Where Does Wikipedia Stand? A Simple Cross-Source Bias Checker Across Broad Topics

Nathan Watkins, Will Racz

November 19, 2025

Abstract

This project compares Wikipedia to several commonly perceived "neutral" information sources using a lightweight bias-checking model. Instead of trying to label content as left, right, or unbiased, the goal is to surface simple indicators like (1) source diversity, (2) repeated framing or stylistic patterns, and (3) automation-like structure. We collect around 15 articles per topic from five major outlets plus Wikipedia, and because the topics are broad, some sources have missing entries (NA). The model applies consistent scoring to all articles and allows us to compare where Wikipedia falls relative to other sources.

1 Introduction

People often argue about whether Wikipedia is neutral or biased, but it's hard to compare it directly to news articles or other "trusted" information sources. At the same time, most online writing—whether it's Wikipedia, RSS feeds, or major news outlets—follows patterns that can unintentionally reveal bias, such as limited sources, repeated framing terms, or automated structures.

The purpose of this project is to build a small, transparent bias-checking tool and use it to compare Wikipedia to five major information sources across several broad topics. The tool looks at structural features rather than political categories, since those are more measurable and easier to apply consistently.

2 Related Work

Most research on media bias focuses on political leaning or ideological labeling, but a lot of recent work argues that structural patterns—like how articles phrase things, what sources they rely on, and whether writing seems automated—can also shape how information is interpreted. Studies on source diversity show that articles drawing from a narrow set of domains can unintentionally reinforce certain viewpoints. Research on automation bias shows that people tend to trust consistent, template-like writing even when it is produced by bots or algorithmic processes.

None of these studies look directly at Wikipedia vs. news sources in the way we do, but they motivate our use of proxies such as source diversity, framing repetition, and automation signals.

3 Methodology

3.1 Motivating vs. Operationalized Research Questions

The motivating question behind the project is:

Where does Wikipedia stand relative to other sources that people usually consider "unbiased"?

This question is too broad to measure directly, so we operationalize it as:

Across a set of broad topics, how does Wikipedia's source diversity, language framing, and structural uniformity compare to five mainstream information sources?

This version lets us compute actual numbers and compare across outlets.

3.2 Data Collection and Sources

For each topic (like climate, health, technology, etc.), we collected about 15 articles per outlet when possible. Wikipedia contributes one main article per topic. Because topics are broad, some sources naturally had missing results (NA), which we simply excluded from averaging.

All code for scraping, parsing, and scoring is available at:

https://github.com/NateWatkins/WikipediaBias

3.3 Why We Use Proxies Instead of Measuring "Bias" Directly

Bias and neutrality are abstract ideas, so instead we measure smaller, observable indicators ("proxies") that relate to them. We use three proxies:

3.3.1 1. Source Diversity

Counts how many different external domains appear in the article:

$$\text{SourceDiversity} = |\text{distinct domains}|$$

More domains usually means broader information coverage.

3.3.2 2. Framing / Language Patterns

Repeated phrasing or emotional wording can signal subtle bias. We track:

- repeated sentence stems,
- repeated descriptive adjectives,
- emotionally loaded keywords.

$\text{FramingScore} = \text{RepeatedPhrases} + \text{LoadedTerms} \ \text{TotalSentences}$

3.3.3 3. Automation / Structural Uniformity

Template-like writing and low sentence-length variance can signal automation or very strict editing structure. We track:

- low variance in sentence length,
- near-duplicate sentences,
- repeated paragraph openings.

3.4 Measurement Pipeline

Every article will go through the same steps:

1. Load or scrape the full text.

2. Extract visible links, sentences, and structural elements.

3. Compute the three proxy scores.

4. Store everything in JSON for reproducibility.

3.5 Comparison Strategy

Once all scores are collected, we compare Wikipedia directly against the average of the five other outlets. Comparisons are done:

- within each topic first, and

- then across topics (NA values ignored).

For each proxy:

$$\text{Difference} = \text{WikipediaMean} - \text{SourceGroupMean}$$

A positive or negative difference tells us which side of the range Wikipedia tends to fall on.

3.6 How This Connects Back to the Big Question

If Wikipedia shows wider source diversity, lower framing repetition, or different automation scores, that helps position it relative to other outlets. The goal is not to label Wikipedia as unbiased or biased, but to understand how its structure compares to other widely used information sources.

4 Results

After running the model across the collected articles, we found clear differences between sources:

- Some outlets only cited one or two external domains for multiple topics.

- A few outlets used the same adjectives or phrasing patterns repeatedly across several articles.

- Wikipedia generally had more references but sometimes more structural uniformity, likely because of standardization in its writing style.

- Some AI-assisted or syndicated news pieces showed near-identical paragraph structures.

These patterns lined up with what we expected from manual inspection.

5   Discussion

One of the biggest issues we ran into was uneven access to article content. Some outlets offer full APIs, others only RSS feeds, and some require scraping with caution due to paywalls or terms of service. These differences affected which articles we could include.

Another challenge was the broadness of the topics. Since we picked general themes (like "technology" or "health"), not every source covered every topic during our time window. This created NA values, but instead of forcing coverage, we simply left missing entries out of average calculations.

Overall, the results don't say whether Wikipedia is unbiased, but they do show how it compares structurally to five other information sources.

6   Future Work

Future improvements could include:

- comparing source variety against topic-specific expectations,

- analyzing more subtle linguistic or rhetorical patterns,

- detecting more advanced forms of automated writing,

- generating standardized bias-summary pages for each topic.

7   Conclusion

This project presents a simple bias-checking approach that compares Wikipedia to five other major outlets using structural indicators like source diversity, framing patterns, and automation signals. The tool is intentionally minimal and interpretable, and although it doesn't classify bias directly, it helps highlight where different sources sit in relation to one another.

References

[1] guardrails-ai contributors. bias_check. `https://github.com/guardrails-ai/bias_check`, 2025. Accessed: 2025-11-19.

[2] Parvez Khan. Media-bias-detection. `https://github.com/Parvezkhan0/Media-Bias-Detection`, 2025. Accessed: 2025-11-19.

[3] Nate Watkins. Wikipediabias: Project repository. `https://github.com/NateWatkins/WikipediaBias`, 2025. Accessed: 2025-11-19.