Final assignment:Sentence Level Bais Detection in News Article

Bernard (Adala) Wanyande
Leiden University
awanyande@gmail.com

Ekansh Khanulia
Leiden University
jatinkhanulia@gmail.com

1 Introduction

Media being bias has been an important problem in both journalism studies and natural language processing, as biased reporting can shape public opinion, influence political alignments, and affect democratic decision-making. Prior work on biased sentence detection, including the dataset introduced by Lim et al. (2020), focuses on identifying whether a sentence is biased based on human annotations. However, such annotations often capture multiple forms of bias simultaneously, without distinguishing *why* a sentence is perceived as biased.

In real-world news writing, not all bias serves the same purpose. Some sentences are labeled as biased because they contain emotionally charged or impactful words that intensify the message or are used to make the news article concise. While such language increases emotional salience, it does not necessarily mean that the intention is to persuade or manipulate the reader. Other sentences, however, exhibit bias through framing and structuring of information/words, where facts are selectively presented, emphasized to persuade opinion.In such examples we cannot get rid of baisess even with removing impactful in the article . From a societal perspective, this second form of bias is more problematic, as it reflects deliberate persuasion .

The central motivation of this work is to distinguish between these two forms of bias. While prior work treats bias as a single phenomenon, we argue that separating *emotion-driven bias* from *framing-driven bias* is crucial for understanding persuasive intent in news language. Since author intent cannot be directly observed through the dataset, we approach this challenge computationally by modeling bias intensity from two complementary perspectives: emotional tone and lexical–structural patterns. By comparing human bias judgments with predictions from tone-based and content-based models, we aim to identify sentences whose perceived bias cannot be explained by emotional language alone. Such sentences serve as a proxy for intent-like, framing-driven bias, which aligns more closely with persuasive intent than with expressive style.

1.1 Research Question

This study is guided by the following research question:

> **To what extent can sentences perceived as biased due to emotionally strong wording be distinguished from sentences whose bias arises from ideological framing in the dataset 3 refer?**

By addressing this question, this work moves beyond binary bias detection and contributes to a more nuanced understanding of how bias manifests in news text. It explicitly connects a real-world concern intentional persuasion in media to concrete natural language processing methods that aim to separate stylistic emotionality from structural framing.

2 Related work

The computational detection of media bias has evolved from basic frequency-based counts to multi-dimensional models that analyze both linguistic nuances and structural hierarchies. This section talks about six key technical advancements that provide the foundation for distinguishing between emotionality and intentional structural framing.

2.1 The Technical Landscape and Problem Definition

A critical baseline for this field is established by the recent systematic review by Castillo-Campos

et al. , which analyzed 28 relevant peer-reviewed articles from 2019 to 2023. Their findings reveal a significant "heterogeneity" in how media bias is technically defined and operationalized, leading to inconsistencies in outcome measurement across different studies. This lack of a technical consensus on standardized metrics underscores the necessity of research that can explicitly separate different manifestations of bias a core objective of this study.

2.2 Structural Framing and Hierarchical Salience

According to framing theory, biased communicators select and emphasize specific facts to suit a specific narrative. Early work by Morstatter et al. addressed this by developing a multi-level model to identify both the type and polarity of frames within sentences. Their research suggests that these frames are often built through subtle patterns and linguistic redundancies that can be detected by models,when moving beyond traditional "agenda setting" counts of story occurrences.

Advancing this perspective, Yi et al. proposed a tree-structured hierarchical model where the title, body, and individual sentences serve as nodes. Their model introduces a new method to extract "central sentences" to establish primary-secondary relationships between different parts of a news article. This hierarchical approach is significant because it provides a mechanism for identifying framing-driven bias based on structural importance rather than just lexical .

2.3 Decoupling Emotional Tone from Ideological Lean

A significant advancement in identifying these distinct forms of bias is the *Media Bias Detector* introduced by Wang et al. . By integrating LLMs, this tool provides near real-time, granular insights into editorial choices. Crucially, their framework separates "Tone" (the emotional language used) from "Political Lean" (the ideological framing). Their methodology emphasizes that framing involves the selective inclusion or omission of specific details and context, which can mislead readers even in the absence of outright falsehoods.

2.4 Linguistic Mechanisms and Element Networking

The linguistic mechanisms of bias are further explored by Collins and Boyd , who developed an automated method to detect *Linguistic Intergroup Bias* (LIB). Utilizing the *Linguistic Category Model* (LCM), they demonstrate that bias is subtly embedded in the level of abstraction used ,where abstract words are often applied to "ingroup" virtues and "outgroup" vices. Their research provides a technical bridge to understanding how implicit attitudes are revealed through linguistic asymmetries that go beyond simple sentiment.

Finally, Jiang et al. proposed a different approach that identifies frames as packages of specific "framing elements," such as actors and topics. By using community detection in framing element networks, this method moves beyond traditional topic-focused keywords to uncover the high-level associations that define a news frame. This unsupervised approach provides a robust way to map the structural "frame packages" that journalists use to shape public thinking.

3 Data

We use a sentence-level news bias dataset(Dataset 3 from the assignment) designed for *fine-grained bias analysis* in news articles. Each instance in the dataset corresponds to a news article with sentence-level bias annotations. The task supported by this dataset is **sentence-level bias classification**, where the goal is to assign a bias intensity label to individual sentences. Bias is defined as statement-level , arising from framing choices, word selection, and emphasis, rather than from outlet-level ideology alone.It is organized around four real-world news events, identified by the columns `id_event` and `event`:**NFL player protests** (`event = NFL`) ,**Facebook–Cambridge Analytica data misuse** (`event = Facebook`) , **U.S.–North Korea diplomatic tensions** (`event = NorthKorea`) , **Dan Johnson suicide case** (`event = Johnson`)

Each article is associated with a publication date (`date_event`) enabling relative bias analysis by comparing how different outlets describe the same real life event.

Each news article is uniquely identified by `id_article` and linked to source-level metadata, including the publishing outlet (`source`), its political leaning (`source_bias`), and the article URL (`url`). This information supports analysis of how bias relates to source characteristics.

For each event, a reference article from a

widely regarded neutral news agency is provided. Reference related fields include `ref`, `ref_url`, `reftitle`, and `reftext`. These reference articles provide contextual grounding and serve as a neutral, allowing bias in target articles to be judged relative to a shared factual description of the event. The target article's content is preserved using `doctitle` (headline) and `docbody` (full article text).Articles are segmented into sentences, including the title and the first twenty sentences of the article body. Sentence-level bias labels are stored in column `t` for the title and in columns `0–19` for the corresponding body sentences. The actual sentence texts are stored in `s0–s19`. Bias is annotated at two levels. The column `article_bias` captures the perceived overall bias of the article, while sentence-level annotations (`t`, `0–19`) use a four-point ordinal scale: *neutral*, *slightly biased*, *biased*, and *very biased*. This ordinal formulation reflects the non-binary nature of bias intensity. The column `preknow` records whether annotators had prior knowledge of the news event. This captures the subjective nature of bias perception and enables analysis of disagreement related to reader background.

3.1 Challenges

Subjectivity and annotator disagreement. Bias perception is subjective and influenced by individual sensitivity and background knowledge. This is reflected in disagreement among annotators, even when judging the same sentence, and is explicitly captured by the presence of multiple annotations and the `preknow` field.

Context dependence. Sentences cannot be reliably judged in isolation. Understanding whether a sentence is biased often requires knowledge of the broader event and comparison with the reference article. This makes purely sentence-level modeling challenging without contextual information.

Ordinal label ambiguity. The four-point bias scale introduces fuzzy boundaries between adjacent classes (e.g., *slightly biased* vs. *biased*). These soft distinctions increase classification difficulty and can lead to confusion during training and evaluation.This challenge is taken care in the preprocessing stage and baseline model implementation

3.2 Exploratory Data Analysis

Table **??** shows that most sentences are labeled as *biased* or *very biased*, with very few labeled *neutral* across all events. Facebook and North Korea articles have higher proportions of strongly biased content, while Johnson and NFL events show a more even distribution. This indicates variation in framing intensity depending on the event.

Table **??** reveals that bias intensity is highest in the article title and tends to decrease over sentence positions. This suggests that articles often frontload subjective or emotionally charged language in early parts of the text.

Table **??** compares annotators with and without prior knowledge of the event. Annotators without prior knowledge assign higher average bias scores, suggesting unfamiliar readers may perceive more bias, while knowledgeable annotators interpret content more neutrally.

These patterns highlight the influence of event type, sentence position, and annotator familiarity on perceived bias.

4 Methods

4.1 Preprocessing

In preprocessing we transform the dataset into a structured format for sentence-level bias classification.The original dataset has sentence annotations spread across multiple columns and included noise (e.g., HTML tags, whitespaces,NaN values), so the preprocessing was done to clean textual input, consistent label alignment, and annotator agreement analysis.

We began by standardizing the textual content across all fields—removing HTML tags, unescaping encoded entities, and normalizing whitespace. This ensured that the models would not be affected by these formatting inconsistencies, especially critical in bias detection tasks where subtle phrasing matters. The dataset originally stored sentences as (s0–s19) alongside positional labels (0–19) and title-level labels (t). To support position-aware modeling and per-sentence analysis, we converted this wide format into a long format, where each row represented a single sentence with its metadata, position, and label. This restructuring was essential to enable structured aggregation. Another motivation was to preserve and analyze annotator disagreement. Each sentence have multiple bias labels, and instead of collapsing them prematurely, we retained all annotations

Table 1: Event-wise sentence-level bias distribution (count and percentage).

Event	Neutral	Slightly Biased	Biased	Very Biased
Facebook	260 (21.7%)	424 (35.4%)	452 (37.8%)	61 (5.1%)
Johnson	323 (30.8%)	295 (28.1%)	299 (28.5%)	133 (12.7%)
NFL	426 (35.0%)	440 (36.1%)	286 (23.5%)	66 (5.4%)
North Korea	344 (32.8%)	393 (37.4%)	235 (22.4%)	78 (7.4%)

Table 2: Average bias intensity by sentence position (0–3 scale).

Position	Mean Bias	Std. Dev.	N
Title (t)	1.321	1.016	215
0	1.335	0.942	215
1	1.219	0.944	215
2	1.088	0.946	215
3	1.167	0.932	215
4	1.084	0.882	215
5	1.191	0.930	215
6	1.140	0.906	215
7	1.121	0.909	215
8	1.144	0.887	215
9	1.112	0.900	215
10	1.144	0.949	215
11	1.158	0.963	215
12	1.153	0.896	215
13	1.116	0.917	215
14	1.098	0.904	215
15	1.084	0.923	215
16	1.042	0.898	215
17	1.000	0.932	215
18	1.009	0.927	215
19	1.042	0.963	215

Table 3: Effect of annotator prior knowledge on perceived bias.

Pre-knowledge	N	Mean Bias	Variance	Std. Dev.
No	156	1.2189	0.2884	0.5370
Yes	59	0.9015	0.3263	0.5712

and computed both majority vote and label agreement scores. These scores helped us filter training data based on annotation reliability (e.g., thresholding agreement 0.75) .

4.2 Baseline

The original 4-point scale (neutral, slightly biased, biased, very biased) were transformed into binary categories : a **strict** mapping and an **inclusive** mapping.

The strict mapping grouped labels 1 (neutral) and 2 (slightly biased) into a single *non-biased* class (0), and labels 3 (biased) and 4 (very biased) into the *biased* class (1). This approach turns this task as a classification task and emphasizes high-precision detection by focusing only on clearly biased instances. It was primarily used to train models (Logistic Regression and Linear SVM) and evaluate performance under rigorous definitions of bias. However, initial experiments using this conservative mapping yielded poor classifiers performance where the dataset became heavily skewed toward the unbiased class (results). This imbalance caused classifiers to predict only the dominant class. As a result, while overall accuracy remained high—due to correctly predicting the majority class—precision, recall, and F1 score dropped to zero in case of logistic regression , reflecting the model's failure to identify any true positives from the minority class.

So we tried using inclusive mapping where we treat label 1 as *non-biased* (0), and grouped 2, 3, and 4 into the *biased* class (1).In this case models showed strong performance which is again largely due to the class imbalance introduced by the inclusive mapping (571 biased vs 33 non-biased samples). The dominance of the biased class makes it easier for the model to correctly classify most examples, especially when recall is prioritized. The inclusive mapping also reduces label ambiguity by merging moderately and strongly biased annotations, making the decision boundary simpler and more learnable. So therefore we then tried both the mappings with classweight='balanced' setting which help the model focus more on minority class (results).

Together, these mappings were essential for stress-testing the model across different assumptions and ensured that evaluation results were not contingent on a single threshold. Finally, we produced an EDA summary containing label distributions and annotator agreement patterns. These summaries guided both our experimental choices (e.g., thresholds for training data selection) and our understanding of dataset noise.And then we

tested Roberta on the same dataset

4.3 Tone vs Framing

In the final stage of our analysis, we move beyond binary bias classification to explore how and why a sentence is perceived as biased. Specifically, we aim to address our central research question: to what extent can sentences perceived as biased due to emotionally strong wording be distinguished from those whose bias arises from ideological framing? Earlier stages treated bias as a binary label — either present or absent — but this oversimplifies the nuanced ways bias can manifest. In this phase, we model bias as a continuous intensity score (derived from annotator labels), and contrast two types of linguistic signals: tone-based features (e.g. sentiment, subjectivity) and textual content features (via TF-IDF). By comparing how well each signal predicts perceived bias intensity, we aim to uncover whether certain sentences are judged as biased more due to emotional tone, or more due to subtle framing strategies — and quantify that distinction. We then compare two sets of input features to predict this continuous bias score: one based on the textual content (TF-IDF vectors), and another capturing tonal cues (sentiment and subjectivity via VADER). Both feature sets are used to train Ridge regression models. We evaluate their predictive performance using mean absolute error (MAE) and root mean squared error (RMSE). The goal is to assess whether content or tone is more informative for estimating perceived bias.

Building on this, we also explore the distinction between tone-driven and framing-driven bias. We hypothesize that some sentences may appear neutral in tone but are still perceived as highly biased due to their framing. To detect such cases, we compare the predictions of the TF-IDF-based model against the tone-based model across all reliable sentences. A large gap between these two predictions suggests that the sentence may be persuasive or biased through its framing rather than tone. Sentences with high human bias scores, high TF-IDF predictions, and low tone predictions are flagged as potential intent-driven cases. We also experiment with a percentile-based threshold, selecting the top 10 percent of gaps among sentences rated as highly biased by annotators. These selected sentences serve as a proxy set for exploring intent-like framing in language, providing deeper insight into how bias may manifest beyond surface-level sentiment.

5 Results

Table 4: Label distribution after filtering sentences using a reliability threshold (at least 3 annotators and sufficient agreement.

Mapping Type	Total Rows	Label 0 Count	Label 1 Count
Conservative	485	384	101
Inclusive	604	33	571

Table 5 presents the performance of binary bias classification models across different label mapping strategies and classifiers. The table compares Logistic Regression and Linear SVM models under four settings: Conservative(labels 0 and 1 are mapped to *not biased* i.e 0, while labels 2 and 3 are mapped to *biased* i.e 1), Inclusive(label 0 is mapped to *not biased* i.e 0, while labels 1, 2, and 3 are mapped to *biased* i.e 1), Conservative with balancing, and Inclusive with balancing. For each combination, standard classification metrics are reported, including Accuracy, Precision, Recall, and F1 Score. The table shows how different mapping strategies and class balancing impact each model's ability to distinguish biased from unbiased sentences.

Table 6 shows the performance of Ridge regression models trained on two feature types: TF-IDF vectors representing textual content and tone features derived from VADER sentiment scores. The model trained on TF-IDF achieved a lower Mean Absolute Error and Root Mean Square Error compared to the tone-based model .

To detect subtle forms of bias not rooted in emotional polarized language, we introduce a proxy for intent-based bias by measuring the gap between two predictive models: one based on TF-IDF features (word usage patterns) and another based on tone features (sentiment and subjectivity). Sentences where the TF-IDF model strongly overestimates the bias compared to the tone model are likely to exhibit framing-driven bias — where the structure and phrasing imply bias rather than emotionally charged words. By selecting the top 20 percentage of such high-gap sentences, we aim to isolate cases where intent-based bias may be present.

To detect intent-driven bias beyond emotional tone, we identify sentences where the TF-IDF model predicts significantly higher bias than the tone model. Sentences above the 80th percentile

Table 5: Binary Bias Classification Results Across Mapping Strategies and Classifiers

Mapping Type	Model	Accuracy	Precision	Recall	F1 Score
Conservative	Logistic Regression	0.794	0.000	0.000	0.000
Conservative	Linear SVM	0.773	0.250	0.050	0.083
Inclusive	Logistic Regression	0.942	0.942	1.000	0.970
Inclusive	Linear SVM	0.959	0.958	1.000	0.979
Conservative + Balanced	Logistic Regression	0.784	0.462	0.300	0.364
Conservative + Balanced	Linear SVM	0.773	0.400	0.200	0.267
Inclusive + Balanced	Logistic Regression	0.950	0.966	0.982	0.974
Inclusive + Balanced	Linear SVM	0.934	0.957	0.974	0.965
Inclusive	**RoBERTa**	**0.9752**	**0.9752**	**1.0**	**0.9874**

Table 6: Component B: Ridge Regression Performance on TF-IDF and Tone Features

Feature Set	MAE	RMSE
TF-IDF (text content)	0.3828	0.4755
Tone (VADER scores)	0.4118	0.4916

Table 7: Component C: Intent-like Proxy Summary (Top 20% Gap)

Source Bias	Count	Intent-like Rate
Left-center	26	0.124
Left	3	0.033
Right	4	0.023
Total	**33**	-

Table 8: Summary of intent-like proxy detection using the percentile gap method.

Metric	Value
Total reliable sentences	601
Human-biased sentences (mean_label ≥ 2.5)	165
Intent-like proxy sentences	33
Gap percentile threshold	80th

gap are marked as intent-like proxies. Summary statistics are shown in Table 8.

The sentences in Table 9 were selected as the top 5 intent-like bias cases, based on the largest gaps between TF-IDF and tone-based predictions. These high-gap instances suggest framing-based bias—where the structure and emphasis imply bias without relying on overtly emotional or polarizing words.

6 Discussion

Table 4 highlights the severe class imbalance in both the Conservative and Inclusive mappings. Under the Conservative mapping (labels 1–2 $\rightarrow$ 0; 3–4 $\rightarrow$ 1), only 101 out of 485 filtered samples were labeled as biased. This imbalance heavily affected classifier performance, as seen in Table 5: both Logistic Regression and SVM exhibit poor recall and F1 scores (F1 = 0.000 and 0.083, respectively), despite decent accuracy. The models predominantly predicted the majority class (unbiased), failing to learn meaningful bias signals. To mitigate this, we applied class balancing to improve attention of the model on minority class. The Conservative + Balanced results show modest improvements: Logistic Regression achieved an F1 of 0.364, and SVM improved to 0.267. While this confirms some effectiveness of weighting, performance remained limited due to the intrinsic sparsity of biased examples. The Inclusive mapping (labels 0 $\rightarrow$ 0; 1–4 $\rightarrow$ 1) reverses the imbalance here, 571 out of 604 samples are labeled as biased (Table 4),this causes the model to overfit to majority class,leading to inflated accuracy. Linear SVM and Logistic Regression achieved F1 scores above 0.97. Even without balancing, the dominant biased class provided sufficient signal to learn a decision bound-

Table 9: Top 5 sentences flagged as intent-like bias by the percentile-gap method. These have a high difference between TF-IDF and tone predictions, indicating framing rather than lexical bias.

ID	Sentence	Mean Label	TF-IDF Pred	Tone Pred
11	Asked about the Russian oil company, a spokesman for SCL said that in 2014 the firm's commercial division "discussed helping Lukoil Turkey better engage with its loyalty-card customers at gas stations."	3.4	2.831	2.073
19	The tumult began Monday, when the Kentucky Center for Investigative Reporting published allegations that Johnson sexually assaulted his daughter's friend during a sleepover in 2013.	3.75	2.905	2.178
12	Who knows what Trump's real gripe is with the NFL.	3.5	2.845	2.132
6	How the NFL responds to Donald Trump's spit-foaming is hardly a test case for whether the republic will stand.	3.5	2.807	2.132
7	Nevertheless, the league is a maker of manners in this country, so it means something that Commissioner Roger Goodell and others are getting it right, striking the perfect calm but resistant tone in response to Trump's gutter-mouthing, a tone that sa[...]	3.5	2.756	2.105

ary. Adding class weights in Inclusive + Balanced gave marginal improvements, with F1s remaining above 0.96. This suggests that when the biased class dominates, weighting is not as crucial.

On otherhand RoBERTa outperforms all traditional metods, achieving **F1 = 0.9874** and **Accuracy = 0.9752** under the *Inclusive* mapping. This confirms its strong capability to learn semantic nuances and framing bias, not just keyword presence.

The results presented in Tables 6,7,8,9 collectively address the research question of whether bias can be detected beyond emotionally charged language, capturing intent or framing driven bias rather than purely lexical sentiment.

Table 6 compares Ridge regression models trained on TF-IDF vectors, capturing word usage and textual structure, and tone features derived from VADER, capturing sentiment polarity and subjectivity. The model trained on TF-IDF achieves lower Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) than the tone-based model. This indicates that textual content provides a more reliable signal for predicting annotator-assigned bias intensity than sentiment alone. These results suggest that bias in the dataset is not primarily driven by emotional language, but by how information is phrased and framed.

To further isolate framing-driven bias, we analyze cases where the TF-IDF model predicts substantially higher bias than the tone-based model. The summary in Table 8 shows that, out of 601 reliable sentences, 165 are considered biased by human annotators that has meanlabel greater than equal to 2.5, yet only 33 sentences exceed the 80th percentile gap threshold. This demonstrates that intent-like bias represents a narrower subset of biased content, reinforcing the distinction between emotional bias(VADER) and framing-based bias.

Table 7 presents the distribution of these intent-like proxy sentences across source bias categories. A higher concentration is observed in left-center sources, suggesting that editorial framing styles, rather than sentiment intensity, may play a stronger role in conveying bias in certain outlets. This pattern should be interpreted with caution, as it reflects stylistic tendencies in language use rather than direct ideological causation.

The qualitative examples shown in Table 9 further support this interpretation. These sentences exhibit moderate tone-based predictions while consistently receiving higher TF-IDF-based bias estimates. This discrepancy indicates that bias arises through rhetorical emphasis, implication, or phrasing rather than explicit emotional wording. Such cases illustrate framing-driven bias, where meaning is conveyed implicitly through structure and narrative choice.

Several limitations should be acknowledged. The percentile-gap threshold used to identify intent-like bias is heuristic, and alternative thresholds may yield different subsets of sentences. Additionally, VADER provides a coarse representation of tone and may fail to capture subtle affective cues. While TF-IDF offers interpretability, it cannot model deeper discourse structure or long-range semantic dependencies.

Overall, the results across Tables 6,7,8,9 demonstrate that contrasting lexical and tone-based models provides a practical approach to approximating intent-driven bias. This highlights the limitations of sentiment-only methods and motivates the use of more expressive models for capturing framing effects in political and news text.

7 Conclusion

This study addressed the research question: *to what extent can sentences perceived as biased due to emotionally strong wording be distinguished from sentences whose bias arises from ideological framing?* Our results show that these two sources of perceived bias are meaningfully separable. While binary bias classification achieves high performance under inclusive mappings, these results are strongly influenced by label imbalance and provide limited insight into why a sentence is perceived as biased. By modeling bias as a continuous intensity signal, we find that content-based features (TF-IDF) more closely align with human bias judgments than tone-based features derived from sentiment, indicating that perceived bias in the dataset is driven more by framing and lexical choice than by emotional language alone. Further, the gap between content- and tone-based predictions enables the identification of sentences that appear neutral in tone yet are judged as highly biased, providing empirical evidence of intent-like, framing-driven bias. More broadly, these findings suggest that bias detection systems should move beyond sentiment-focused approaches and incorporate representations that capture framing and semantic structure. Future work could extend this analysis using discourse-level models, richer intent representations, and more balanced datasets to further refine the distinction between emotional and framing-based bias.

8 Reflection

8.1 Contribution

We divided the workload and ensured continuous discussion and synchronization throughout the process.

- **Bernard (Adala) Wanyande**: Responsible for data preprocessing (`process.py`) and implementing the binary classification models (`baseline.py` and `baselinesota.py`). This included designing the conservative and inclusive label mappings, running experiments with logistic regression and linear SVM, and reporting the metrics in structured result tables.

- **Ekansh khanulia**: Handled the exploratory data analysis (`eda.py`) and the bias intensity regression and proxy detection components (`biasintensity.py`). This involved computing sentence-level mean bias scores, evaluating the TF-IDF and tone-based regression models, and detecting intent-like bias using the percentile-gap method.

The Research Question ,entire report and codebase were jointly written and reviewed to ensure coherence and quality.

References

Mar Castillo-Campos, David Becerra-Alonso, and Hajo G. Boomgaarden. 2025. Automated detection of media bias using artificial intelligence and natural language processing: A systematic review. *Social Science Computer Review*, 0(0):1–20.

Katherine A. Collins and Ryan L. Boyd. 2025. Automating the detection of linguistic intergroup bias through computerized language analysis. *Journal of Language and Social Psychology*, 44(3-4):343–366.

Yanru Jiang, Sha Lai, Lei Guo, Prakash Ishwar, Derry Wijaya, and Margrit Betke. 2025. Exploring an alternative computational approach for news framing analysis through community detection in framing element networks. *Journalism & Mass Communication Quarterly*, pages 1–32.

Fred Morstatter, Liang Wu, Uraz Yavanoglu, Steven R. Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1.

Jenny S. Wang, Samar Haider, Amir Tohidi, Anushkaa Gupta, Yuxuan Zhang, Chris Callison-Burch, David Rothschild, and Duncan J. Watts. 2025. Media bias detector: Designing and implementing a tool for real-time selection and framing bias analysis in news coverage. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM.

JinCheng Yi, ShaoHua Jiang, and QiPeng Wen. 2025. Detecting news bias with sentence salience and hierarchical structures. *Hunan Normal University Technical Report*.