DSS/Worldie Project Summary

Background: In Autumn 2024, DSS partnered with Worldie – Social Media for Good, a nonprofit that helps victims of sexual misconduct and abuse in the visual and performing arts. The goal of the project was to identify bias and signs of media manipulation against victims and women by performing sentiment analysis upon a previously compiled dataset of 511 case articles, which we divided into 4 categories based on whether they were Control or Non-Control and AMI or Non-AMI. To clarify these terms, in Control case articles, bias and media manipulation is expected (e.g., sexual misconduct lawsuits), whereas in Non-Control ones, it is not expected (e.g., standard divorces). Here, AMI case articles are those published by American Media, Inc., now rebranded as A360 Media. The purpose of this work is to demonstrate and quantify media bias against victims and women, to encourage publications to exercise more oversight over their reporting of sexual misconduct and abuse, and to inspire public critique and legal regulation against publications that perpetuate these harmful reporting practices.

Research Questions: Does the case article dataset showcase bias against women and victims of sexual misconduct and abuse? How do these biases vary between the 4 case article categories?

Methods: To scrape the article text, the team used BeautifulSoup. Coreference resolution was applied to identify entities in text with a rule-based spaCy approach and pre-trained Spark NLP models. Separately, single-case analysis was performed on 29 case articles, which constituted reading and analyzing them to identify signs of linguistic and tonal bias. Finally, the Non-Control portion of the dataset was analyzed GPT-4o by (1) evaluating GPT-4o's ability to assess bias fairly in articles, (2) observing what the model prioritizes when making bias evaluations, (3) creating well-defined model features or bias categories based on GPT-4o's priorities, (4) using OpenAI's API to train a GPT-4o model that quantifies article bias accordingly, and (5) feeding the scraped article text into this trained model to assign bias scores to each Non-Control case article, as Control articles lack victims.

Results: As a result of step (3) in the GPT-4o article analysis, 6 model features were derived: Coverage Bias, Evidence Disparity, Language Favorability, Tone, Contextual Framing, and Legal and Procedural Accuracy. In step (5), the team assigned Overall Bias scores in [-10, 10] to each Non-Control case article where greater scores indicated favorability towards the victim (see Figure 1). Demonstrating the robustness of our model features, Overall Bias scores were strongly positively correlated with Coverage Bias, Evidence Disparity, and Language Favorability. While most articles had moderate bias (46% of scores in [-4, -2] or [2, 4]), there were far more articles heavily biased toward victims (16.6% of scores in [6, 10]) than perpetrators (0.5% of scores in [-10, -6]). As for gendered bias, when Perpetrator = Female and Victim = Male, coverage bias is -0.049, but when roles are reversed, coverage bias is 0.297 (see Figure 2). Due to accuracy problems in coreference resolution and the sparsity of the single-case analysis approach, no results were gleaned from these methods.

Discussion: The models' detection of coverage bias towards victims and gender bias towards women does not necessarily imply that there is no underlying bias towards men and perpetrators. Given the subject matter of the Non-Control cases and their primarily male perpetrators, one should expect a baseline skew against men and perpetrators.The existence of true bias against men and perpetrators is contingent upon where detected bias is relative to the negative baseline. Further research is needed to contextualize these results, clarify the true direction of the bias, and identify the interaction between bias and the Control/Non-Control and AMI/Non-AMI case classifications.

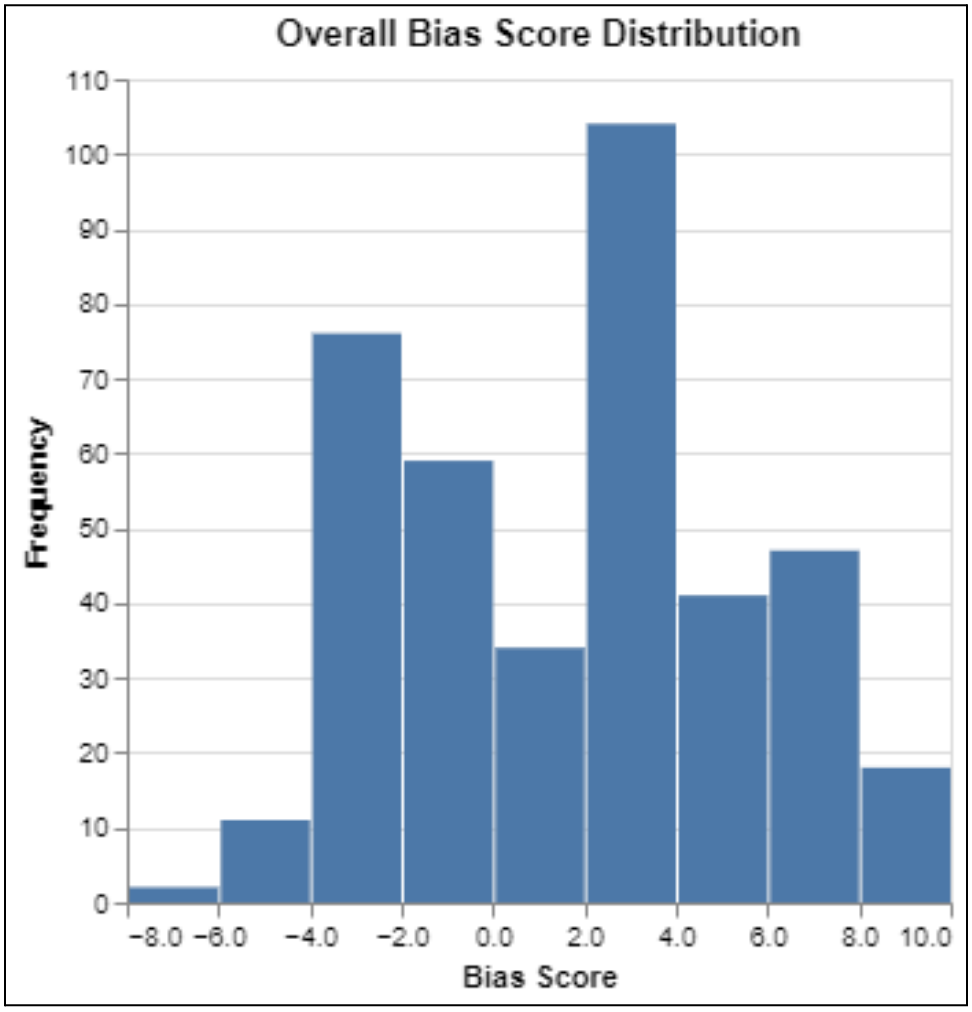

Figure 1. Histogram of Overall Bias scores for Non-Control case articles with left-skewed distribution and greater frequencies of positive bias scores, suggesting bias towards victims.

	Perpetrator Gender	Victim Gender	Overall Bias [-10, 10]	Coverage Bias [-1, 1]
0	Female	Female	-0.500	-0.080
1	Female	Male	0.173	-0.049
2	Female	Unknown	-1.292	-0.400
3	Male	Female	2.280	0.297
4	Male	Male	1.364	0.209
5	Male	Unknown	3.221	0.430
6	Unknown	Female	1.667	0.233
7	Unknown	Male	-0.294	-0.153
8	Unknown	Unknown	2.000	0.275

Figure 2. Table of Overall Bias and Coverage Bias scores of Non-Control case articles grouped by the gender of the victim and perpetrator.