Data Science Society/Worldie Project Final Presentation

Chris Frazer, Cynthia Di, Cynthia Zeng, Francy Hsu,
Jeremy Zhao, Victoria Karai, and Violet Reed

Background

- Partnership with Worldie – Social Media for Good, a nonprofit that helps victims of sexual misconduct and abuse in the visual and performing arts
- Project goal: To identify bias and signs of media manipulation against victims and women by performing sentiment analysis upon 511 case articles
- Specifically, we would like to demonstrate differences between Control and Non-Control case articles as well as AMI and Non-AMI case articles.
  - In Control case articles, bias/manipulation is <u>not</u> expected (e.g., standard divorces).
  - In Non-Control case articles, bias/manipulation is expected (e.g., sexual misconduct).
  - AMI case articles are those published by American Media, Inc., now rebranded as A360 Media.

Methods

- Scraped data from 511 case articles across 10+ domains using BeautifulSoup for Control/Non-Control and AMI/Non-AMI classification
- Identified entities in text using coreference resolution with a rule-based spaCy approach and pre-trained Spark NLP models
- Employed single-case analysis, which constituted reading and analyzing individual case articles to identify signs of linguistic and tonal bias
    - This analysis considers the article authors, their larger body of worker, and omission of case information to push a narrative (e.g., implying a perpetrator has just one victim when there are truly multiple).

Literature on Linguistic Bias

Recasens, Danescu-Niculescu-Mizil, and Jurafsky (2013) examined the edits made to Wikipedia articles to lessen bias and found:

1. Framing Bias
    a. Intensifiers: “fantastic” or “accurate”
2. Epistemological Bias
    a. Factive verbs: words that presuppose truth, “realized”
    b. Entailments: “murdered” instead of “killed”
    c. Assertives: verbs that help to assert a proposition, “pointed out”
    d. Hedges: “may”, “could”

GPT-4o Article Analysis

Overview

Objective: Analyze the bias in Non-Control case articles using a sophisticated Large Language Model

Steps:

1. Evaluate GPT-4’s ability to fairly assess bias in articles
2. Take note of what the model prioritizes when making bias evaluations
3. Create well-defined categories based on these priorities
    a. These categories become the “features” of the model
4. Use OpenAI’s API to train a GPT-4 model that quantifies article bias
5. Feed model scraped article text to generate a dataset of article bias scores

Model Features

1. Coverage Bias (-1 to 1): Proportion of the article’s text spent covering allegations against perpetrator & background of victim vs. defense & background of perpetrator
    a. -1 = Heavy coverage of perpetrator’s defense, 0 = equal coverage, 1 = heavy coverage of allegations
2. Evidence Disparity (-1 to 1): The amount of evidence presented indicting the perpetrator vs. defending the perpetrator or questioning the victim’s credibility
    a. -1 = primarily evidence defending perpetrator, 0 = equal evidence, 1 = evidence indicating perpetrator
3. Language Favorability (-1 to 1): The connotation of the language used when describing the actions of the perpetrator and any sympathetic or critical language used toward either side
    a. -1 = Language favorable toward perpetrator/unfavorable toward victim, 0 = neutral, 1 = language favorable toward victim, unfavorable toward perpetrator

Model Features cont.

4. Tone (0 to 1): Objectivity of the article
    a. 0 = objective tone, 1 = significantly charged tone
5. Contextual Framing (0 to 1): Degree to which the article provides context of broader social movements, systemic issues, or patterns of behavior
    a. 0 = minimal added context, 1 = context established and heavily emphasized
6. Legal and Procedural Accuracy (0 to 1): How well the article explains legal procedures or outcomes
    a. 0 = little explanation of procedures/outcomes, 1 = detailed explanation of procedures/outcomes

“Training” the Model

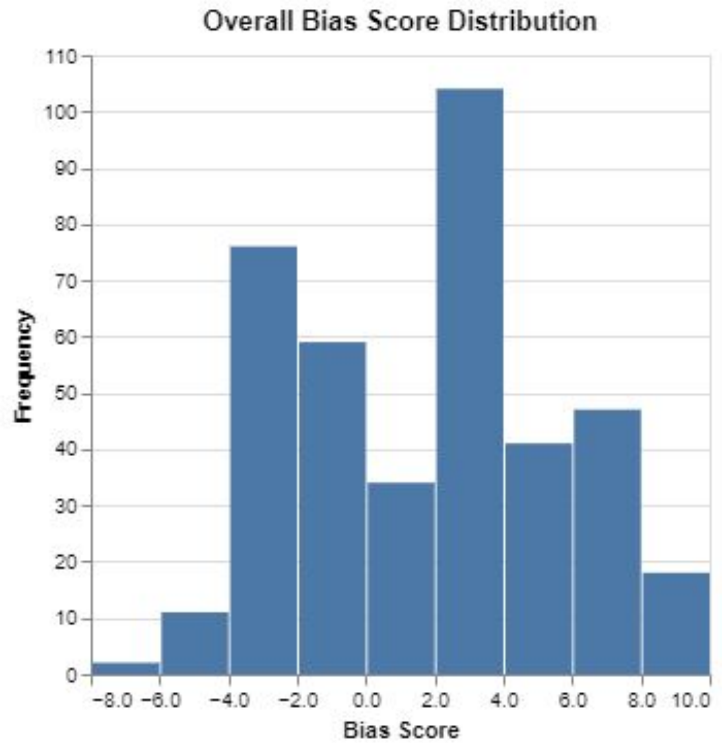
1. Provide a description of model’s intended function
   a. “You are an article bias analyst. I will provide you with news articles and the name of a perpetrator and victim, and you are to quantify potential bias in the article toward the given perpetrator or victim.”
2. Provide an ordered list of instructions that the model should follow
   a. Evaluate articles on the following criteria
   b. Generate an overall bias score from -10 to 10
   c. Generate category scores for each criterion in step one
3. Provide the general structure of an input to the model
   a. Text from an article preceded by the line “Perpetrator: *perpetrator name*, Victim: *victim name*”

“Training” the Model cont.

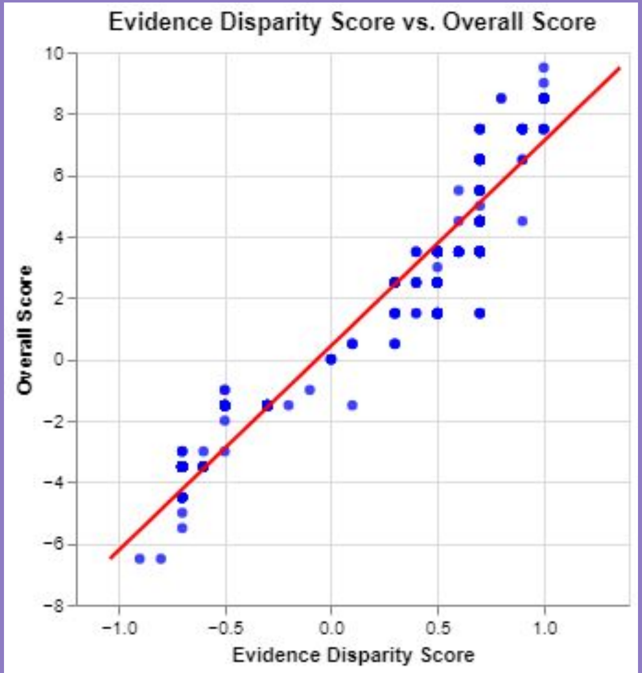
4. Provide the specific structure of what the model should output
    a. {“Overall:” *overall bias score*,
       “Coverage Bias”: *coverage bias score* …}
5. Test the model to see if it works
6. If something needs to be fine-tuned, provide a list of pointed instructions at the end of the model instructions file

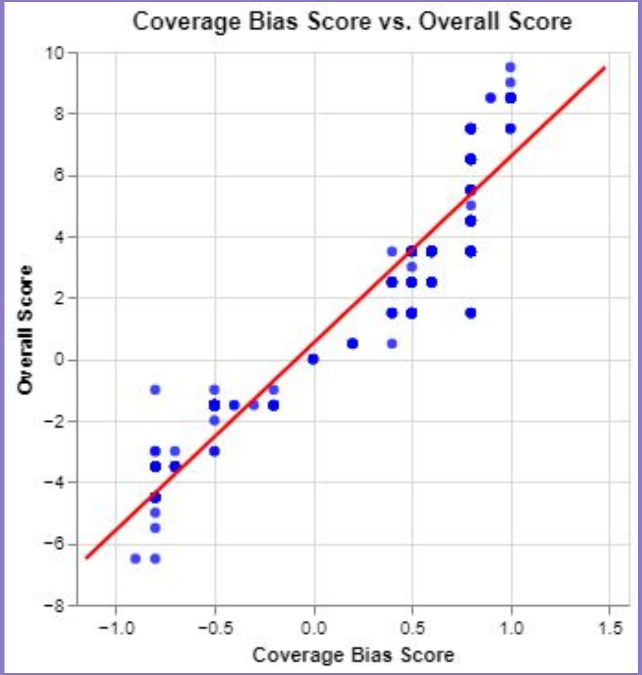
Results

Article Biases



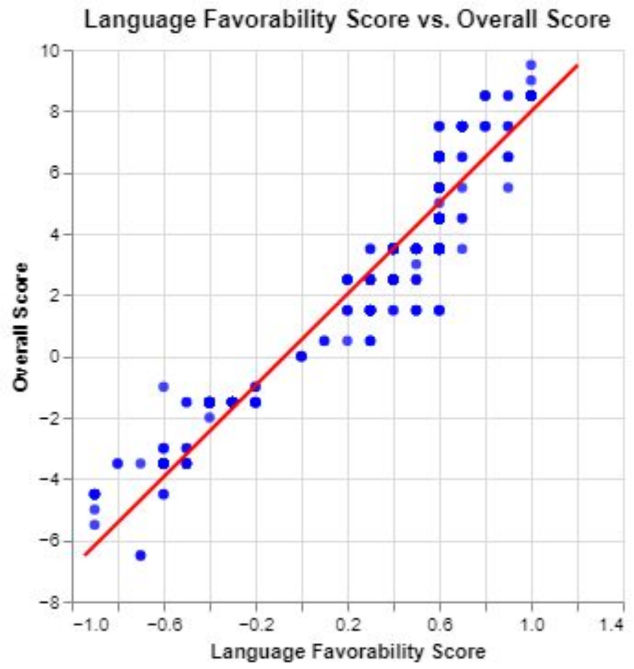
- 46% of articles had moderate bias
  - Scores between 2-4 or -2- -4
- Only 24% of articles had little to no bias (scores between -2 and 2)
- There were significantly more articles heavily biased toward victims than perpetrators
  - 0.5% of scores < -6
  - 16.6% of scores > 6

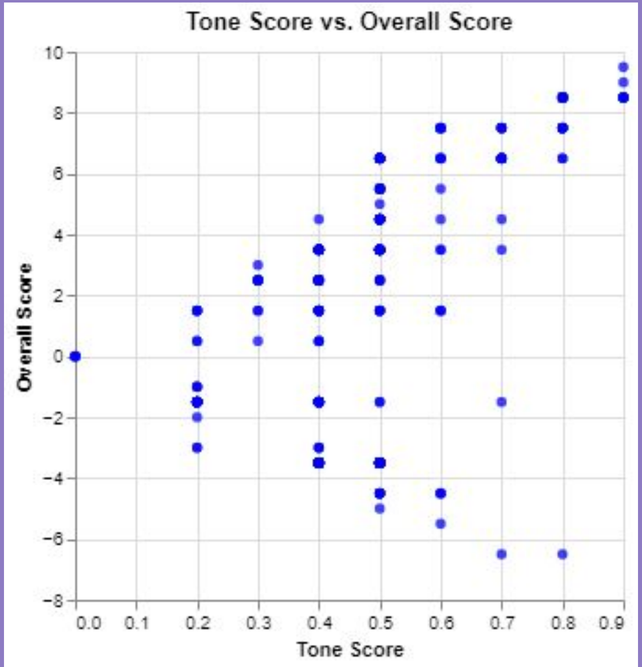




- Strong positive correlation
- **Intuition:** More evidence defending one side tends to indicate bias toward that side

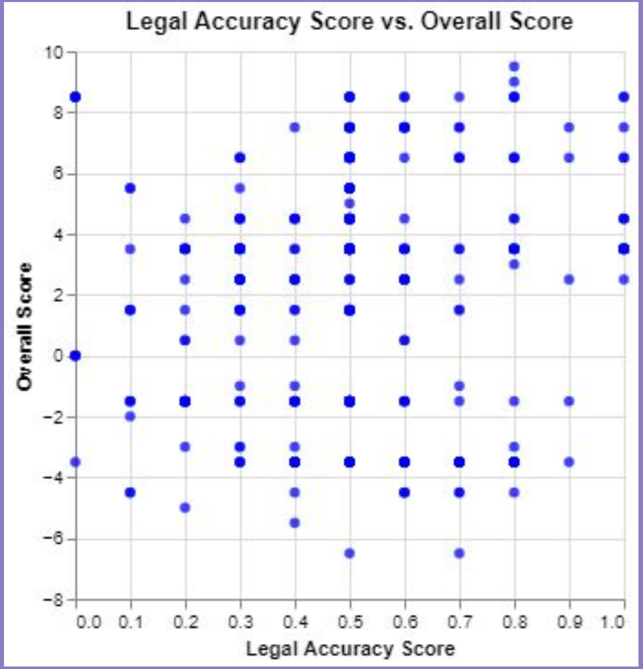
- Strong positive correlation
- **Intuition:** More coverage of one side tends to indicate bias toward that side

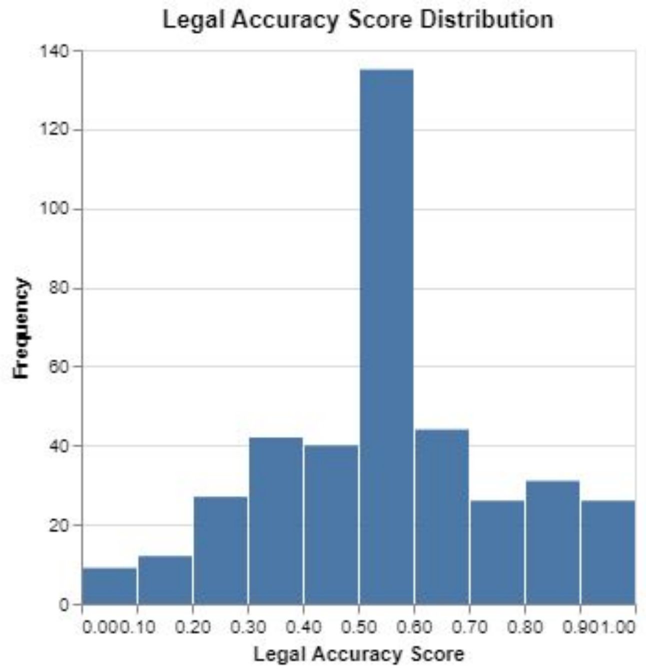




- Strong positive correlation
- **Intuition:** Disparity of favorable language describing one side indicates bias

- Parabolic relationship
- **Intuition:** Articles with more objective (lower) tone scores tend to be less biased (overall score ~ 0)

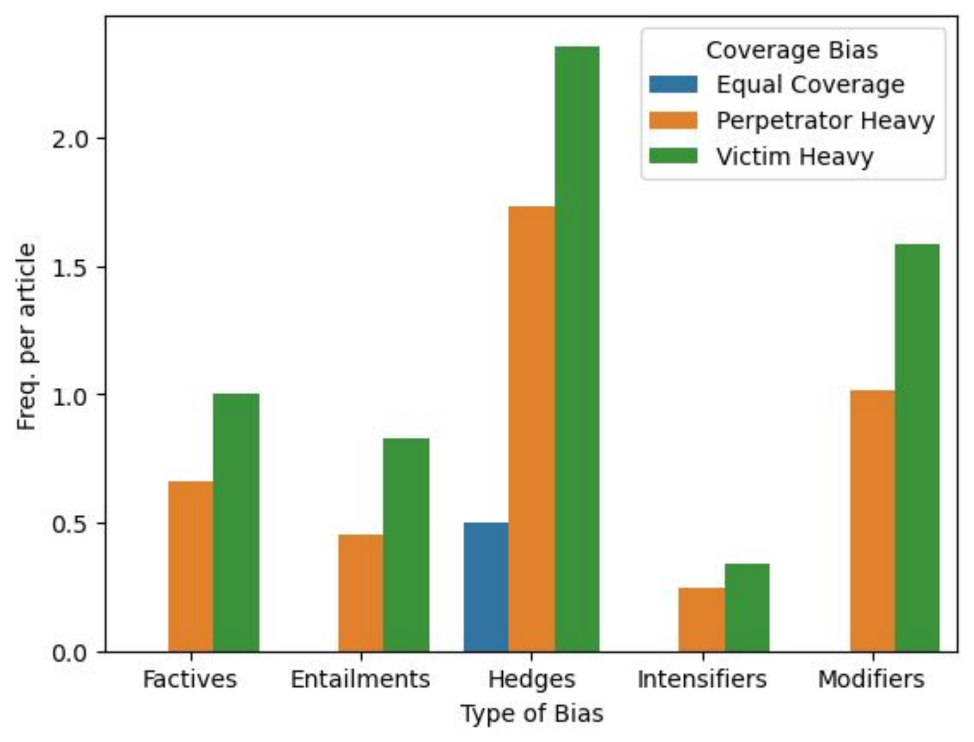




- No correlation
- Appears that quality of legal explanations has no discernable bearing on bias

- Relatively normal distribution
- Overwhelming majority of scores are between 0.5 and 0.6

Coverage vs. Biased Words

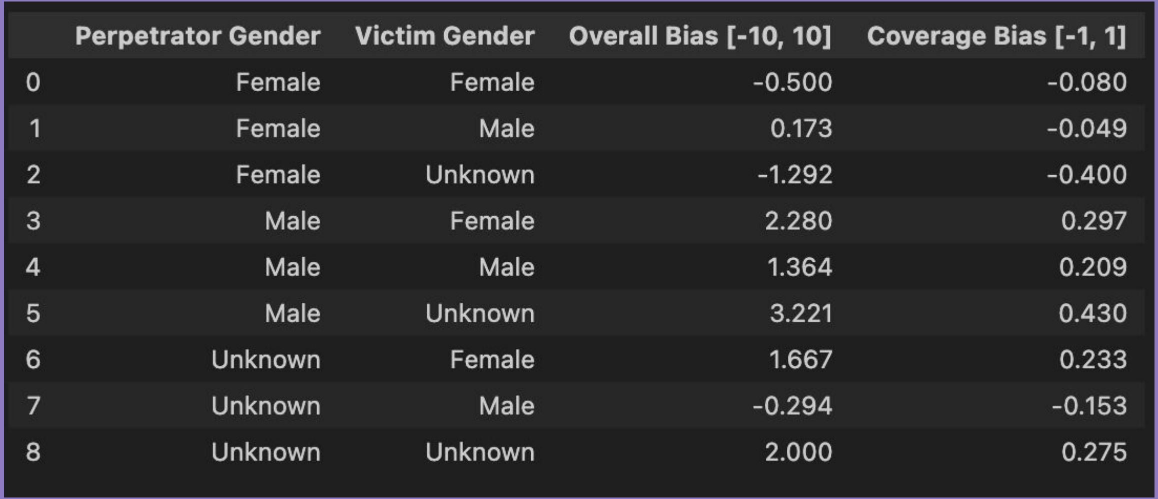


Coverage Bias: Does the article favor one side in how much time it spends defending or giving background for them?

Gender and Bias

This is the output of the average bias, grouped by gender pairs.

Before any significance testing, there seems to be little discrepancies in the Overall Biases.

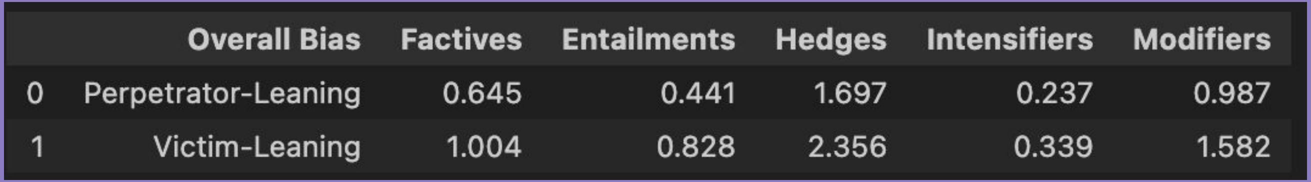


	Perpetrator Gender	Victim Gender	Overall Bias [-10, 10]	Coverage Bias [-1, 1]
0	Female	Female	-0.500	-0.080
1	Female	Male	0.173	-0.049
2	Female	Unknown	-1.292	-0.400
3	Male	Female	2.280	0.297
4	Male	Male	1.364	0.209
5	Male	Unknown	3.221	0.430
6	Unknown	Female	1.667	0.233
7	Unknown	Male	-0.294	-0.153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

Coverage Bias ranges from [-1,1] with -1 being heavy coverage of perpetrator defense/background and 1 being heavy coverage of victim defense/background. There seems to be evidence suggesting that articles, on average, cover more of the female's side of the case. When Perpetrator = Female and Victim = Male, coverage bias is -0.049. When roles are reversed, coverage bias is 0.297.

Biased Words and Article Lean

The lean variable shown here is a binary variable derived from Overall Bias. There are more biased words, on average, across all categories in Victim-Leaning articles.



	Overall Bias	Factives	Entailments	Hedges	Intensifiers	Modifiers
0	Perpetrator-Leaning	0.645	0.441	1.697	0.237	0.987
1	Victim-Leaning	1.004	0.828	2.356	0.339	1.582

This suggests that more often than not, the biased words are used to:

1. Attack or doubt the perpetrator’s story
2. Defend the victim’s story

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, we 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
  - Identify the interaction between bias and the Control/Non-Control and AMI/Non-AMI case classifications.

Thank you!