, we aim to evaluate how the methodology and its effectiveness may differ depending on the nature of the corpora.

2 Related Work

In recent years, substantial effort has been invested in detecting biases in different datasets. This is mostly motivated by the intention of identifying and rectifying broad biases in datasets such that deep learning models do not perpetuate the same biases from the corpus. Accordingly, prior work

often analyzes gender bias at the corpus level in order to control for bias degenerative models, for instance, Dinan et al. classify gender bias in eight large corpora Dinan et al. (2020). While this research does highlight broad insights regarding bias detection, it is only applicable at the corpus level. As such, the current research is not applicable for use to control biases at the comment level on online platforms.

Prior work also heavily relies on human annotation in order to define gender bias for detection models. The earlier work of Dinan et al. crowdsource evaluations of gender bias. Other works also rely only on human judgment for creating datasets to build detection models (Fast et al., 2016). By relying on human judgment for annotations, different datasets will ultimately lead to differing intrinsic definitions of gender bias among models. This will ultimately lead to a natural conflict regarding how bias should be defined, illustrating the need for an unsupervised approach.

A crucial aspect of this project is utilizing sentiment analysis to explain differences in the effectiveness of the discovery framework. Sentiment analysis can be used to determine the polarity of a text, on a scale ranging from positive to negative (Mejova, 2009). The sentiment of a text is inherently tied to levels of bias, as bias is the opposite of objectivity. Since sentiment is a gauge of the objectivity of a piece of text, it is natural to utilize sentiment analysis to evaluate the bias detection framework. To perform the sentiment analysis, we utilize the FLAIR deep-learning text classifier, which outperforms many of its counterparts for this task (Yimam et al., 2020).

3 Methodology

Field et al. attempt to create a model that can identify subtle biases in comments without relying on subjective human judgments or being influenced by confounding factors. This is done by controlling for observed confounding variables through propensity matching, controlling for latent confounding variables using adversarial training, and addressing overt signals by substituting gendered terms with neutral language.

In our work, beyond replicating the unsupervised bias detection methods of Field et al. and applying it to new data, we additionally perform sentiment analysis on the data to explore our hypothesized correlation between sentiment and implicit bias. We wished to investigate whether more negative sentiment correlated to higher rates of implicit bias.

Let us define the following variables we will use in methodology:

- `OW`: "Original Writer", the person who wrote the original text, e.g., the addressee
- `O_TXT`: The content of the original text
- `W_GEN`: The gender (`M` or `F`) of `OW`
- `W_TRAITS`: Traits of `OW` other than gender
- `COM_TXT`: Replies to `O_TXT`

3.1 Controlling Observed Confounding Variables through Propensity Matching

This technique is used in the paper to control for potential biases introduced by observed confounding variables. This can be broken into three steps:

Identifying confounding variables: First, the authors identify observed confounding variables, such as the original writer's age group, country, self-reported occupation, education level, and whether the writer has children. These variables can introduce biases and potentially confound the relationship between `W_GEN` and the dependent variable (language features).

Calculating propensity scores: For each writer, the authors compute a propensity score, which is the probability of being in the female group (`W_GEN=F`) given the observed confounding variables. They use logistic regression to model the probability of `W_GEN=F` as a function of the observed confounding variables.

Matching: Once the propensity scores are calculated, the authors match each writer in the female group with a writer in the male group who has a similar propensity score. This step aims to create a balanced dataset where writers from both groups have similar observed confounding variables.

These steps are to ensure that the differences in language features they identify between male and female groups are less likely to be attributed to these observed confounders.

3.2 Controlling Latent Confounding Variables through Adversarial Training

While propensity matching can control observed confounding variables, it cannot control latent confounding variables such as the writer's traits (`W_TRAITS`) which are not possible to match. This

adversarial training technique is used to control latent confounding variables, encouraging the model to ignore W_TRAITS. The goal is to obtain a model that can predict the gender of the addressee (W_GEN) without being influenced by the writer's traits (W_TRAITS).

Finding Representations: First, Field et al. find the confounding representations: They cannot explicitly enumerate W_TRAITS, but they know these traits are associated with the identity of the Original Writer (OW). To infer W_TRAITS from comments (COM_TXT) addressed to OW, they use log-odds scores to calculate associations between OW and words in COM_TXT. For each input COM_TXT, they obtain a vector whose elements represent the comment's association with each OW individual.

Adversarial Training: Now, the authors aim to create a model that can predict W_GEN but cannot predict the latent confounds represented by the vector obtained in the previous step. They follow an alternate Generative Adversarial Network (GAN)-like procedure:

First, the input comment ($x \in$ COM_TXT) is encoded using an encoder neural network $h(x; \theta_h)$ to obtain a hidden representation (h_x).

Then, the hidden representation (h_x) is passed through two feedforward networks:

- $c(h(x); \theta_c)$ to predict the gender label $y \in M, F$.
- An adversary network $adv(h(x); \theta_a)$ to predict the vector representation of the latent confounds.

The encoder is trained so that h_x does not contain any information predictive of the confound vector but does contain information predictive of the target attribute (gender). The primary training objective for the encoder is to minimize the loss that considers both the gender prediction and the adversary network's prediction of the latent confounds, represented by:

$$\min_{c,h} \frac{1}{N} \sum_{i=1}^{N} \text{CE}(c(h_{x_i}), y_i) + \text{KL}(\text{adv}(h_{x_i}), U_K)$$

where U is a normal distribution, CE is cross-entropy loss, and KL is KL-divergence.

By leveraging adversarial training, Field et al. encourage the model to focus on features that are indicative of the gender group (W_GEN) and not on features specific to individual members of the group (e.g., some group members are from Canada). This way, the model can predict W_GEN without being influenced by the latent confounding variables (W_TRAITS).

3.3 Exploration of Sentiment and its Implications

We wished to perform textual analysis on the corpora to better understand why the classifier indicating implicit bias would perform better or worse on some texts. To do this, we used the pretrained FLAIR framework to evaluate each of the corpora. We used the standard English FLAIR sentiment classifier to perform sentiment analysis on the responses in each dataset, and then from the results, we could draw conclusions regarding the association between sentiment and implicit bias detection.

4 Implementation

4.1 Dataset

In our analysis, we used the RtGender dataset Voigt et al. (2018) which provides five diverse corpora derived from different social media platforms: Facebook, Fitocracy, and Reddit.

4.1.1 Facebook Public Figures Dataset

The first Facebook dataset pertains to public figures from the realms of journalism, fiction writing, television, film, and athletics. It contains posts and associated top-level comments for 105 such public figures, drawn from various sets of Wikipedia categories including American television news anchors, American film actresses/actors, American male/female tennis players, and 21st-century American novelists, among others.

4.1.2 Facebook Politicians Dataset

The second Facebook dataset is centered around politicians. It includes all posts and associated top-level comments from the 412 current members of the United States Senate and House who have public Facebook pages meeting certain criteria.

Both Facebook datasets include only top-level comments, ensuring that each comment is a direct response to the original poster, and anonymize all user information for privacy.

4.1.3 Fitocracy Dataset

Fitocracy, a social media fitness website, provides a dataset of status updates and their corresponding top-level comments. These comments typically

Dataset	Source Individuals		Response Count
Facebook (Public Figures)	M: 41	W: 64	10,667,500
Facebook (Politicans)	M: 306	W: 96	13,866,507
Reddit	M: 19,010	W: 11,116	1,453,512
Fitocracy	M: 52,432	W: 47,498	318,535

Table 1: Statistics of RtGender corpora

pertain to fitness-related activities and progress updates. The first comment after a post is used as it is necessarily a direct response to the original post, ensuring the relevance of our analysis.

4.1.4 Reddit Dataset

The Reddit dataset is a collection of post-response pairs from various subreddits for which the gender of the source poster is known. This dataset covers a wide variety of subreddits, allowing for a diverse analysis of gender bias across a multitude of discussion topics.

4.2 Implicit Bias Detection

For our modeling step, we replicated Field et al's implicit bias detection model. The modeling in this stage includes preprocessing the datasets to replace overly-gendered words with more gender-neutral words, controlling observed confounding variables through propensity matching, and also using adversarial training to control latent confounding variables. These steps improve the model's robustness to the confounding variables in the dataset's to give a more unbiased exploration of gender bias. Field et al ran this model on Facebook Public Figures Dataset and Facebook Politicians dataset, which we accompanied with our own evaluation on these datasets along with the addition of the Fitocracy and and Reddit Datasets.

4.3 Sentiment Analysis

To aid the bias detection model, we used FLAIR, which calculated a sentiment score in the range $[-1, 1]$ for each in our response post in our dataset. Sentiment scores measure the emotion of text: where positive emotions have scores closer to 1, and negative emotions are closer to -1 with neutral scores being around 0. As using FLAIR's pretrained sentiment classifier is very computationally expensive, we randomly sample a proportion of each dataset, stratifying on gender of the author of the response post, and ran the FLAIR model on that subset of response posts.

5 Results

We evaluate the ability of the classifier to detect implicit bias through its accuracy rate of predicting the gender that some text was addressed to. This reflects the causal framework Field & Tsvetkov as if the classifier predicts with high confidence that some text is intended for a woman, it is likely the text contains bias. Similarly, if some text is predicted with high accuracy to be intended for a man, there must be some implicit gender bias leading the classifier to make the prediction. Accordingly, the ability to predict gender is a metric by which to evaluate the effectiveness of the model, the authors of the original paper utilize the same metric. Table 1 gives the accuracy values from Field and Tsvetkov's regarding the Public Figures and Politicians dataset as well as our findings related to the Reddit and Fitocracy corpora.

Pub. Figs.	Politicians	Reddit	Fitocracy
65.1	77.1	77.0	55.9

Table 2: Accuracy rate of predicting the gender of the addressee

The datasets are balanced in terms of addressee gender during preprocessing. Thus any improvement beyond $acc = 0.5$ is indicative of predictive performance. We find that the framework was equally effective in detecting implicit gender bias in the Reddit corpus as the Politicians corpus. Interestingly, however, the framework failed to detect bias in the Fitocracy corpus. Through sentiment analysis, we further explore the differences in effectiveness. The mean sentiments by gender of all the corpora are displayed in Figure 1.

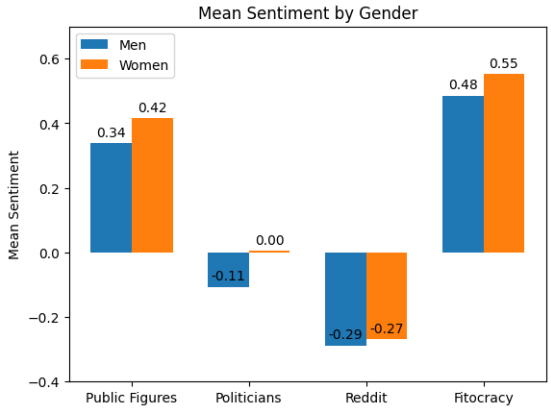

Figure 1: Mean Sentiments by Gender

The mean sentiments of the corpora match expectations regarding known information of the platforms the data is representative of. For instance, users generally speak positively about public figures such as movie and music stars. Similarly, a platform related to body positivity and fitness regimes is likely to contain positive, uplifting language which is reflected in the mean sentiment. Conversely, users may express disdain about politicians leading to its negative sentiment. Pairing this sentiment analysis with the effectiveness in implicit bias detection, a clear correlation emerges. Figure 2 shows the mean sentiment plotted against the Bias Classifier accuracy for each corpus.

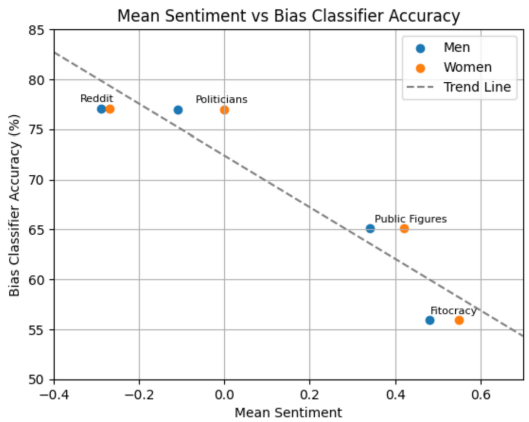

Figure 2: Mean Sentiment Vs. Accuracy For Each Corpus

Figure 2 illustrates a correlation between the average sentiment of a corpus and the ability to detect implicit bias within it. As the mean sentiment trends positive for a corpus, the accuracy of the bias classifier decreases. Our findings quantitatively establish a relationship between sentiment and the level of implicit bias within text. In a forum where the mean sentiment is more negative, the implicit bias is easier to detect. This relationship that we have found lends credence to our hypothesis that negative sentiment leads to higher degrees of implicit bias. This relationship suggests that on a text forum where people are generally more positive, they speak to men and women in a more similar way than people do on a more negative forum.

6 Conclusion

6.1 Effectiveness and Insights

By applying Field and Tsvetkov's framework for detecting implicit gender bias to new corpora, we reaffirm its effectiveness while also highlighting potential areas of weakness. Through sentiment analysis, this work gives insights into how these systems may be evaluated in practice. We provide evidence that users producing text with negative average sentiment are likely to also exhibit higher levels of detectable implicit gender bias. Social media platforms often cite their large scale as barriers to effective moderation. The correlation identified between sentiment and bias can help guide moderation towards forums/communities where gender bias is more likely to be present.

6.2 Limitations and Future Work

While these four corpora (Public Figures, Politicians, Reddit, and Fitocracy) may include thousands of users and millions of text responses, the conclusions drawn are limited to these specific platforms. In order to draw more concrete conclusions, future work would be needed to expand the source dataset with more text from different platforms (Twitter, Quora) for better generalization. On a similar note regarding dataset limitations, future work where gender is not restricted to the binary male or female may also produce interesting findings about how implicit gender bias presents itself. Furthermore, the RtGender dataset provides information regarding the gender of the responder. Analysis of how levels of bias differ depending on the gender of the speaker could also reveal important attributes about gender bias as a concept, but also how content moderation can be specifically targeted to minimize its presence on online platforms.

Acknowledgements

We would like to thank Professor Chen for her support and our advisor Austin Wang for his feedback and advice throughout the project.

References

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification.

Ethan Fast, Tina Vachovsky, and Michael Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media*. volume 10, pages 112–120.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. *arXiv preprint arXiv:2004.08361* .

Yelena Mejova. 2009. Sentiment analysis: An overview. *University of Iowa, Computer Science Department* .

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. RtGender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1445.

Seid Muhie Yimam, Hizkiel Mitiku Alemayehu, Abinew Ayele, and Chris Biemann. 2020. Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics*. pages 1048–1060.