Detecting Media Bias: Effects of Tool-Based and Cognitive Support on Human Bias Detection

Michael Winter
University of Regensburg
Regensburg, Germany
Michael1.Winterl@stud.uni-regensburg.de

Mark Zänglein
University of Regensburg
Regensburg, Germany
Mark.Zaenglein@stud.uni-regensburg.de

ABSTRACT

This study investigates individuals' ability to detect bias in journalistic texts and examines the effectiveness of different forms of support in enhancing this ability. Participants were divided into three groups: one received assistance from a bias-detection software, another was provided with bias-related definitions, and a control group received no support. Results show that while software-assisted detection initially led to high precision, it also fostered a strong dependency. In contrast, definitional support yielded more consistent and potentially sustainable improvements. The control group's performance remained stable. A significant difference in precision across groups in the second phase—after support was removed—suggests possible learning effects. However, due to the small sample size (N = 20), these findings should be interpreted with caution. Future research should explore these effects with larger samples and over longer periods.

KEYWORDS

Bias detection, Bias, Journalism, Media literacy, Media perception

1 INTRODUCTION

In today's information-rich environment, various media platforms such as broadcasting, cable television, online outlets, and social media have become central tools for accessing information [8]. News media play a crucial role in shaping public opinion and social norms by influencing attitudes and behaviors, as well as promoting transparency [1, 10]. However, these media often exhibit biases influenced by factors such as geography, ideology, institutional affiliations, or the specific medium used. The manifestation of bias is shaped by elements like word choice, omissions, agenda-setting, and the selection of sources [2, 13]. Despite extensive academic research confirming biased reporting, many individuals still perceive news as trustworthy[7, 12]. Additionally, various individual factors influence bias detection, such as demographic data, self-assessment, bias tolerance and private and social prejudices[14]. Recognizing bias in journalistic texts is essential, as it contributes to shaping informed public opinion, reducing polarization through the inclusion of diverse perspectives, and fostering democratic processes by promoting transparency, accountability, and active citizen engagement [11]. Despite the substantial body of theoretical research on media bias recognition, empirical investigations into individuals' ability to detect such biases—particularly through the use of tools like biasscanner.org, designed to identify biased sentences in news articles—remain scarce. This highlights the need for further systematic exploration of how these tools can enhance bias-detection skills and whether individuals can effectively learn and improve this ability. To address this gap, this study investigates whether the method of bias detection (BiasScanner, definitions or personal judgment) influences participants' ability to detect biased content (RQ1), and whether the use of the BiasScanner leads to a measurable learning effect (RQ2).

Our research aims to contribute to this exploration by examining how individuals can identify bias in texts and evaluating methods—especially through the use of tools like *BiasScanner.org*—to improve this skill.

2 RELATED WORK

Media bias is not a phenomenon exclusive to the 21st century, although the advent of the internet and social media has amplified its prevalence and impact [4]. As early as 1950, the relevance of subjective decision-making in news selection was highlighted in White's gatekeeping study. In this case study, White analyzed the decision-making processes of a news editor who selected articles based on personal preferences. This study demonstrated that individual judgments and social influences significantly shape the perception and dissemination of news [15].

Since then, extensive research has been conducted on the detection and categorization of media bias using computational tools. One such tool is BiasScanner, which leverages a trained large language model to classify and identify bias, enabling readers to critically assess the articles they consume. It supports the detection of different bias types, provides explanations, and quantifies the bias intensity in a given text [6].

An early approach to bias detection can be found in the work of Lin et al. (2006), titled "Which Side Are You On? Identifying Perspectives at the Document and Sentence Levels". In this study, the authors developed statistical models to identify the perspective from which a document was written, analyzing bias at both document and sentence levels. Their results demonstrated that machine learning models could successfully recognize how bias manifests in word choices and identify the biased perspective of a document with high accuracy [5].

Beyond BiasScanner, additional tools exist for the automated detection of media bias, utilizing various technological approaches. A widely adopted method involves rule-based systems that rely on predefined linguistic features and keyword-based identification. While these systems often yield transparent and interpretable results, they lack flexibility and struggle to capture contextual nuances[9].

A systematic review of bias detection methodologies underscores the significant room for improvement in automated bias detection. Due to the diversity of existing approaches, there is a pressing need for a standardized benchmark and consistent evaluation metrics[9]. Consequently, relying solely on automated tools for media bias assessment is insufficient; rather, it is essential to develop critical media literacy skills to recognize bias independently.

Hybrid approaches combining multiple methodologies have also emerged. Media Bias Fact Check, for instance, employs both human evaluations and algorithmic analyses to assess the credibility and

potential bias of news sources. Its methodology involves a multi-step evaluation process in which analysts assess articles based on criteria such as word choice, source usage, and factual accuracy. Additionally, external sources and independent fact-checking organizations are consulted to ensure the reliability of the evaluations. The tool categorizes news sources along a political spectrum from left to right and flags potential misinformation or highly biased reporting[3].

Ultimately, the necessity of advancing automated bias detection methods remains evident. However, this recognition also implies that individuals must cultivate their ability to identify media bias independently. Can users of these tools enhance their own competencies in recognizing bias through their assistance? Our research aims to answer this question.

3 RESEARCH METHODOLOGY

3.1 Hypotheses

Based on the research questions, the following hypotheses were formulated to guide the experimental design and analysis:

- **H1:** Participants supported by the BiasScanner will achieve higher precision in bias detection during the support phase compared to the other groups.
- **H2:** Participants receiving definitional support will retain higher precision in bias detection after the support is removed compared to the control group.
- **H3:** The control group will demonstrate the most consistent performance across both phases due to their independent engagement with the material.

3.2 Experimental Setup

To evaluate whether technological support—specifically, the use of the BiasScanner—enhances the ability to detect bias in news articles, an experiment was conducted with three independent groups. A dedicated online platform was developed using React for the frontend and a Neon database for backend data storage. The platform ensured a standardized presentation of the articles and efficient data collection. Participants accessed the study via a unique link, and their classification responses were securely logged.

3.3 Experimental Groups

Participants were randomly assigned to one of three experimental groups, each differing in the level of technological support provided:

- **BiasScanner Assistance Group**: Participants in this group received direct support from the BiasScanner, including explicit feedback and detailed explanations regarding potential bias indicators beneath each text segment. This condition was designed to assess whether comprehensive technological assistance enhances the accuracy of bias detection.
- **Bias Definition Group**: Participants in this group were provided with a set of definitions covering different types of bias, each of which appeared at least once in the given articles. This condition aimed to evaluate the extent to which theoretical knowledge influences bias detection performance.
- **Control Group (No Technological Support)**: Participants in this group classified text segments solely based on their own judgment, without any technological assistance. This condition served as a baseline to assess human bias detection capabilities in the absence of external support.

3.4 Procedure

Each participant completed five trials on the website, classifying text segments as biased or unbiased across five different articles. The articles were presented in a fixed order to ensure consistency across all participants. Prior to the task, participants received standardized instructions but did not undergo additional training. On average, participants completed the study within 20–30 minutes, with each trial taking approximately 5 minutes.

3.5 Selection of Articles

The selection of articles adhered to two key criteria:

(1) Each article had to be divisible into 10 to 14 evenly distributed text segments to allow for a standardized classification process across participants.
(2) A minimum of 50–65% of the sentences within each article had to exhibit bias, ensuring that participants were sufficiently exposed to biased content.

Articles were sourced from various reputable German news outlets, ensuring representation of different journalistic standards and political perspectives. Each article was analyzed using the BiasScanner.org tool, whose results for individual text segments served as the gold standard for evaluating participant responses.

3.6 Bias Categories

To ensure a structured and consistent assessment, the study focused on 11 distinct bias types that appeared across the selected articles. While no single article contained all 11 bias types, each type was present in at least one of the texts. This approach ensured that participants were exposed to a diverse spectrum of biases, enhancing the study's validity while maintaining comparability across trials.

3.7 Data Analysis

To assess classification performance, Precision, Recall, and F1-scores were computed for each group. These metrics quantify the accuracy of bias detection by comparing participants' classifications to a predefined gold standard, namely the BiasScanner. Precision measures the proportion of correctly identified biased segments among all segments classified as biased by participants, whereas Recall quantifies the proportion of biased segments correctly identified out of all biased segments present in the dataset. The F1-score, as the harmonic mean of Precision and Recall, is used to provide a single, balanced measure of performance that accounts for both false positives and false negatives. It is particularly useful when the class distribution is imbalanced or when a trade-off between Precision and Recall must be considered.

Additionally, a one-way omnibus ANOVA was performed to determine whether statistically significant differences in classification performance existed across the three groups. This statistical test was chosen as it allows for the comparison of mean differences among multiple independent groups while controlling for Type I error inflation. The ANOVA results provided insights into whether technological support influenced bias detection performance and, if so, which groups significantly differed from one another.

4 RESULTS

4.1 General Findings

To evaluate potential learning effects in participants' ability to detect biased content, individual scores for the metrics *precision*,

recall, and *F1-score* were computed. These metrics were chosen to provide a comprehensive view of participants' classification performance, accounting for both the accuracy of bias identification and the balance between sensitivity and specificity.

It is important to acknowledge that, due to time constraints, the final sample consisted of only $N = 20$ participants, falling short of the originally targeted $N = 69$. As a result, statistical power is limited, and findings should be interpreted with caution. Nevertheless, the observed trends offer valuable initial insights and point to potential effects that merit closer examination in future studies with larger and more representative samples.

Participants (N = 20) ranged in age from 23 to 27 (M = 24.4), with a slight majority identifying as male. The sample included individuals with diverse educational and professional backgrounds, such as students, technical professionals, and vocational workers, providing a varied perspective on media bias perception.

Table 1 displays the average differences (Phase 2 minus Phase 1) and corresponding standard deviations for each metric, reported separately by experimental group. These difference scores serve as indicators of performance change following the removal of external assistance, thereby capturing the extent to which participants were able to transfer or retain bias detection competencies.

4.2 Group Results

4.2.1 Group 0 – BiasScanner. The interpretation of results for Group 0 is ambivalent. On the one hand, the large negative differences in all assessed metrics clearly indicate that participants in Phase 1 relied heavily on the software. Particularly striking is the metric *precision*, which reached nearly 100% for almost all participants during Phase 1. This can be attributed to the fact that all correct responses were provided by the BiasScanner, making independent decision-making largely unnecessary. Thus, the high negative difference values in comparison to Phase 2 should not be interpreted as a true decline in performance, but rather as a consequence of the artificially elevated baseline in Phase 1.

The results from Phase 2, in which no support from the BiasScanner was provided, therefore reflect the participants' actual ability to detect bias. Interestingly, the group's average performance in this phase, despite the decline, is not substantially lower than that of the other groups. This may suggest that the use of the tool had at least a sensitizing effect. Potential differences across groups are explored in the next section.

These findings offer partial support for **Hypothesis H1**, which predicted that participants supported by the BiasScanner would achieve higher precision in bias detection compared to the other groups. Although the nearly perfect values during Phase 1 were expected due to direct tool feedback, a more meaningful evaluation lies in Phase 2. After support was removed, Group 0 achieved a mean precision of 0.54—higher than the definition group (approx. 0.45), but lower than the control group (approx. 0.63). This suggests that while the tool may have had a sensitizing effect, it did not lead to a lasting performance advantage. Therefore, H1 is only partially supported.

4.2.2 Group 1 – Definition Group. Participants in Group 1 received supporting definitions during Phase 1 to aid in bias detection. Unlike Group 0, whose responses in Phase 1 were largely predetermined by the software, this form of support required more active cognitive engagement with the material.

The differences between Phases 1 and 2 in this group are overall moderate. The metric *precision* shows an average decrease, while the *F1-score* remains nearly stable. Notably, *recall* displays a positive mean difference of 0.07, which may indicate an improved ability to identify biased content even in the absence of assistance.

These relatively stable results suggest that the definitional support in Phase 1 may have contributed, at least partially, to a sustainable improvement in competence. However, in Phase 2 the group's precision was significantly lower than that of the control group, as shown by the post-hoc Tukey test.

Therefore, **Hypothesis H2**, which proposed that participants receiving definitional support would retain higher precision after support was removed compared to the control group, **is not supported** by the data. Although the group performed reasonably well, the control group ultimately outperformed them in precision without any assistance.

4.2.3 Group 2 – Control Group. Participants in Group 2 received no support in either phase for identifying bias. Their performance thus most closely reflects unassisted, individual bias detection ability.

The differences between phases are minimal across all three metrics. *Precision* increased slightly on average, *recall* decreased slightly, and the *F1-score* remained almost unchanged. These values suggest that participants in this group exhibited relatively stable performance across both phases despite the lack of support.

This observed stability supports **Hypothesis H3**, which assumed that the control group would demonstrate the most consistent performance due to their independent engagement with the material. The slight improvement in precision further reinforces this assumption, indicating that self-reliant strategies may foster durable skills in bias detection.

Group	Precision	Recall	F1-Score
Group 0 (BiasScanner)	-0.43 ± 0.14	-0.28 ± 0.21	-0.36 ± 0.19
Group 1 (Definitions)	-0.09 ± 0.11	0.07 ± 0.22	-0.01 ± 0.17
Group 2 (Control)	0.09 ± 0.19	-0.07 ± 0.20	-0.01 ± 0.13

Table 1: Performance change (Phase 2 – Phase 1) by group (M ± SD)

4.3 Group Comparison

To further investigate potential performance differences between the three groups in Phase 2 (without assistance), a one-way ANOVA was conducted. For the metrics **recall** and **F1-score**, no significant differences were found. However, for **precision**, a statistically significant group difference emerged ($F(2, 17) = 4.51$, $p = .027$).

This finding suggests that the groups differ significantly in their ability to correctly identify biased statements. Notably, Group 0, which received support from the BiasScanner in Phase 1, demonstrated comparatively high precision in Phase 2 (without support).

This could indicate a potential learning effect resulting from the use of the software. However, this interpretation should be made cautiously given the limited sample size.

4.4 Post-hoc Analysis

Following the significant difference in precision scores across the three groups in Phase 2 (without assistance), a Tukey HSD post-hoc

Metric	F-value	p-value
Precision	4.51	.027
Recall	0.47	.63
F1-Score	0.41	.67

Table 2: ANOVA results for group differences in Phase 2 (without assistance)

test was conducted to identify the specific group differences driving this effect.

The analysis revealed a statistically significant difference between Group 1 (Definitions) and Group 2 (No Assistance), with participants in the no-support group achieving higher precision. No significant differences were found between the software-supported group (Group 0) and the other two groups.

This finding suggests that definitional support alone may be less effective in fostering lasting improvements in accurate bias detection. Surprisingly, participants without any assistance outperformed those who received theoretical definitions. This could indicate that independent engagement with the material encourages deeper cognitive processing and the development of more transferable evaluation strategies. The software-supported group showed intermediate results, hinting at a possible sensitization effect, but the lack of significant differences limits firm conclusions.

These results should be interpreted cautiously given the small sample size. Nevertheless, they highlight important tendencies that warrant further investigation.

Comparison	Mean Difference	p-value	Significant
Group 0 vs. Group 1	−0.1075	.210	No
Group 0 vs. Group 2	0.0749	.424	No
Group 1 vs. Group 2	0.1824	**.021**	**Yes**

Table 3: Tukey HSD post-hoc test results for *precision* in Phase 2

5 CONCLUSION

This study examined how different forms of support—automated, definitional, and none—affect individuals' ability to recognize bias in journalistic texts. Despite the limited sample size, several key insights emerged:

- The software-supported group demonstrated high initial precision, but this sharply declined once the tool was removed, suggesting a reliance on the system rather than the development of internalized skills. Nonetheless, their performance in the second phase was still comparatively strong, possibly indicating increased awareness or sensitization to bias.
- The definitional group showed stable performance across both phases, with a slight increase in recall, pointing to a modest, but potentially more sustainable improvement in bias detection capabilities.
- The control group, which received no support, maintained a consistent performance throughout the study. Interestingly, this group achieved the highest precision in the second phase, outperforming the definitional group. This may imply that self-directed engagement fosters stronger, more transferable skills.

Overall, the findings suggest that while technological tools can support bias detection, they may unintentionally hinder the development of independent skills if users become overly reliant on them. In contrast, minimal support may encourage deeper cognitive engagement. These insights underline the importance of balancing assistance and autonomy in media literacy interventions.

Future research should build on these findings with larger, more diverse samples and explore long-term effects of support types on critical media skills. It may also be valuable to investigate hybrid approaches that combine technological tools with reflective training to foster both immediate support and sustainable learning.

REFERENCES

[1] Eric Arias. 2019. How Does Media Influence Social Norms? Experimental Evidence on the Role of Common Knowledge. *Political Science Research and Methods* 7, 3 (2019), 561–578. https://doi.org/10.1017/psrm.2018.1

[2] Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication* 43, 4 (1993), 51–58.

[3] M FactCheck. 2020. MediaBiasFactCheck. http://mediabiasfactcheck.com/methodology/

[4] Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: an interdisciplinary literature review. *International Journal on Digital Libraries* 20 (2019), 391–415. https://doi.org/10.1007/s00799-018-0261-y

[5] Wei-Hao Lin, Theresa Wilson, Janyce Wiebe, and Alexander G Hauptmann. 2006. Which side are you on? Identifying perspectives at the document and sentence levels. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*. 109–116.

[6] Tim Menzner and Jochen L. Leidner. 2024. BiasScanner: Automatic Detection and Classification of News Bias to Strengthen Democracy. arXiv:2407.10829 [cs.CL] https://arxiv.org/abs/2407.10829

[7] N Newman, R Fletcher, K Eddy, CT Robertson, and RK Nielsen. 2023. *Digital news report 2023*. Technical Report.

[8] Michael S. Pollard, Sean Grant, and Jessica Saunders. 2020. Profiles of News Consumption: Platform Choices, Perceptions of Reliability, and Partisanship. https://www.rand.org/pubs/research_reports/RR4212.html Zugriff auf die PDF-Datei über die offizielle Webseite der RAND Corporation.

[9] Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications* 237 (2024), 121641.

[10] Dietram A. Scheufele and David Tewksbury. 2007. Framing, Agenda Setting, and Priming: The Evolution of Three Media Effects Models. *Journal of Communication* 57, 1 (2007), 9–20. https://doi.org/10.1111/j.0021-9916.2007.00326.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0021-9916.2007.00326.x

[11] Gülistan H Serapşah and Mehtap Sevgihan. 2023. The Influence of Media on Political Communication: A Review of Literature. *American Journal of Law and Political Science* 2, 3 (Nov. 2023), 42–54. https://gprjournals.org/journals/index.php/AJLPS/article/view/214

[12] Timo Spinde, Christina Kreuter, Wolfgang Gaissmaier, Felix Hamborg, Bela Gipp, and Helge Giese. 2021. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. 61–69. https://doi.org/10.1109/JCDL52503.2021.00018

[13] University of Michigan. 2014. *News Bias Explored*. https://websites.umich.edu/~newsbias/index.html Accessed: 2024-12-05.

[14] Qi Wang and Hee Jin Jeon. 2020. Bias in bias recognition: People view others but not themselves as biased by preexisting beliefs and social stigmas. *PLOS ONE* 15, 10 (10 2020), 1–18. https://doi.org/10.1371/journal.pone.0240232

[15] David Manning White. 1950. The "Gate Keeper": A Case Study in the Selection of News. *Journalism Quarterly* 27, 4 (1950), 383–390. https://doi.org/10.1177/107769905002700403