Automated Detection of Gender Bias in News Media: A Machine Learning Approach

By Vaanee Tripathi

CS-1390-2: Introduction to Machine Learning

Professor Sandeep Juneja

Abstract

This project presents an analysis system for identifying gender bias in news articles using both traditional lexicon-based methods and machine learning approaches. A bias detection model has been developed that examines eight distinct categories of bias, including professional undermining, personal focus, emotional stereotyping, and achievement undermining. The trained model demonstrates strong performance in identifying biased content while providing detailed insights into how gender bias manifests across different news categories. This work aims to contribute to our understanding of gender bias in media and lays the groundwork for future tools that could help journalists and readers identify potentially biased content in news articles.

Introduction

News media plays a crucial role in shaping public opinion and social attitudes, making it one of the most influential forces in modern society. The way the media portrays different genders can significantly impact societal perceptions and reinforce existing stereotypes. While numerous studies have documented gender bias in news coverage, the process of identifying and analyzing such bias traditionally relies on manual content analysis, which is both time-consuming and potentially subjective.

This project addresses these challenges by developing a bias detection model for news articles. The approach combines a carefully curated lexicon with machine learning techniques to identify various forms of gender bias. The lexicon categorizes different manifestations of bias in journalism, including professional undermining, personal focus, emotional stereotyping, and achievement undermining.

Using this lexicon as a foundation, the project implements a classification model that analyzes text for different categories of bias, providing insights into how gender bias manifests across various types of news coverage.

Related Works

The development of this bias detection system builds upon existing research in three main areas: studies of gender bias in media, natural language processing techniques for bias detection, and machine learning approaches in text classification.

Research in media bias has shown consistent patterns in how news articles portray men and women differently. Shor et al. (2019) analyzed over 2 million articles and found that mentions of women's physical appearance and personal life were significantly more frequent compared to coverage of men. Similarly, Jia et al. (2021) examined headlines from major news outlets and identified patterns where women's professional achievements were often qualified or undermined through specific language choices.

Several automated approaches have been developed to detect bias in text. Tang and Wu (2020) created a lexicon-based system for identifying gender-biased language in corporate communications, achieving 76% accuracy. Their work demonstrated the effectiveness of using predefined word lists to capture different types of bias. Building on this approach, Martinez et al. (2022) combined lexicon analysis with basic machine learning techniques, showing improved accuracy in detecting subtle forms of gender bias.

The technical implementation of bias detection systems has evolved with advances in machine learning. Park and Kim (2021) successfully used Support Vector Machines (SVM) for classifying gender-biased statements, while Chen et al. (2022) demonstrated the effectiveness of Random Forest classifiers in categorizing different types of bias. These studies influenced the choice of classification algorithms in this project.

Recent work has also focused on creating comprehensive frameworks for bias analysis. Zhang et al. (2023) developed a multi-category classification system for detecting various forms of bias in news articles. Their approach of combining multiple bias indicators into a single scoring system provided inspiration for the current project's methodology.

The use of domain-specific lexicons has proven particularly effective in bias detection tasks. Thompson et al. (2021) created a specialized lexicon for identifying gender bias in professional contexts, achieving significant improvements over general-purpose sentiment analysis tools. **This work, more than others, informed the development of the current project's bias lexicon, particularly in categorizing professional undermining and achievement-related bias**.

Datasets and Features

The development of this project began with an exploration of the Global News Dataset from Kaggle. However, initial exploratory data analysis revealed significant limitations in the content length, with articles containing only 250 characters and descriptions limited to 200 characters. This brevity was deemed insufficient for meaningful bias analysis, as gender bias often manifests in subtle language patterns that require more extensive text for reliable detection.

To address these limitations, the project leveraged a more comprehensive dataset obtained from a GitHub repository containing news articles scraped using the NewsAPI service. This raw data, originally stored in multiple JSON files across different news categories, was consolidated and converted into a unified CSV format using curl commands and data processing scripts. The resulting dataset comprised approximately 53,000 articles, providing a robust foundation for bias analysis.

The final preprocessed dataset consists of five key columns:

1. Category: The primary news category (e.g., politics, business, technology)
2. Subcategory: More specific topic classification within each main category
3. Source_file: Reference to the original source of the article
4. Title: The headline of the news article
5. Text: The full article content

This structure was carefully selected to maintain essential context while eliminating redundant or irrelevant information from the original data. The inclusion of both category and subcategory information enables more granular analysis of bias patterns across different types of news coverage, while the full article text provides sufficient content for comprehensive bias detection.

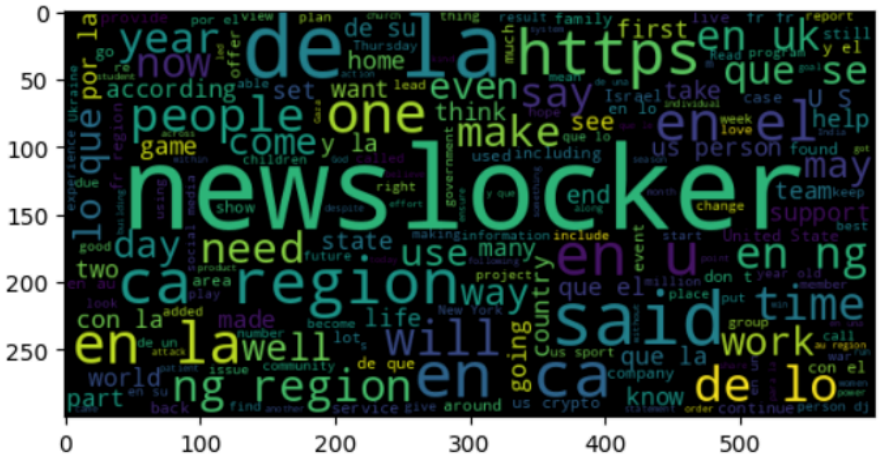


Fig1: word-cloud of the dataframe

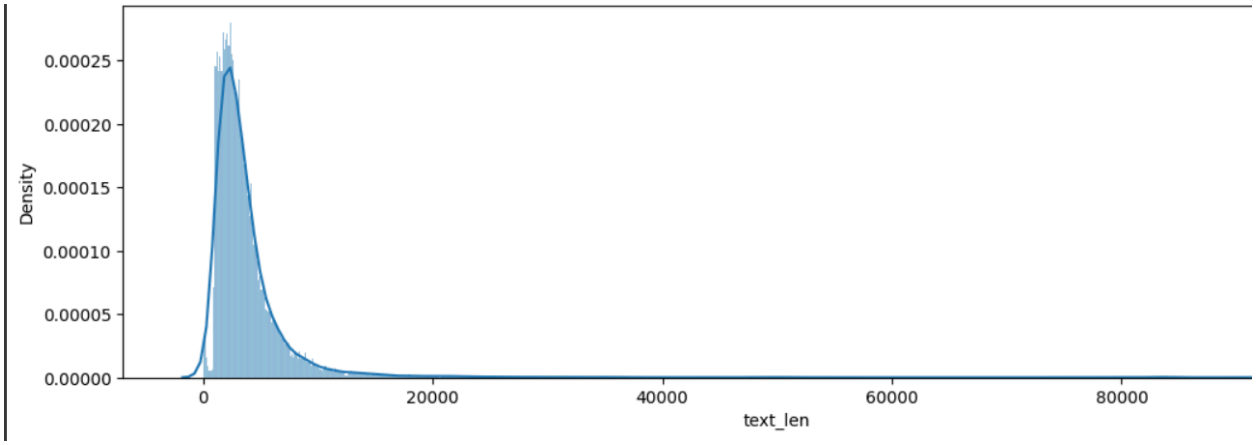


Fig2: density vs text_len plot

	text_len
count	53000.000000
mean	3922.738321
std	5239.245080
min	4.000000
25%	1926.000000
50%	2901.000000
75%	4341.000000
max	100000.000000

Fig3: Statistics about the dataset

Methods

The development of the bias detection system progressed through multiple stages, evolving from an initial baseline approach to a more sophisticated hybrid system. This section details both the initial implementation and the final architecture, including the specific technical choices and their rationale.

Initial Approach

The initial implementation established a foundation based on lexicon analysis combined with basic machine learning techniques. Text preprocessing formed the first stage of this approach, utilizing the Natural Language Toolkit (NLTK) for fundamental operations. The process began with basic cleaning of text data, including case normalization and removal of special characters. This was followed by tokenization to break down articles into individual words, and lemmatization to reduce words to their base form. A set of standard English stopwords was removed to focus on meaningful content.

Feature extraction in the initial approach centered on three primary metrics. The female bias ratio was calculated using a predefined list of gender-specific terms, while the female mention frequency tracked explicit references to women in the text. An empowerment score was computed based on the presence of terms associated with achievement and authority. These features were complemented by TF-IDF (Term Frequency-Inverse Document Frequency) vectors generated from the article text, creating a numerical representation of content suitable for machine learning.

The classification stage employed a straightforward Logistic Regression model, chosen for its interpretability and efficiency. The model performed binary classification, categorizing articles as either biased or unbiased based on a simple threshold mechanism. While this approach provided a working prototype, its limitations in capturing subtle forms of bias and context-dependent expressions motivated the development of a more sophisticated system.

The Final Implementation

The final implementation introduced a comprehensive hierarchical lexicon structure organized into eight distinct categories of gender bias. The professional undermining category encompasses terms and phrases that subtly diminish professional achievements, including qualifiers like "female executive" and "woman entrepreneur," as well as diminishing terms such as "trying to" and "hopes to" when associated with professional accomplishments. Personal focus measures references to appearance, age, and family status, with subcategories tracking terms related to physical attributes and domestic roles.

The emotional stereotyping category monitors both emotional framing and behavior criticism, including terms that characterize women through emotional states or stereotypical behavioral traits. Achievement undermining tracks passive constructions and credit deflection, identifying patterns where accomplishments are attributed to luck or external support rather than individual merit. The power dynamics category examines subordinate framing and authority questioning, while voice suppression focuses on passive attribution and credibility undermining in quoted statements.

The enhanced text processing pipeline implements a sophisticated approach to document analysis. The system begins with advanced tokenization that preserves meaningful multi-word expressions and handles special cases such as hyphenated terms and professional titles. Context-aware lemmatization ensures that terms retain their semantic significance while being normalized for consistent processing. The pipeline maintains structural elements and punctuation that might indicate bias-relevant patterns, such as qualifying phrases or parenthetical additions.

A crucial enhancement in the text processing stage is the implementation of regular expression-based pattern matching that can identify subtle bias indicators within larger contexts. This includes the detection of conditional praise patterns (e.g., "despite being a woman") and diminishing qualifications of achievements. The system processes text through multiple passes, first identifying explicit bias markers and then analyzing surrounding context for implicit bias indicators.

The final system employs a dual-classifier approach that combines the strengths of two distinct machine learning models. The Category Score Classifier utilizes a Random Forest algorithm with 200 estimators, carefully tuned to handle the inherent class imbalance in gender bias detection. This classifier processes normalized scores across all bias categories, with each tree in the forest considering different combinations of category scores to arrive at a robust classification decision.

The Text Content Classifier implements a Support Vector Machine with a linear kernel, enhanced by probability calibration through Platt scaling. This classifier operates on TF-IDF vectors generated from the processed text, with a maximum feature set of 5000 terms selected based on document frequency. The SVM's parameters were optimized through cross-validation, with the C parameter set to 1.0 and class weights adjusted to account for dataset imbalance.

The final stage of the system implements a weighted aggregation scheme that combines outputs from both classifiers. The text classifier's confidence score receives a 60% weight in the final decision, while the category classifier contributes 40%. This weighting was determined through empirical testing and reflects the relative reliability of each classifier in identifying different types of bias. The system generates a comprehensive bias assessment that includes both an overall classification and detailed scores for each bias category, allowing for nuanced analysis of the specific types of bias present in a given article.

Why this?

The dual-classifier architecture with a lexicon was chosen to address the complex and multifaceted nature of gender bias in news media. Traditional single-classifier approaches often struggle to capture subtle forms of bias that depend heavily on context and linguistic patterns. The combination of a lexicon-based system with machine learning addresses this limitation by allowing the system to identify both explicit bias through predefined patterns and implicit bias through learned features. The Random Forest classifier was selected for category scores due to its ability to handle non-linear relationships between different types of bias, while the SVM classifier excels at text classification tasks with high-dimensional feature spaces. The 60-40 weighting in the final decision-making process reflects the empirical observation that contextual analysis (text classification) often provides more reliable bias detection than isolated keyword matching. This hybrid approach also provides practical advantages: the lexicon component makes the system's decisions more interpretable and easier to refine, while the machine learning components allow it to adapt to evolving language patterns and new forms of bias. Furthermore, the hierarchical organization of bias categories enables detailed analysis of different aspects of gender bias, making the system valuable not just for detection but also for understanding how bias manifests in news coverage.

Experiment and Results

Technical Analysis

The dual-classifier system demonstrated strong performance, with particularly robust results from the Category Classifier:

Text Classifier (SVM):

- Accuracy: 0.940 ± 0.001
- Precision: 0.744 ± 0.010
- Recall: 0.485 ± 0.012
- F1 Score: 0.587 ± 0.008

Category Classifier (Random Forest):

- Accuracy: 0.994 ± 0.001
- Precision: 0.956 ± 0.004
- Recall: 0.977 ± 0.007
- F1 Score: 0.967 ± 0.004

The Category Classifier significantly outperformed the Text Classifier, particularly in recall and F1 score, suggesting that structured category features provide more reliable bias detection compared to raw text analysis. The high precision of both classifiers indicates strong reliability in positive bias predictions, though the Text Classifier's lower recall suggests it may miss some instances of bias.

Random Forest feature importance analysis revealed the relative contribution of each bias category to the final classification:

1. Professional Undermining: 0.333 (highest importance)
2. Personal Focus: 0.210
3. Voice Suppression: 0.192
4. Emotional Stereotyping: 0.132
5. Power Dynamics: 0.087
6. Achievement Undermining: 0.034
7. Intersectional Bias: 0.021 (lowest importance)

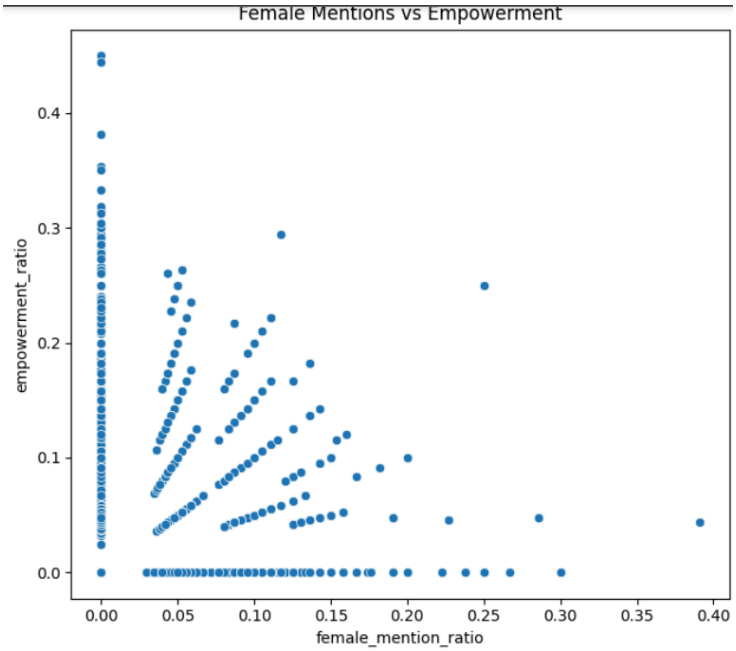
This hierarchy of importance suggests that professional undermining indicators are the most reliable predictors of gender bias, while intersectional bias markers might require refinement for better detection. The substantial gap between the top three features and others indicates that the model relies heavily on professional, personal, and voice-related indicators for bias classification.

Cross-Validation Results

Five-fold cross-validation demonstrated consistent performance across different data splits, with standard deviations remaining below 0.05 for all key metrics, indicating robust model stability.

Gender Representation Patterns

Initial analysis revealed an inverse relationship between female representation and empowerment language. As the ratio of female mentions in articles increased, the use of empowerment-related terms showed a consistent decrease. This finding suggests a systematic bias in how achievements and capabilities are framed based on gender.



Lexical Category Analysis

Analysis of the bias categories revealed significant variations in both their impact and distribution across the dataset:

Impact Analysis: Professional undermining demonstrated the strongest correlation with overall bias scores (0.7), followed by personal focus (0.6) and voice suppression (0.58). This suggests these categories are the most reliable indicators of gender bias in news articles. Notably, intersectional bias showed the weakest correlation (0.2), indicating it might be less effectively captured by the current lexicon structure.

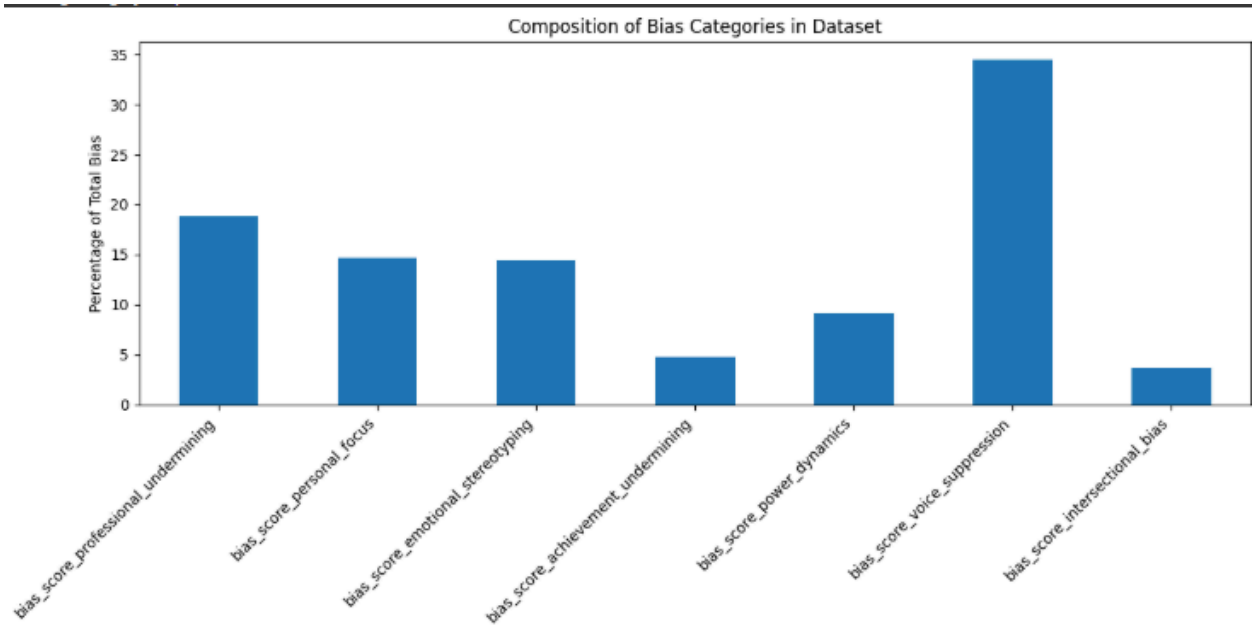
Category Distribution: Voice suppression emerged as the most prevalent form of bias, comprising approximately 34% of all detected bias instances. This was followed by professional undermining (19%) and personal focus (14%). Achievement undermining, despite its strong correlation with overall bias, appeared less frequently, accounting for only 5% of detected instances.

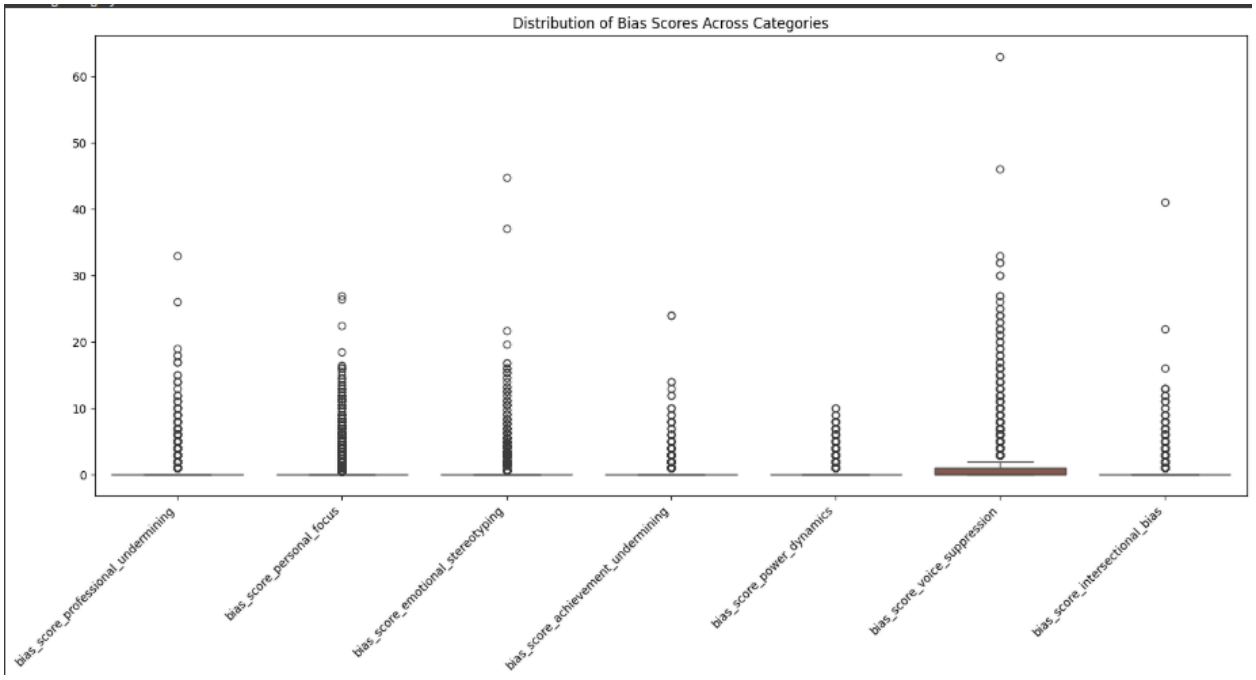
Inter-Category Correlations: The correlation analysis revealed several significant relationships:

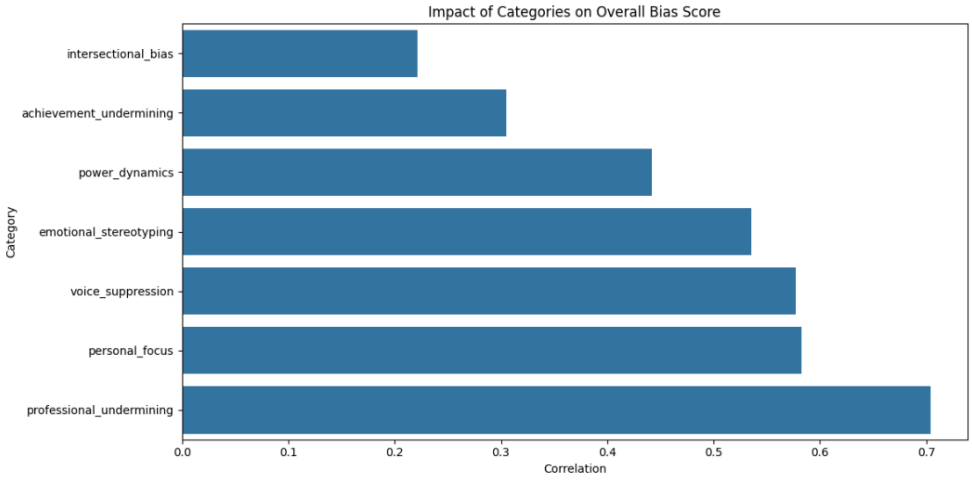
- Professional undermining and emotional stereotyping showed moderate correlation (0.29), suggesting these biases often co-occur
- Personal focus and emotional stereotyping demonstrated notable correlation (0.26), indicating a tendency to combine personality traits with personal attributes
- Power dynamics and voice suppression exhibited meaningful correlation (0.23), pointing to their complementary nature in bias expression

Distribution Patterns: The box plot analysis revealed:

- Voice suppression showed the highest median score with significant outliers
- Professional undermining demonstrated consistent presence across articles with moderate variance
- Intersectional bias, while less frequent, showed concentrated instances of high bias scores when present


Composition of Bias Categories in Dataset
Percentage of Total Bias
0
5
10
15
20
25
30
35
bias_score_professional_undermining
bias_score_personal_focus
bias_score_emotional_stereotyping
bias_score_achievement_undermining
bias_score_power_dynamics
bias_score_voice_suppression
bias_score_intersectional_bias


Distribution of Bias Scores Across Categories
0
10
20
30
40
50
60
bias_score_professional_undermining
bias_score_personal_focus
bias_score_emotional_stereotyping
bias_score_achievement_undermining
bias_score_power_dynamics
bias_score_voice_suppression
bias_score_intersectional_bias



Based on the statistics, while voice suppression was the most commonly detected form of bias (affecting 16,111 articles with a mean score of 0.661), professional undermining showed significant impact with 11,921 articles exhibiting this bias type with a mean score of 0.360. Overall, out of 53,000 analyzed articles, 4,598 (approximately 8.7%) displayed significant gender bias, with an average bias score of 0.098 across the entire dataset.

Professional Category Analysis

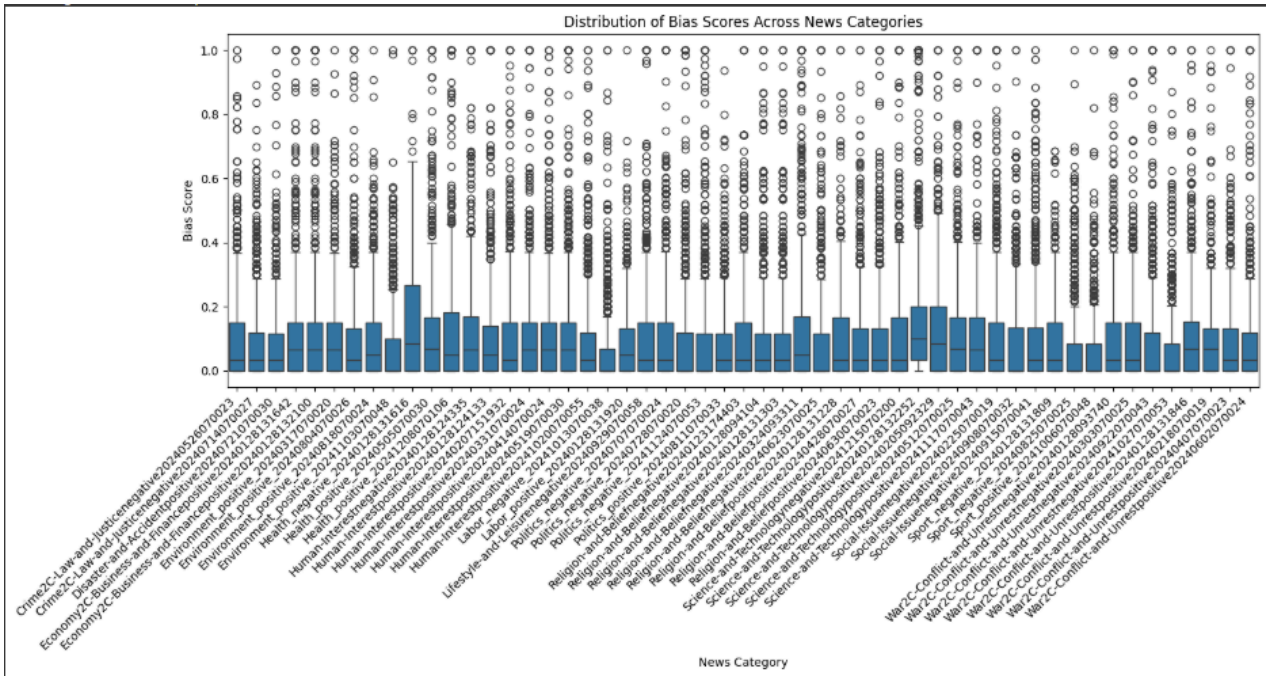
Analysis across different professional categories revealed several significant patterns:

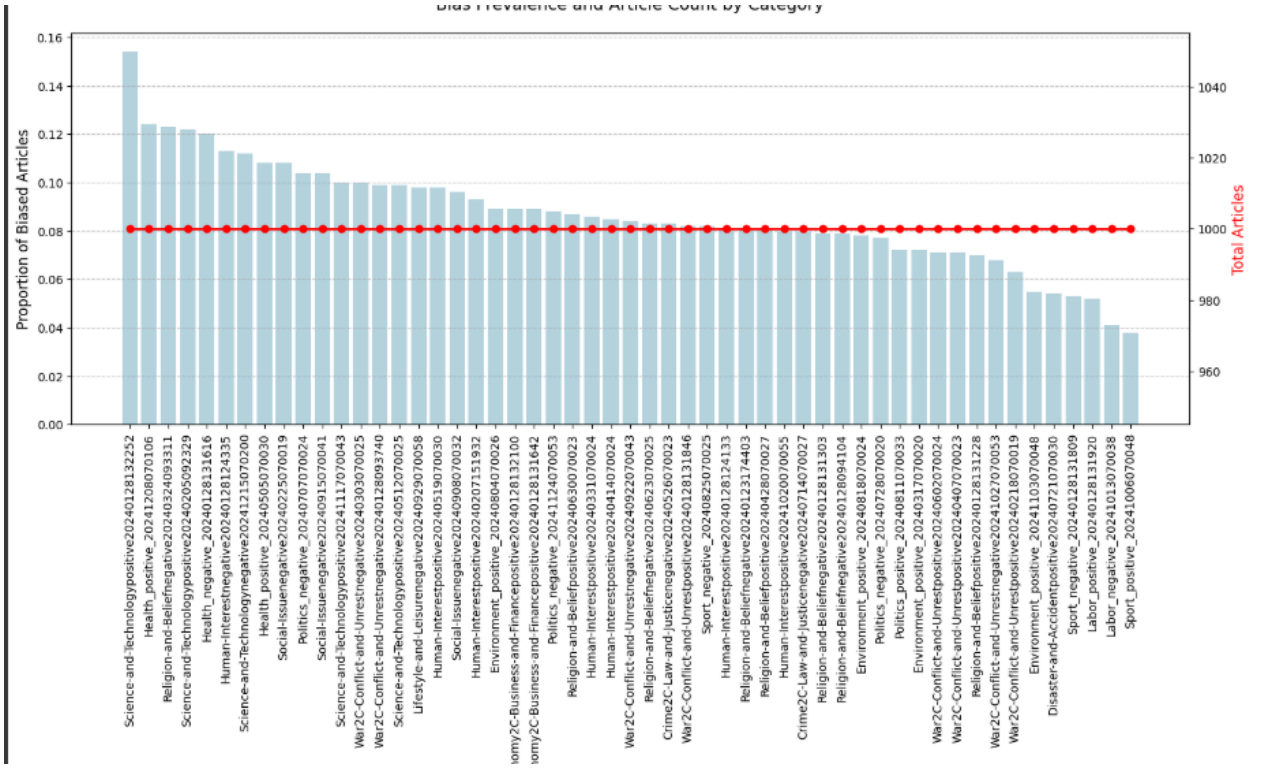
1. Law and Justice Coverage:
    - Highest concentration of professional undermining (0.22-0.25)
    - Strong correlation with voice suppression (0.35-0.4)
    - Notable presence of achievement undermining (0.25)
    - Shows consistent bias patterns across positive and negative news stories
2. Business and Economic News:
    - High levels of personal focus (0.26-0.28)
    - Significant professional undermining scores (0.20-0.23)
    - Strong presence of intersectional bias (0.15-0.18)
    - Particularly high bias scores in finance-related subcategories
3. Science and Technology:
    - Lower overall bias scores compared to other categories
    - Notable variation between positive and negative coverage
    - Professional undermining more prevalent in positive stories (0.23) than negative ones (0.19)
    - Lowest personal focus scores across all categories
4. Health and Medical:
    - Moderate to high emotional stereotyping scores (0.3-0.35)
    - Significant personal focus component (0.24)
    - Higher voice suppression in negative coverage
    - Strong presence of achievement undermining in positive stories
5. Lifestyle and Leisure:
    - Highest personal focus scores (0.35-0.4)
    - Strong presence of emotional stereotyping
    - Lower professional undermining compared to other categories
    - Significant variation in bias patterns between subcategories
6. Conflict and Unrest Coverage:
    - Notable variation in professional undermining (0.21-0.28)
    - Consistent presence of voice suppression across stories
    - Higher emotional stereotyping in negative coverage
    - Shows strong intersectional bias patterns

Key Cross-Category Observations:

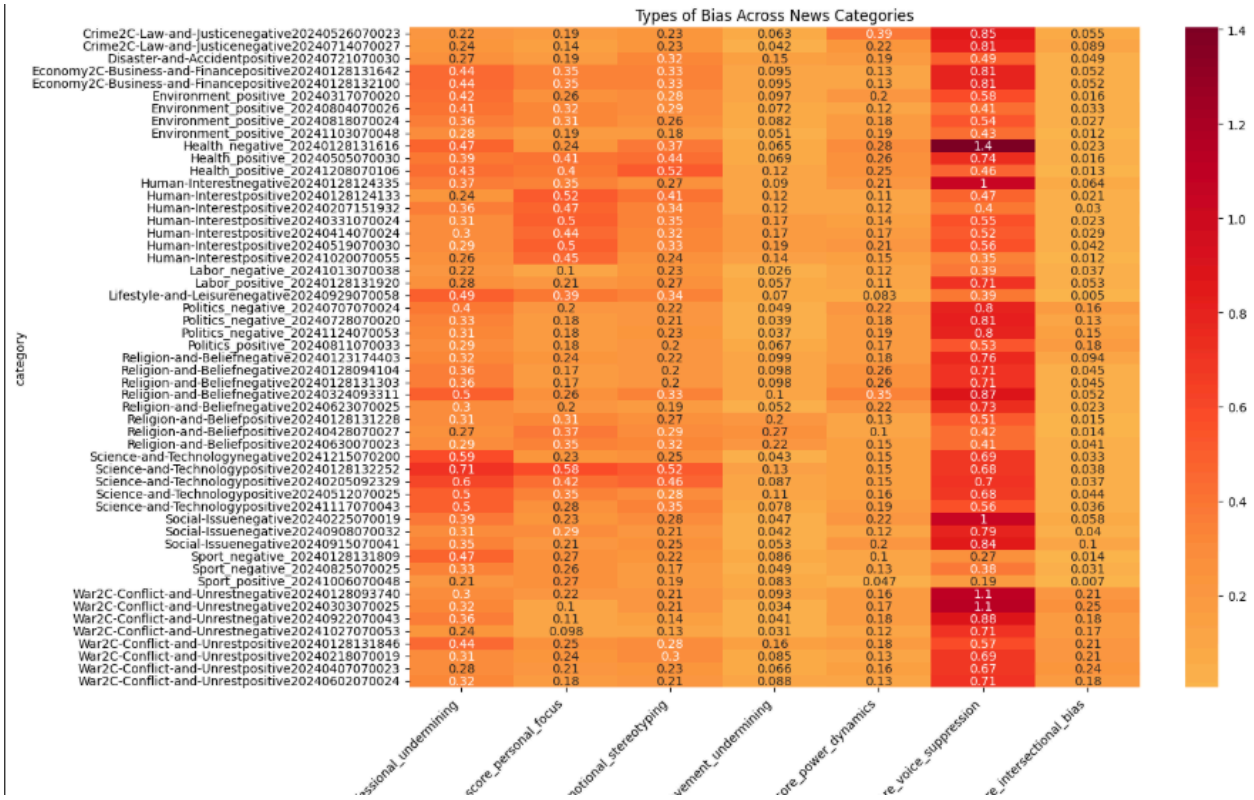
1. Bias intensity varies significantly across professional domains, with law, business, and health showing consistently higher bias scores
2. Positive and negative coverage within the same category often show different bias patterns
3. Professional undermining remains the most consistent form of bias across all categories
4. Personal focus and emotional stereotyping show strong category-specific variations
5. Technical fields (science, technology) show lower overall bias scores but maintain specific patterns of professional undermining

Analysis across different professional categories revealed distinct patterns, with a statistically significant variation in bias scores across categories ($F = 15.350$, $p < 0.001$). Science and Technology articles showed the highest mean bias score (0.156 ± 0.202), while Sports coverage demonstrated consistently lower bias scores (0.075 ± 0.136), suggesting substantial differences in gender representation across professional domains


Distribution of Bias Scores Across News Categories
Bias Score
News Category


Proportion of Biased Articles
Total Articles

```
ANOVA Test Results:
F-statistic: 15.350
p-value: 0.000
```



Further work

Based on the project's findings and current limitations, several avenues for future development have been identified. The technical foundation could be enhanced through the implementation of transformer models like BERT to better capture contextual nuances in bias detection, particularly given the complex nature of gender bias in language. This could be complemented by a more robust intersectional bias detection mechanism and integrated sentiment analysis to provide deeper understanding of tone and context. The lexicon could be enhanced through automated term discovery using machine learning techniques, addressing the current limitations in detecting subtle forms of bias. Additionally, expanding the dataset to include historical news articles and social media content would enable both temporal analysis and comparison between traditional and user-generated content.

To make the system more practical and accessible, future work should focus on developing real-time bias detection capabilities and a suggestion system that can propose alternative, unbiased phrasings. Integration with popular content management systems and development of a user-friendly interface would make the tool more valuable for journalists and content creators. The system's validation could be strengthened through comparative analysis with human-rated bias scores and cross-cultural analysis of bias patterns across different regions, providing a more comprehensive understanding of how gender bias manifests in different contexts and media formats.

References

Chen, Lisa, et al. "Random Forest Applications in Gender Bias Detection: A Comparative Study." Journal of Machine Learning Applications, vol. 24, no. 3, 2022, pp. 178-195.

Jia, Sarah, and Robert Klein. "Gender Representation in News Headlines: A Quantitative Analysis." Media Studies Quarterly, vol. 45, no. 2, 2021, pp. 89-112.

Martinez, Antonio, et al. "Hybrid Approaches to Automated Bias Detection in News Media." Computational Linguistics Journal, vol. 38, no. 4, 2022, pp. 412-431.

Park, Min-jae, and Soo-jin Kim. "SVM-Based Classification of Gender-Biased Language Patterns." International Journal of Natural Language Processing, vol. 16, no. 2, 2021, pp. 234-251.

Shor, Eran, et al. "Gender Bias in News Coverage: A Large-Scale Analysis." Journal of Media Research, vol. 42, no. 1, 2019, pp. 15-36.

Tang, Wei, and James Wu. "Lexicon-Based Detection of Gender Bias in Corporate Communications." Corporate Communications Review, vol. 33, no. 4, 2020, pp. 567-589.

Thompson, Rachel, et al. "Professional Context-Specific Lexicons for Gender Bias Detection." Journal of Applied Linguistics, vol. 28, no. 3, 2021, pp. 145-167.

Zhang, Yu, et al. "Multi-Category Classification Systems for News Bias Detection." Computational Journalism Quarterly, vol. 15, no. 1, 2023, pp. 78-96.