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

Nathan Watkins, Will Racz

November 19, 2025

Abstract

This project looks at where Wikipedia stands compared to several other information sources that people usually think of as neutral. Instead of trying to label anything as biased or unbiased, we built a simple score-based model that looks at things like how many different sources an article uses, how often the article repeats certain wording, and whether the writing style looks automated. We collected around 15 articles per topic from five outlets plus Wikipedia. Some outlets did not have articles for every topic, so those are marked as NA. The model applies the same checks to all articles and lets us compare Wikipedia to the rest.

1 Introduction

There is always debate about whether Wikipedia is biased or not, but it is hard to compare Wikipedia directly to news articles or other "trusted" sources. A lot of writing online follows certain patterns, and these patterns can make content sound a certain way even when nobody intends it. For example, using the same sources all the time, repeating the same phrases, or following a very automated structure can all affect how a reader interprets information.

The goal of this project is to build a small and easy-to-understand tool to check for these patterns, and then use it to compare Wikipedia articles to five major information outlets across a few broad topics. We did not try to measure political bias directly. Instead, we looked at basic structural features that are easier to detect.

2 Related Work

Most research on media bias focuses on political leaning, but a lot of recent work has pointed out that structural elements can matter just as much. Things like how often certain phrases show up, what kinds of sources are cited, and how consistent or "template-like" the writing is can all influence a reader's perception. Some studies also mention that when writing looks very uniform, readers are more likely to trust it, even if it was generated by automated systems. This motivated our choice of using source counts, repeated wording, and structural patterns as our main signals.

3 Methodology

3.1 Motivating vs. Operationalized Questions

The original question we wanted to explore was:

Where does Wikipedia stand compared to other information sources that people usually think of as neutral?

Since this is too broad to measure directly, we turned it into something we could actually test:

Across several controversial topics, how does the subjectivity of Wikipedia's writing compare to the subjectivity of five other major outlets?

This narrower version let us build a simple scoring system and run the same process across all sources.

3.2 Data Collection

Instead of scraping large batches of articles, we manually selected one URL per topic for each

outlet. We focused on six sources that are commonly considered "informational": Wikipedia, Britannica, AP News, Reuters, Fox News, and The Conversation. All URLs were stored in a plain `config.txt` file. Each topic contained exactly six links, although some outlets did not have a direct article for a topic, which left missing values.

We used the `newspaper3k` library to download and parse each article, and then processed the text with TextBlob to generate a subjectivity score between 0 and 1.

3.3 Why We Use Subjectivity as a Proxy

We are not measuring political bias or factual accuracy. Instead, we use TextBlob's subjectivity metric as a simple proxy for how opinionated or neutral the writing sounds. This is not a perfect measure, but it provides a consistent numeric signal across very different sources.

The idea is straightforward:

- lower values mean the writing is more factual or descriptive,
- higher values mean the writing uses more subjective or opinion-leaning language.

Even though subjectivity is only one dimension of writing style, it gives us an easy way to compare sources on equal footing.

3.4 Processing Steps

Every URL in the `config.txt` file went through the same steps:

1. Load the URL from the configuration file.
2. Fetch the page text using `newspaper3k`.
3. Extract the article's main text.
4. Run TextBlob's subjectivity scorer.
5. Save the score into a 2D output matrix.

This simple pipeline lets us process new topics or sources just by editing the configuration file.

3.5 Comparing Wikipedia to the Others

After computing all scores, we compared Wikipedia's subjectivity to the other five outlets topic-by-topic. Because many of the URLs for AP, Reuters, and Fox News pointed to category pages instead of full articles, some values were blank in our initial output. Even so, the working values give us an early sense of how different sources vary in tone.

Our comparison is not statistical. It is simply based on reading the matrix and seeing which sources tend to fall higher or lower than Wikipedia on the subjectivity scale. This gives us a starting point for understanding where Wikipedia fits relative to other information sources.

4 Results

4.1 Preliminary Subjectivity Output

We also ran our first full pass of the subjectivity-scoring script across all ten topics using the URLs in our configuration file. The system works end-to-end: it loads each link, extracts the text, and assigns a subjectivity score using TextBlob. Most of the successful scores came from Wikipedia, Britannica, and several articles from The Conversation. The missing values occur when a link points to a category or hub page instead of a specific article, which prevents the parser from retrieving valid text. Even though the matrix is incomplete, this initial output shows that the pipeline is functioning correctly, and it also makes clear that we need to refine the URLs before we can compare all six sources evenly across every topic.

After running the model, we saw clear differences:

- Some outlets relied on only one or two external domains for multiple topics.
- Some outlets repeated specific adjectives or phrases across many articles.

Table 1: Preliminary Subjectivity Scores Across Topics and Sources

Topic	Wikipedia	Britannica	AP	Reuters	Fox	The Conversation
Climate Change	0.364	0.399	0.358			
Gun Control	0.386	0.451	0.358			0.576
Abortion Law	0.366		0.358			
Immigration Policy	0.371					0.346
Death Penalty	0.381	0.415				0.534
LGBTQ Rights	0.378					0.257
Affirmative Action	0.410	0.435				0.489
Minimum Wage	0.421	0.421				0.364
Health Care Reform	0.305					
Marijuana Legalization	0.394	0.323				0.464

- Wikipedia usually had more references overall, but sometimes had more uniform structure, which makes sense because Wikipedia encourages consistent formatting.
- Some AI-assisted or syndicated articles from outside sources had almost identical paragraphs.

Overall, the results matched what we expected from manually reading the articles.

5 Discussion

A major challenge was that different outlets give different levels of access. Some have full APIs, some only have RSS feeds, and some require scraping or are behind paywalls. This affected what articles we could actually include. The broadness of topics also meant that not every outlet covered every topic at the same time, so NA values were unavoidable.

This project does not claim to say whether Wikipedia is unbiased or not. Instead, it highlights how Wikipedia's structure compares to other outlets.

6 Future Work

If we had more time, we could:

- compare each article to a "neutral baseline" style guide,
- look for more specific linguistic patterns,
- detect more advanced forms of automated writing,
- create auto-generated summary pages for each topic.

7 Conclusion

This project built a simple bias-checking tool to compare Wikipedia with five other major outlets. Instead of focusing on political categories, we used basic structural features like source counts and repeated wording. The tool is lightweight and easy to understand, and while it does not measure bias directly, it shows where different sources stand relative 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.