NewsUnfold: Creating a News-Reading Application That Indicates Linguistic Media Bias and Collects Feedback

Smi Hinterreiter[1], Martin Paul Wessel[2], Fabian Schliski[3], Isao Echizen[4], Marc Erich Latoschik[1], Timo Spinde[5]

[1]University of Würzburg, Germany
[2]Technical University of Munich, Germany
[3]University of Passau, Germany
[4]National Institute of Informatics, Japan
[5]University of Göttingen, Germany
smi.hinterreiter@uni-wuerzburg.de, m.wessel@media-bias-research.org, schliski@fim.uni-passau.de, iechizen@nii.ac.jp, marc.latoschik@uni-wuerzburg.de, t.spinde@media-bias-research.org

Abstract

Media bias is a multifaceted problem, leading to one-sided views and impacting decision-making. A way to address digital media bias is to detect and indicate it automatically through machine-learning methods. However, such detection is limited due to the difficulty of obtaining reliable training data. Human-in-the-loop-based feedback mechanisms have proven an effective way to facilitate the data-gathering process. Therefore, we introduce and test feedback mechanisms for the media bias domain, which we then implement on NewsUnfold, a news-reading web application to collect reader feedback on machine-generated bias highlights within online news articles. Our approach augments dataset quality by significantly increasing inter-annotator agreement by 26.31% and improving classifier performance by 2.49%. As the first human-in-the-loop application for media bias, the feedback mechanism shows that a user-centric approach to media bias data collection can return reliable data while being scalable and evaluated as easy to use. NewsUnfold demonstrates that feedback mechanisms are a promising strategy to reduce data collection expenses and continuously update datasets to changes in context.

1 Introduction

Media bias, slanted or one-sided media content, impacts public opinion and decision-making processes, especially on web platforms and social media (Ardèvol-Abreu and Zúñiga 2017; Eberl, Boomgaarden, and Wagner 2017; Spinde et al. 2023). News consumers are frequently unaware of the extent and influence of bias (Kause, Townsend, and Gaissmaier 2019; Spinde et al. 2020; Ribeiro et al. 2018), leading to limited awareness of specific issues and narrow, one-sided points of view (Ardèvol-Abreu and Zúñiga 2017; Eberl, Boomgaarden, and Wagner 2017). As promoting media bias awareness has beneficial effects (Park et al. 2009; Spinde et al. 2022), emphasis on the need for methods that automatically detect media bias is growing (Wessel et al. 2023). Such methods potentially impact user behavior, as they facilitate the development of systems that analyze various subtypes of bias comprehensively and in real-time (Spinde et al. 2021a).

Several approaches have been developed for automated media bias classification (Wessel et al. 2023; Spinde et al. 2024; Liu et al. 2021; Hube and Fetahu 2019; Vraga and Tully 2015). However, they share a challenge: While datasets are vital for training machine-learning models, the intricate and subjective nature of media bias makes the manual creation of these datasets time-consuming and expensive (Spinde et al. 2021b). Crowdsourcing is cost-effective but can yield unreliable annotations with low annotator agreement (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013). In contrast, expert raters ensure consistency but lead to substantial costs (Spinde et al. 2021b),[1] making scaling data collection challenging (Spinde et al. 2021b). Consequently, the media bias domain lacks reliable datasets for effective training of automatic detection systems (Wessel et al. 2023). Successful Human-in-the-loop (HITL) approaches addressing similar challenges (Mosqueira-Rey et al. 2022; Karmakharm, Aletras, and Bontcheva 2019) remain untested for media bias, particularly visual methods (Karmakharm, Aletras, and Bontcheva 2019).

We propose a HITL feedback mechanism showcased on NewsUnfold, a news-reading platform that visually indicates linguistic bias to readers and collects user input to improve dataset quality. NewsUnfold is the first approach employing feedback collection to gather a media bias dataset. In the first of three phases (Figure 1), since visual HITL (Human-in-the-Loop) methods for media bias annotation have not previously been tested, we conducted a study comparing two feedback mechanisms (Section 3). Second, we implement a feedback mechanism on NewsUnfold (Section 4). Third, we use NewsUnfold with 12 articles to curate the NewsUnfold Dataset (NUDA), comprising approximately 2000 annotations (Section 4). Notably, the collected feedback annotations exhibit a 90.97% agreement with expert annotations and a 26.31% higher inter-annotator agreement (IAA) than the baseline, the expert-annotated BABE dataset (Spinde et al. 2021b).[2] This increase is also visible

[1]For example, in the expert-based BABE dataset, one sentence label costs four to six euros, varying with rater count.

[2]The IAA evaluates how consistently different individuals assess or classify the same dataset (Hayes and Krippendorff 2007).

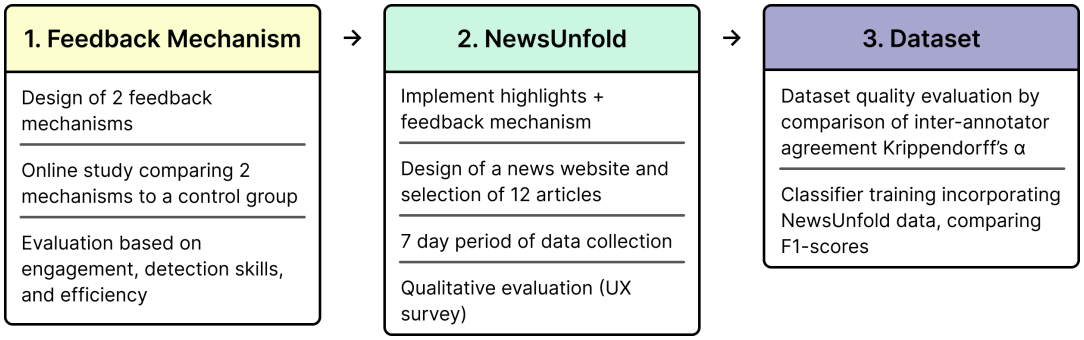

Figure 1: Three-step process of the NewsUnfold Development and Evaluation.

when the dataset is used in classifier training, resulting in an F1-score of .824, an increase of 2.49% compared to the baseline BABE performance. While the platform's design is adaptable to diverse subtypes of bias, we facilitate our evaluation by focusing on linguistic bias. Linguistic bias is defined by Spinde et al. (2024) as a bias by word choice to transmit a perspective that manifests prejudice or favoritism towards a specific group or idea (Spinde et al. 2024). Despite being neither objective nor binary, collecting binary labels is a promising solution regarding the challenges arising from its ambiguous and complex nature (Spinde et al. 2021b). A UX study involving 13 participants highlights high ease of use and enthusiasm for the concept. Participants also reported a strong perceived impact on critical reading and expressed positive sentiment toward the highlights.

In this work, we:

1. Explore feedback mechanisms for the first time in the context of automated media bias detection methods.

2. Introduce and evaluate NewsUnfold, a news-reading platform highlighting bias in news articles, making media bias detection models accessible for everyday news consumers. NewsUnfold collects feedback on bias highlights to improve its automatic detection.[3]

3. Generate the NewsUnfold Dataset (NUDA) incorporating approximately 2,000 annotations.

4. Present classifiers trained using NUDA and benchmarked against existing methodologies, enhancing performance when combined with other datasets.

This paper proposes a design for a cost-effective HITL system to improve and scale media bias datasets. Such feedback mechanisms can be integrated into various media platforms to highlight media bias and related concepts. Further, the system can adapt to changes in language and context, facilitating applied endeavors to run models on news sites and social media to understand and mitigate media bias and increase readers' awareness.

[3]Data and code are publicly available at https://github.com/media-bias-group/newsunfold

2 Related Work

Media Bias

Various studies (Lee et al. 2022; Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013; Raza, Reji, and Ding 2022; Hube and Fetahu 2019; Ardèvol-Abreu and Zúñiga 2017; Eberl, Boomgaarden, and Wagner 2017) highlight the complex nature of media bias, or, more specifically, linguistic bias (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013; Wessel et al. 2023; Spinde et al. 2024). Individual backgrounds, such as demographics, news consumption habits, and political ideology, influence the perception of media bias (Druckman and Parkin 2005; Eveland Jr. and Shah 2003; Ardèvol-Abreu and Zúñiga 2017; Kause, Townsend, and Gaissmaier 2019). Content resonating with a reader's beliefs is often viewed as neutral, while dissenting content is perceived as biased (Kause, Townsend, and Gaissmaier 2019; Feldman 2011). Enhancing awareness of media bias can improve the ability to detect bias at various levels — word-level, sentence-level, article-level, or outlet-level (Spinde et al. 2022; Baumer et al. 2015).

While misinformation is closely connected to media bias and has received much research attention, most news articles do not fall into strict categories of veracity (Weil and Wolfe 2022). Instead, they frequently exhibit varying degrees of bias, underlining the importance of media bias research.

Automatic Media Bias Detection

NLP methods can automate bias detection, enabling large-scale bias analysis and mitigation systems (Wessel et al. 2023; Spinde et al. 2021b; Liu et al. 2021; Lee et al. 2022; Pryzant et al. 2020; He, Majumder, and McAuley 2021). Yet, current bias models' reliability for end-consumer applications is limited (Spinde et al. 2021b) due to their dependency on the training dataset's quality. These models often rely on small, handcrafted, and domain-specific datasets, frequently using crowdsourcing (Wessel et al. 2023), which cost-effectively delegates annotation to a diverse, non-expert community (Xintong et al. 2014). The subjective nature of bias and potential inaccuracies from non-experts can result in lower agreement, more noise (Spinde et al. 2021c), and the perpetuation of harmful stereotypes (Otterbacher 2015). Conversely, expert-curated datasets offer higher agreement but come at a greater cost (Spinde et al. 2024).

Datasets used for automated media bias detection need to stay updated (Wessel et al. 2023), annotations should be collected across demographics (Pryzant et al. 2020), and media bias awareness reduces misclassification (Spinde et al. 2021b). The limited range of topics and periods covered by current datasets and the complexities involved in annotating bias decrease the accuracy of media bias detection tools. This, in turn, impedes their widespread adoption and accessibility for everyday users (Spinde et al. 2024). To make the data collection process less resource-intensive and optimize gathering human feedback, we raise media bias awareness by algorithmically highlighting bias and gathering feedback from readers.

Media Bias Awareness

News-reading websites like AllSides[4] or GroundNews[5] offer approaches for media bias awareness at article and topic levels (Spinde et al. 2022; An et al. 2021; Park et al. 2009). However, research on these approaches is sparse. One approach uses ideological classifications (An et al. 2021; Park et al. 2009; Yaqub et al. 2020) to show contrasting views at the article level. At the text level, studies use visual bias indicators like bias highlights (Spinde et al. 2020, 2022; Baumer et al. 2015) with learning effects persisting post-highlight removal (Spinde et al. 2022). As the creation of media bias datasets does not include media bias awareness research, NewsUnfold connects these research areas.

HITL Platforms for Crowdsourcing Annotations

HITL learning improves machine learning algorithms through user feedback, refining existing classifiers instead of creating new labels (Mosqueira-Rey et al. 2022; Sheng and Zhang 2019). Enhanced classifier precision can be achieved by combining crowdsourcing and HITL approaches, leveraging user feedback to generate labels via repeated-labeling, and increasing the number of annotations (Xintong et al. 2014; Karmakharm, Aletras, and Bontcheva 2019; Sheng and Zhang 2019; Stumpf et al. 2007). For instance, "Journalists-In-The-Loop" (Karmakharm, Aletras, and Bontcheva 2019) continuously refines rumor detection by soliciting visual veracity ratings from journalist's feedback. Similarly, Mavridis et al. (2018) suggest a HITL system to detect media bias in videos. They plan to extract bias cues through comparative analysis and sentiment analysis and rely on scholars to validate the output. However, their system stays in the conceptual phase. Brew, Greene, and Cunningham (2010)'s web platform crowdsources news article sentiments and re-trains classifiers based on non-expert majority votes, emphasizing the effectiveness of diversified annotations and user demographics over mere annotator consensus. Demartini, Mizzaro, and Spina (2020) propose combining automatic methods, crowdsourced workers, and experts to balance cost, quality, volume, and speed. Their concept uses automated methods to identify and classify misinformation, passing some to the crowd and experts for verification in unclear cases. Similar to Mavridis et al. (2018), they do not implement their system and describe no UI details.

As no HITL system has been implemented to address media bias, we aim to close this gap by integrating automatic bias highlights based on expert annotation data readers can review. To mitigate possible anchoring bias and uncritical acceptance of machine judgments, we test a second feedback mechanism aimed at increasing critical thinking (Vaccaro and Waldo 2019; Furnham and Boo 2011; Jakesch et al. 2023; Shaw, Horton, and Chen 2011).

3 Feedback Mechanisms

As the evaluation of feedback mechanisms for media bias remains unexplored, in a preliminary study, we design and assess two HITL feedback mechanisms for their suitability for data collection. Using sentences from news articles labeled by the classifier from Spinde, Hamborg, and Gipp (2020), we compare the mechanisms *Highlights*, *Comparison*, and a control group without visual highlights. Our analysis focuses on (1) dataset quality, assessed using Krippendorff's α; (2) engagement, quantified by feedback given on each sentence[6]; (3) agreement with expert annotations, evaluated through F1 scores; and (4) feedback efficiency, measured by the time required in combination with engagement and agreement.

In the *Highlights* mechanism, biased sentences are colored yellow, and non-biased ones are grey, inspired by Spinde et al. (2022). Participants indicate their agreement or disagreement with these classifications through a floating module (Figure 2). The *Comparison* mechanism displays sentence pairs. For the first sentence, participants provide feedback on the AI's classification as in *Highlights*. The second sentence has no color coding, prompting users with "What do you think?" (Figure 3), thereby aiming to foster an independent bias assessment and mitigate anchoring effects. Participants in the control group do not see any highlights, solely encountering the feedback module with the second question from *Comparison*.

We use the BABE classifier trained by Spinde et al. (2021b) to generate the sentence labels and highlights. Currently, the classifier showcases the highest performance by fine-tuning the large language model RoBERTa with an extensive dataset on linguistic bias annotated by experts on both sentence and word levels. The BABE-based model on Huggingface[7] generates the probability of a sentence being biased or not biased for each article. We accordingly assign the label with the higher probability.

Study Design

To assess the two mechanisms, we recruit 240 participants, balanced regarding gender, from Prolific.[8] On the study website built for this purpose, depicted in Figure 13, they view two articles from different political orientations paired with one feedback mechanism per group. During the study,

[4]https://www.allsides.com/

[5]https://ground.news/

[6]Readers can edit annotations anytime, but each unique sentence counts as one interaction in our feedback metric.

[7]https://huggingface.co/mediabiasgroup/da-roberta-babe-ft

[8]https://www.prolific.co

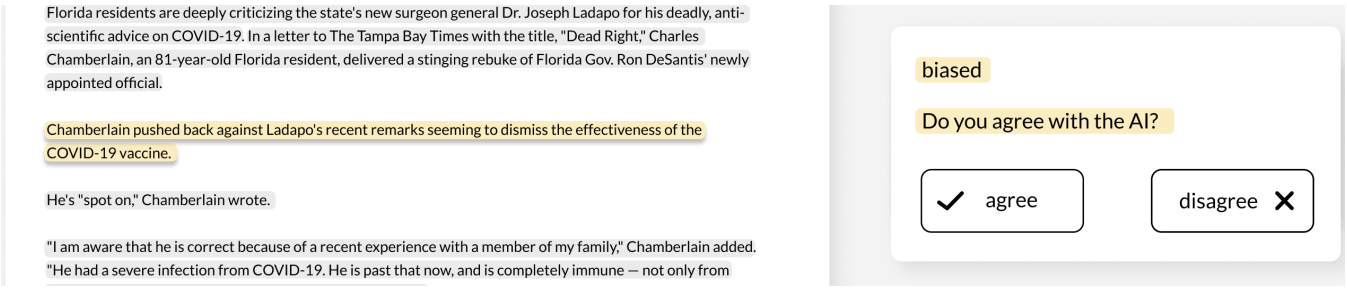

Figure 2: The feedback mechanism *Highlights* uses the BABE classifier to highlight biased sentences in yellow and not biased sentences in grey. Readers can agree or disagree with this classification through the feedback module on the right.

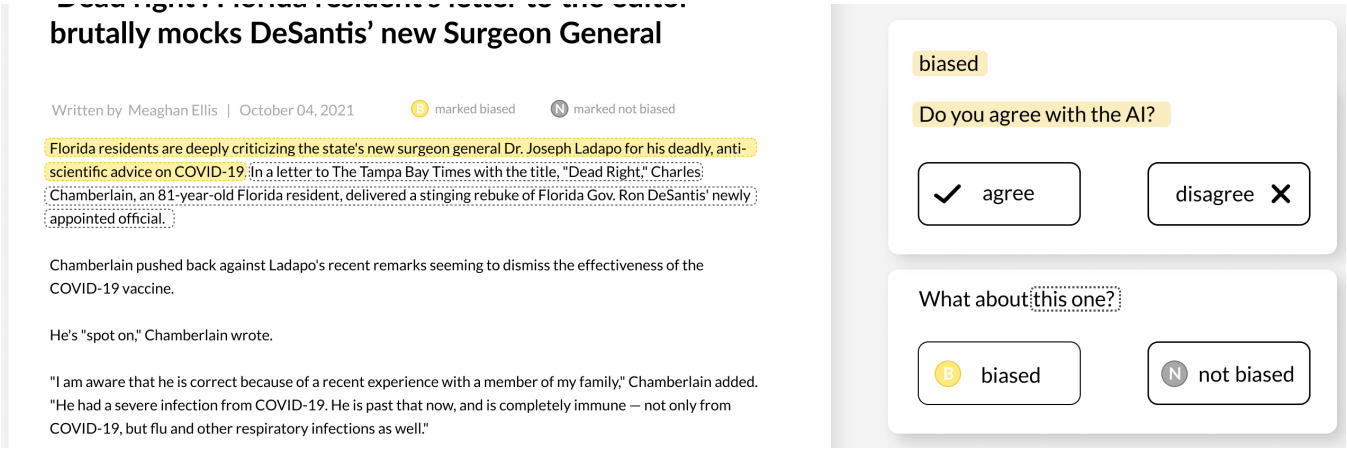

Figure 3: The feedback mechanism *Comparison* operates on sentence pairs and uses the BABE classifier to highlight the first sentence as biased in yellow. Readers can agree or disagree with this classification through the feedback module on the right. The next sentence is merely outlined. Here, the feedback module asks for a bias rating without the classifier anchor.

users freely determine their annotation count and time spent, with a progress bar showing the number of annotated sentences. Not interacting with any sentences prompts a pop-up, but they can click 'next' to proceed.

Curated from AllSides, articles match the baseline dataset's topics (Spinde et al. 2021b) and were annotated by four experts.[9] Table 5 compares the classifier and expert annotations. To measure IAA, we use Krippendorff's α, an evaluation metric often used in the media bias domain that assesses dataset quality by determining annotator agreement beyond chance (Hayes and Krippendorff 2007). As higher engagement yields more data, we measure engagement through the number of decisions made with the feedback mechanism. An efficient feedback mechanism reduces the task's tedium while ensuring data quality. Efficiency is calculated with the Bonferroni correction: $\frac{Engagement}{Time} * F1$.

We guarantee GDPR conformity through a preliminary data processing agreement. A demographic survey and an introduction to media bias follow (Appendix A). A post-introduction attention test confirms participants' understanding of media bias, which, if failed twice, results in study exclusion. Then, participants read through a description of the study task and proceed to give feedback on the two articles. Lastly, a concluding trustworthiness question ensures data reliability. If participants clicked through the study inattentively, they could indicate that their data is not usable for research (Draws et al. 2021) while still receiving full pay (Spinde et al. 2022).

[9]Experts have at least six months experience in media bias. Consensus was achieved through majority or discussion.

Results

The 240 participants in the study spent an average of 11:24 minutes, with a compensation rate of £7.89/hr. Twelve participants failed the attention test once, but only one was excluded for a second failure. We further excluded 33 participants who flagged their data as unsuitable for research. Therefore, the analysis includes data from 206 participants: 69 control group participants, 66 *Comparison* group participants, and 71 *Highlights* group participants ($p = .84, f = .23, \alpha = .05$). 104 participants identified as female, 99 as male, and 3 as other, with an average age of 36.62 years ($SD = 13.74$). The sample, on average, exhibits a left slant (Figure 11 and Figure 12) with higher education (Figure 7). 196 participants indicated advanced English levels, 9 intermediate, and 1 beginner (Figure 9). News reading frequency averaged around once a day (Figure 10).

Notably, we observe a high overall engagement, with even

Group	Feedback[a]	Engagement	IAA	F1-Score	Efficiency[b]
Highlights	5564	$.9329 \pm .1642$	.229	$.5720 \pm .1266$	$.1252 \pm .0951$
Comparison	4484	$.8088 \pm .3266$	.22	$.5736 \pm .1339$	$.0813 \pm .0421$
Control	5037	$.8690 \pm .2853$	.2	$.5769 \pm .1566$	$.1116 \pm .0678$

[a] Number of feedback-related interactions
[b] Calculated based on the Bonferroni correction

Table 1: Overview of Feedback Interactions per Group.

the least annotated sentences receiving feedback from 70% of the participants. We detail the results of the feedback mechanism study, including engagement, IAA, F1 scores, and efficiency, in Table 1. The *Highlights* group exhibits higher engagement than the *Comparison* group, containing more collected data. Also, *Highlights* demonstrates higher efficiency by collecting more feedback data in less time without compromising quality measured by IAA and agreement with the expert standard.

The increases in engagement and efficiency are significant at a .05 significance level. Due to variance inhomogeneity indicated by a significant Levene test ($p < .05$), we applied Welch's ANOVA for unequal variances. Post-hoc Holm-Bonferroni adjustments revealed significant differences between the CONTROL and HIGHLIGHTS groups, with $p < .0167$ for efficiency and $p < .025$ for engagement. The Games-Howell post-hoc test confirmed these results.As in previous research, IAA and F1 scores from crowdsourcers are low due to the complex and subjective task (Spinde et al. 2021c). F1 score differences are not significant (ANOVA with Holm-Bonferroni, $p > .05$). Given the comparable IAA and F1 scores across groups, we integrate *Highlights* within NewsUnfold to optimize data collection efficiency.

4 The NewsUnfold Platform

Tailored toward news readers, NewsUnfold highlights potentially biased sentences in articles (1 in Figure 4) and incorporates the *Highlights* feedback module (2 in Figure 4) assessed in Section 3 to create a comprehensive, cost-effective, crowdsourced dataset through reader-feedback. The feedback mechanism further includes a free-text field (3 in Figure 4) where readers can justify their feedback.

Application Design

NewsUnfold's responsive design draws inspiration from news aggregators[10] to represent an environment where users, given updating content, return to regularly. By clarifying the purpose of our research, the societal importance of media bias, and giving access to automated bias classification, we encourage voluntary feedback contributions.

The landing page states NewsUnfold's mission: encouraging bias-aware reading and collecting feedback to refine bias detection to mitigate its negative effects. To further motivate contributions, it emphasizes the value of individual users' feedback in enhancing bias-detection capabilities. Clicking a call-to-action button guides users to the *Article Overview Page* (Figure 14). As a preliminary stage, this page displays 12 static articles spanning nine subjects, balanced by the bias amount and political orientation. Different articles enable readers to compare the amount of bias in one article. Selecting an article directs users to NewsUnfold's *Article Reading Page*, which integrates the bias highlights and feedback mechanism. Table 2 outlines its essential components. The sparkles highlight controversial sentences or sentences that received the least feedback to enable balanced feedback collection (4 in Figure 4). From the *Article Overview Page* (Figure 14), users can additionally initiate a tutorial (7 in Figure 14) guiding them through the bias highlights (1 in Figure 4), the feedback mechanism (2 in Figure 4), and concluding with a pointer to the UX survey (5 in Figure 16). After each article, we show three recommended articles (6 in Figure 16).

Study Design

Our primary objectives for testing NewsUnfold in a real-world setting are:

1. **Engagement:** Measure the amount of voluntary feedback from readers without monetary incentives.
2. **Data Quality:** Assessing quality of feedback.
3. **Classifier:** Investigating classifier performance when integrating feedback-generated labels.
4. **User Experience:** Evaluating user experience and perception of NewsUnfold, focusing on bias highlights (1 in Figure 4) and feedback (2 in Figure 4) for a user-centered design approach.

During the study, readers can freely explore the platform, select articles, decide to provide anonymous feedback, and choose to participate in the UX survey. Unlike the preliminary study, participants are not sourced from crowdworking platforms but reached via LinkedIn, Instagram, and university boards. The outreach briefly introduces NewsUnfold with a link to its landing page. Readers are informed of feedback data collection beforehand.

To understand the readers' experiences, a voluntary UX survey (5 in Figure 16) is available after reading an article.[11] In this study, we prioritize identifying UX issues among readers to boost participation and feedback efficiency, focusing on UX-oriented data collection over comprehensive quantitative analysis. To obtain user analytics,

[10] E.g., Google News (https://news.google.com).

[11] The survey consists of 9 questions: two scales and eight optional open-ended queries. Appendix B contains a detailed breakdown of the survey and its results.

Element	Description	Reference
1 Article text with bias highlights	Presents the text with bias indicators, enhancing media bias awareness. Includes headline, author, outlet, and metadata.	1 in Section 3, Figure 4, Figure 15
2 Feedback mechanism *Highlights*	Integrated *Highlights* mechanism, prompting users to consider the sentence and feedback on the classification.	2 in Section 3, Figure 4, Figure 15
3 Free-text field for reasoning	Field where readers explain their bias assessments for more thorough feedback.	3 in Section 4, Figure 4, Figure 15
4 *Sparkles* indicators	Emphasizing unannotated or controversial sentences, prompting further feedback.	4 in Section 4, Figure 4, Figure 15
5 UX survey button	Prompt for a UX survey, collecting feedback on app usability and satisfaction.	5 in Section 4, Figure 16
6 Recommended articles	Displays three suggested articles, prompting continued reading based on user behavior and article annotations.	6 in Section 4, Figure 16
7 Tutorial	Gives readers a tour of NewsUnfold, explains media bias, and the feedback mechanism	7 in Section 4, Figure 14

Table 2: Key Elements of NewsUnfold. Yellow numbers appear in NewsUnfold figures.

we use *Umami*[12], a privacy-centric tool logging the number of clicks, unique visitors, country, language settings, device types, most-visited pages, and the number of tutorial initiations while keeping the anonymity of visitors.

Dataset Creation and Evaluation

NewsUnfold collects anonymous feedback on bias highlights on 12 articles with 357 sentences at the sentence level. The data is stored on university servers. Articles cover topics consistent with the baseline dataset (e.g., gender equality, black lives matter, and climate change (Spinde et al. 2021b)), represent different political slants, and are balanced regarding bias strengths. NewsUnfold uses a repeated-labeling method (Sheng and Zhang 2019), employing a majority-vote system with a minimum of five votes per sentence to establish sentence labels. The labels are stored in the same structure as BABE (Spinde et al. 2021b) to enable the merging of the two datasets. We apply a spam detection method by Raykar and Yu (2011) to filter out unreliable annotations. We calculate a score between 0 and 1 for each annotator and eliminate annotators in the 0.05th percentile. We assess the quality of the resulting dataset, similar to Section 3, using the IAA metric Krippendorff's α and manual analysis.

As HITL systems center around iteratively improving machine performance through user input, we evaluate the integration of feedback data into classifier training. The training process adopts hyperparameter configurations from Spinde et al. (2021b) with a pre-trained model from Hugging Face.[13] We train and evaluate the model with data from NUDA added to the 3700 BABE sentences and compare it against the baseline classifier (Spinde et al. 2021b) using the F1-Score (Powers 2008).

[12]https://umami.is

[13]https://huggingface.co/mediabiasgroup/DA-RoBERTa-BABE

5 Results

From March 4th to March 11th (2023), NewsUnfold had 187 unique visitors. 158 read articles, 33 (20.89%) provided sentence feedback, and eight offered 25 additional reasons for feedback, mainly on sentences perceived as biased (84%) but highlighted as not biased (80%). 45 (28.48%) completed the tutorial, and 13 (6.9%) the UX survey. Geographically, 61% were from Germany, 25% from Japan, 6% from the United States. Language-wise, 45% preferred English, 42% preferred German. Notably, 52% accessed via mobile, highlighting mobile optimization's importance.

The 357 sentences collectively received 1997 individual annotations, representing either agreement or disagreement with the presented classifier outcome. We identify two spammers within the 5% spammer score range and remove 47 annotations, leaving 1950 valid annotations in the dataset. 316 sentences attain a label through the repeated-labeling method. A sentence is categorized as **decided** if there is a majority, **controversial** if the biased-to-unbiased feedback ratio lies between 40-60%[14], and **undecided** if the ratio stands at an exact 50% as listed in Figure 6. 310 **decided** sentences spanning nine topics form NUDA.[15]

Data Quality

To evaluate if NewsUnfold increases data quality, we calculate the Inter-Annotator agreement score Krippendrorff's α. The NUDA dataset achieves a Krippendorff's α of .504. The 26.31% increase in IAA compared to the baseline's IAA of .399 (Spinde et al. 2021b) is statistically significant, as demonstrated in Figure 5 by the non-overlapping bootstrapped confidence intervals. To demonstrate that the IAA does not merely increase with the sample size but through

[14]A sentence can be **decided** and **controversial** at the same time.

[15]Data and statistics: https://doi.org/10.5281/zenodo.8344891

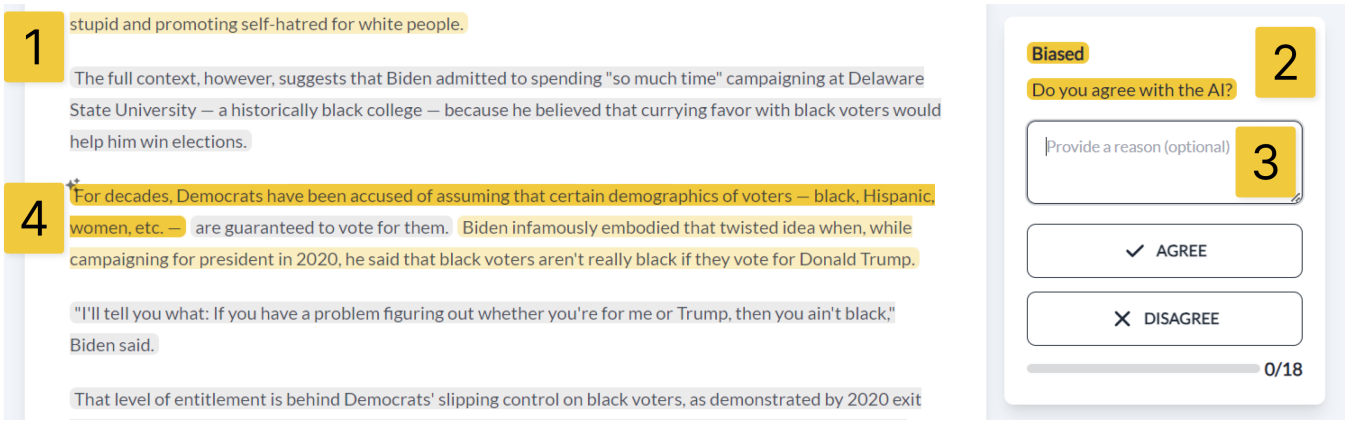

Figure 4: The classifier shows the highlights in yellow (biased) and grey (not biased) on NewsUnfold. The feedback module on the right allows readers to agree or disagree and leave optional feedback. The *Sparkles* draw attention to controversial sentences or sentences that need more feedback. Table 2 explains the elements with yellow numbers.

higher data quality, we take 100 randomly sized dataset samples ($n = 10$ to $n = 1950$), calculate the IAA for each, and employ a regression model.

The model's explanatory power ($R^2 = .009$, $R^2{}_{adjusted} = -.002$) suggests a negligible linear relationship between sample size and the F1 score Table 4. This implies that the model does not explain the variance in F1 scores when accounting for the increase in data points. Moreover, the F-statistic of .8424 ($p = 0.361$) does not provide evidence to reject the null hypothesis that there is no linear relationship between sample size and F1 score ($x_1 = .000004$, $SD = .000004$, $t = -.918$, $CI[.00001, .000004]$). Therefore, we conclude that the collected data is reliable, and increases in quantity do not necessarily translate into increased data quality. Further, we conducted a manual evaluation by annotating 310 sentences and comparing these expert annotations against the labels provided by NUDA. The comparison yielded an agreement of 90.97% across 282 labels, with a disagreement of 9.03% over 28 labels. Specifically, the experts identified 25 sentences as biased, which NUDA had not, whereas only three sentences deemed biased by the experts were classified as unbiased by NUDA. A closer examination of the disagreeing labels revealed that the primary source of discrepancy was sentences containing direct quotes. When we removed 69 sentences predominantly consisting of direct quotes, the agreement increased to 95.44% on 230 labels, with the disagreement rate dropping to 4.56% on 11 labels. Of these, ten sentences experts labeled as biased were not labeled as biased by NUDA, and one sentences experts labeled as biased was labeled not biased by NUDA. This high agreement rate suggests that NewsUnfold can gather high-quality annotations and labels.

Classifier Performance

After merging NUDA with the BABE dataset, the average F1 score (5-fold cross-validation) is .824 (Table 3), showing a 2.49% improvement over the BABE baseline (Spinde et al. 2021b). While this may not constitute a substantial improvement, it is a positive increment towards the anticipated direction. We conduct five 5-fold cross-validations with different distributions to control for potential biases in the F1-Score due to imbalanced dataset distribution. Folds and repetitions show only marginal differences with a variance of .000022, suggesting that the data quality provides reliable results.

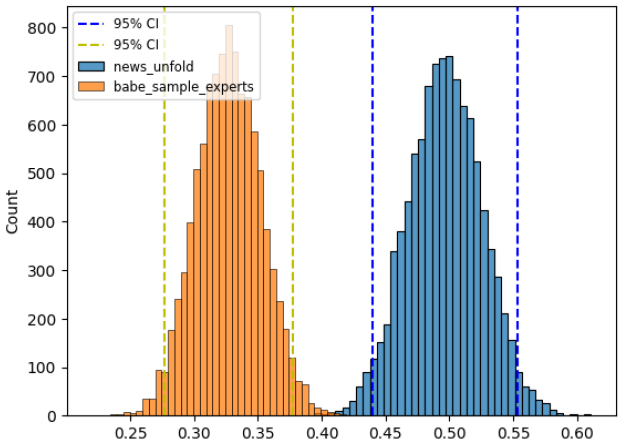

Figure 5: Comparison of the expert-generated dataset with the NUDA dataset. The non-overlapping confidence intervals indicate a significant increase.

Dataset	Sentences	F1-Score (%)
BABE	3700	.804 ± .014
NUDA and BABE	4010	.824 ± .017

Table 3: Comparison of classifiers trained on BABE alone versus BABE combined with NewsUnfold Dataset.

User Experience Survey Results

Thirteen participants took part in the UX survey. They express positive feelings about the platform and bias highlights

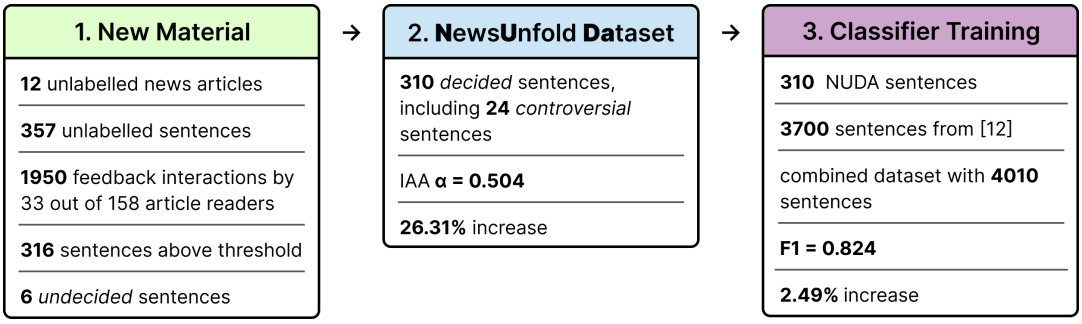

Figure 6: The NewsUnfold dataset creation process. First, readers incorporate and annotate new material. Second, all decided sentences are collected into NUDA, and their IAA is calculated. Third, NUDA is added to BABE for classifier training.

(Appendix B). The platform's ease of use receives a high rating of 8.46 on a 10-point scale, indicating a user-friendly design, affirmed by participants' descriptions of the interface as intuitive and concise. While almost all users state a positive effect on reading more critically, some raise concerns about highlight calibration, their ineffectiveness with unbiased articles, and bias introduced by direct quotes.

Participants exhibit varied opinions when providing feedback, most enjoying it, some undecided, and one finding it work-like (Appendix B). For those interested in giving feedback, the survey indicates an easy process.

One participant mentioned that skipping the tutorial leads to confusion. Thus, one could consider making the tutorial mandatory in future iterations. In conclusion, we expect that the ease of use facilitates higher retention rates and engagement while the self-reported heightened media bias awareness positively correlates with data quality.

6 Discussion

Feedback Mechanisms Study

Although feedback was optional, monetary incentives and a structured study setting prompted participants to share opinions on all highlights, raising questions about engagement in settings without such incentives. Initially, we assumed the *Comparison* method reduced anchoring bias and increased critical thinking. However, F1 scores between *Highlights* and *Comparison* disprove this. Both F1 and IAA scores were expectantly low as media bias perception is highly subjective, and comparable approaches report similar scores (Spinde et al. 2021c; Hube and Fetahu 2019; Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013). *Comparison* was less efficient than *Control*, possibly due to managing two questions simultaneously. Interestingly, the *Highlights* method led to longer engagement times, indicating a mix of focused attention and prolonged article interaction, possibly enhancing contextual critical thinking.

Further, Table 5 highlights issues with the training data for BABE,[16] treated as the ground truth. Discrepancies between expert labels suggest that BABE may not be entirely accurate, especially since the dataset often misclassifies subtly biased sentences as "not biased". However, achieving complete accuracy in bias classification may be unattainable due to the subjective nature of bias and the misleading concept of a single, absolute ground truth (Xu and Diab 2024).

NewsUnfold

NewsUnfold showcases how the feedback mechanism gathers bias annotations in a news-reading environment. The system increases IAA by 26.31%, achieves high agreement with expert labels, and corrects misclassifications through feedback. For example, this sentence was initially deemed non-biased but corrected to biased:

> *"That level of entitlement is behind Democrats' slipping control on black voters, as demonstrated by 2020 exit polls showing that, for example, just 79% of black men voted for Biden, a percentage that has been dropping since 2012."*

Despite having a lower label count than BABE, the feedback dataset demonstrates greater agreement with expert labels. Furthermore, statistical analysis indicates that the improvement in IAA cannot be attributed solely to the annotation count Section 5. Although the increase in the dataset size likely drives the rise in F1-Score, the data has been shown to be reliable. This suggests that readers using the feedback mechanism (2 in Figure 4) offer a reliable alternative to costly expert annotators, facilitating the collection of more extensive data sets. The scarcity of high-quality media bias datasets highlights the need to integrate feedback mechanisms on NewsUnfold or other digital and social platforms. Likewise, other classifiers, such as misinformation classifiers, can use similar mechanisms to gather data and improve accuracy while augmenting the cognitive abilities of readers (Pennycook and Rand 2022; Spinde et al. 2022).

Limited written feedback through the feedback mechanism (3 in Figure 4) might be due to typing disruptions during reading. Most feedback highlighted false negatives, indicating it is simpler to spot bias than explain its absence. Currently, nuances are not well-captured by the binary feedback in the first iteration, as all design decisions are a trade-off between an effortless process that drives engagement and more complex labeling. While more friction can foster deeper thinking, we decided on a simple, binary feedback version (1 in Figure 4). Scales similar to Karmakharm, Ale-

[16] Students primarily annotated the data.

tras, and Bontcheva (2019) could turn binary feedback into a spectrum and include multiple scales for other biases.

Although direct quotes can exhibit bias, they do not inherently impact neutrality (Recasens, Danescu-Niculescu-Mizil, and Jurafsky 2013). In our dataset, we observe significant disagreements regarding quotes (Section 5), indicating confusion among readers regarding their interpretation. Therefore, future iterations should incorporate different visual cues for quotes and may consider excluding them from the bias indication and training dataset.

Expanding the data collection phase could have enlarged the dataset but potentially bear design flaws. Hence, we decide to follow a user-centric design approach with a short collection phase to allow for quick iterations of feedback while showcasing data quality capabilities early on. While the RoBERTa model fine-tuned with BABE was used, NewsUnfold could have tested other models. However, RoBERTa performed superior in a previous study (Spinde et al. 2021b).

A common challenge in projects that rely on community contributions is keeping volunteers motivated over time (Soliman and Tuunainen 2015). With NewsUnfold, we aim to increase motivation by highlighting bias in a news reading application, offering a reason for people to use the platform that goes beyond annotation. The project targets reader groups similar to those interested in AllSides and Ground-News, which have demonstrated the viability of such business concepts. For testing and iterative feedback, we opted for a binary approach, feasible with our resources at the time, predicated on the assumption that feedback would primarily come from individuals valuing unbiased information. In later versions, NewsUnfold will incorporate insights from a recent literature review on media bias detection and mitigation (Xu and Diab 2024). The authors suggest accounting for cultural and group backgrounds in label creation and output generation. By adapting its output to readers' backgrounds, NewsUnfold could extend its appeal beyond those specifically seeking unbiased information. Gamification elements or unlocking additional content through giving feedback could further increase motivation (Zeng, Tang, and Wang 2017).

The use cases of the feedback mechanisms extend beyond NewsUnfold. Any digital and social media that includes text can apply the feedback mechanism to raise readers' awareness and collect feedback. NewsUnfold, as an application, integration, or browser plug-in, could offer an alternative to traditional news platforms. Incorporating feedback mechanisms with customizable classifiers, such as those for detecting misinformation, stereotypes, emotional language, generative content, or opinions, could allow users to analyze the content they consume in greater detail. Simultaneously, they contribute to a community dataset with an open-source purpose, which has shown potential in other applications (Cooper et al. 2010). We believe that by offering something useful, the feedback mechanism on NewsUnfold can gather valuable information in the long run, even if readers do not interact with it daily.

While systems like NewsUnfold can help understand bias and language, educate readers, and foster critical reading, we must closely monitor data quality and include readers' backgrounds while meeting data protection standards. Bias in the reader base or attacks from malicious groups, for example, any politically extreme group interested in shifting the classifier according to their ideology, could lead to a self-enforcing loop that inserts bias and skews classifier results towards a specific perspective, potentially harming minorities. To avoid deliberate attacks, we include spammer detection before training. In the future, we will monitor feedback beyond F1 scores and IAA as they only capture the agreement between raters, which is a standard measure in the media bias domain but does not fully indicate the quality of annotations. They help us set a baseline to check how bias detection systems handle human feedback, backed up by the manual analysis and NewsUnfold's ease of use.Other possibilities include a soft labeling approach (Fornaciari et al. 2021), adding adversarial examples into the training data, (Goyal et al. 2023), employing a HITL approach where experts try to break the model (Wallace et al. 2019), identifying and correcting perturbations (Goyal et al. 2023), or using a more complex probabilistic model for label generation (Law and von Ahn 2011).

Making the system and process transparent is critical to avoid misuse. Given the potential impact of skewed or misclassified bias highlights on reader perceptions, the system must communicate the impossibility of achieving absolute accuracy. Hence, the landing page informs readers about the possible inaccuracy to impart a clear understanding of classification limitations and ask for readers' help.We believe that even with the classification improvements of large language models such as GPT, assessing human perception of bias will always be crucial, and feedback mechanisms to assess such perception are becoming more critical for developing and constantly evaluating fair AI.

Limitations

Our team's Western education might add bias. Similarly, the Prolific study and the recruitment for NewsUnfold via LinkedIn might skew results due to the presence of more academic participants. The age range and education (Figure 8) of participants in Section 3 suggests a bias towards the digitally accustomed and educated, additional to a left slant (Figure 12). Although both studies involve relatively small samples, the results are nevertheless significant. Future research should examine larger and more diverse samples to evaluate how varying backgrounds and political orientations influence feedback behavior and quality. We implemented the feedback mechanism on the NewsUnfold platform to test if readers would give feedback in an environment as close as possible to a real news aggregator. Hence, we decided against a demographic survey to collect annotator data as it might negatively impact readers' experience. The more open study setting (compare Spinde et al. 2021c,d), with users exploring NewsUnfold freely, complicates the identification of factors affecting data quality. While the goal is to gather data from diverse readers, NewsUnfold currently controls for geographical diversity. Unlike the US pre-study, NewsUnfold had significant participation from Japan and Germany. Readers' backgrounds and quality control tasks must be im-

plemented in later iterations, for example, by implementing user accounts. Their data can be used to improve models (Law and von Ahn 2011) and fair classifiers (Cabitza, Campagner, and Basile 2023) accounting for backgrounds and protecting minorities or underprivileged groups. We did not collect demographic data or ask participants which device they used to view NewsUnfold in the UX study. Later studies must control for situations, experiences, attention, and perceptions of bias, which could diverge depending on personal backgrounds and used device.

Future Work

We plan to develop NewsUnfold into a standalone website with constantly updated content. Simultaneously, we aim to evaluate different feedback mechanisms for media bias classifiers and to extend our design's application beyond NewsUnfold. We plan to implement and test the feedback tool (2 in Figure 4) as a browser plugin[17] and social media integration. The value of the feedback mechanism lies in its adaptability across different platforms, using visual cues to enhance datasets for various bias types. Social media websites can add similar bias-highlighting mechanisms to make users more aware of potential biases and their influence (Spinde et al. 2022). Our next phase involves testing its integration in social media environments like X.

We will monitor the impact on user behavior in the long term and explore gamification and designs to increase engagement (Wiethof, Roocks, and Bittner 2022). Also, we will assess varied bias indicators, such as credibility cues (Bhuiyan et al. 2021), which have shown to be effective in similar studies (Yaqub et al. 2020; Kenning, Kelly, and Jones 2018), but need real-world validation. We further plan to add labels for subtypes of bias (Spinde et al. 2024). Advanced models like LLaMA (Touvron et al. 2023), BLOOM (Workshop et al. 2023), and GPT-4 (OpenAI 2023) may offer additional explanations on bias highlights (1 in Figure 4). When both models and data collection improve, it facilitates finding and comparing different outlets' coverage of topics and views of the general population. Simultaneously, controlling for personal backgrounds (Groeling 2013) could assist journalists and researchers in studying and understanding media bias, as well as its formation and expression over time.

7 Conclusion

We present NewsUnfold, a HITL news-reading application that visually highlights media bias for data collection. It augments an existing dataset via a previously evaluated feedback mechanism, improving classifier performance and surpassing the baseline IAA while integrating a UX study. NewsUnfold showcases the potential for diverse data collection in evolving linguistic contexts while considering human factors.

[17] A future experiment will determine if a plugin ensures quality annotations. However, its platform integration expands reach and diversity, making it our preferred choice.

Acknowledgments

This work was supported by the Hanns-Seidel Foundation (https://www.hss.de/), the German Academic Exchange Service (DAAD) (https://www.daad.de/de/), XR Hub Bavaria/Würzburg, and partially supported by JST CREST Grant JPMJCR20D3 Japan. None of the funders played any role in the study design or publication-related decisions. ChatGPT was used for proofreading.

References

An, J.; Cha, M.; Gummadi, K.; Crowcroft, J.; and Quercia, D. 2021. Visualizing Media Bias through Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 6(2): 2–5.

Ardèvol-Abreu, A.; and Zúñiga, H. G. 2017. Effects of Editorial Media Bias Perception and Media Trust on the Use of Traditional, Citizen, and Social Media News. *Journalism & Mass Communication Quarterly*, 94(3): 703–724.

Baumer, E. P. S.; Polletta, F.; Pierski, N.; and Gay, G. K. 2015. A Simple Intervention to Reduce Framing Effects in Perceptions of Global Climate Change. *Environmental Communication*, 11(3): 289–310.

Bhuiyan, M. M.; Horning, M.; Lee, S. W.; and Mitra, T. 2021. NudgeCred: Supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–30.

Brew, A.; Greene, D.; and Cunningham, P. 2010. Using Crowdsourcing and Active Learning to Track Sentiment in Online Media. In *Proceedings of the 2010 Conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, 145–150. NLD: IOS Press. ISBN 9781607506058.

Cabitza, F.; Campagner, A.; and Basile, V. 2023. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6): 6860–6868.

Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popović, Z.; and Players, F. 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307): 756–760.

Demartini, G.; Mizzaro, S.; and Spina, D. 2020. Human-in-the-loop Artificial Intelligence for Fighting Online Misinformation: Challenges and Opportunities. *IEEE Data Eng. Bull.*, 43: 65–74.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9: 48–59.

Druckman, J. N.; and Parkin, M. 2005. The Impact of Media Bias: How Editorial Slant Affects Voters. *The Journal of Politics*, 67(4): 1030–1049.

Eberl, J.-M.; Boomgaarden, H. G.; and Wagner, M. 2017. One Bias Fits All? Three Types of Media Bias and Their Effects on Party Preferences. *Communication Research*, 44(8): 1125–1148.

Eveland Jr., W. P.; and Shah, D. V. 2003. The Impact of Individual and Interpersonal Factors on Perceived News Media Bias. *Political Psychology*, 24(1): 101–117.

Feldman, L. 2011. Partisan Differences in Opinionated News Perceptions: A Test of the Hostile Media Effect. *Political Behavior*, 33(3): 407–432.

Fornaciari, T.; Uma, A.; Paun, S.; Plank, B.; Hovy, D.; and Poesio, M. 2021. Beyond Black & White: Leveraging Annotator Disagreement via Soft-Label Multi-Task Learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2591–2597. Online: Association for Computational Linguistics.

Furnham, A.; and Boo, H. C. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1).

Goyal, S.; Doddapaneni, S.; Khapra, M. M.; and Ravindran, B. 2023. A Survey of Adversarial Defenses and Robustness in NLP. *ACM Computing Surveys*, 55(14s): 332:1–332:39.

Groeling, T. 2013. Media Bias by the Numbers: Challenges and Opportunities in the Empirical Study of Partisan News. *Annual Review of Political Science*, 16(1): 129–151.

Hayes, A. F.; and Krippendorff, K. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1): 77–89.

He, Z.; Majumder, B. P.; and McAuley, J. 2021. Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4173–4181. Association for Computational Linguistics.

Hube, C.; and Fetahu, B. 2019. Neural Based Statement Classification for Biased Language. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, 195–203. Association for Computing Machinery. ISBN 978-1-4503-5940-5.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-Writing with Opinionated Language Models Affects Users' Views. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.

Karmakharm, T.; Aletras, N.; and Bontcheva, K. 2019. Journalist-in-the-Loop: Continuous Learning as a Service for Rumour Analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, 115–120. Hong Kong, China: Association for Computational Linguistics.

Kause, A.; Townsend, T.; and Gaissmaier, W. 2019. Framing Climate Uncertainty: Frame Choices Reveal and Influence Climate Change Beliefs. *Weather, Climate, and Society*, 11(1): 199–215.

Kenning, M. P.; Kelly, R.; and Jones, S. L. 2018. Supporting Credibility Assessment of News in Social Media Using Star Ratings and Alternate Sources. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–6. ACM. ISBN 978-1-4503-5621-3.

Law, E.; and von Ahn, L. 2011. *Human computation*. Springer International. ISBN 978-3-031-01555-7.

Lee, N.; Bang, Y.; Yu, T.; Madotto, A.; and Fung, P. 2022. NeuS: Neutral Multi-News Summarization for Mitigating Framing Bias. *arxiv*.

Liu, R.; Wang, L.; Jia, C.; and Vosoughi, S. 2021. Political Depolarization of News Articles Using Attribute-aware Word Embeddings. *arxiv*.

Mavridis, P.; Jong, M. d.; Aroyo, L.; Bozzon, A.; Vos, J. d.; Oomen, J.; Dimitrova, A.; and Badenoch, A. 2018. A human in the loop approach to capture bias and support media scientists in news video analysis. In *Proceedings of the 1st workshop on subjectivity, ambiguity and disagreement in crowdsourcing, and short paper proceedings of the 1st workshop on disentangling the relation between crowdsourcing and bias management*, volume 2276 of *CEUR workshop proceedings*, 88–92. CEUR-WS.

Mosqueira-Rey, E.; Hernández-Pereira, E.; Alonso-Ríos, D.; Bobes-Bascarán, J.; and Fernández-Leal, A. 2022. Human-in-the-Loop Machine Learning: A State of the Art. *Artificial Intelligence Review*.

OpenAI. 2023. GPT-4 Technical Report. *arxiv*.

Otterbacher, J. 2015. Crowdsourcing Stereotypes: Linguistic Bias in Metadata Generated via GWAP. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, 1955–1964. ACM. ISBN 978-1-4503-3145-6.

Park, S.; Kang, S.; Chung, S.; and Song, J. 2009. NewsCube: Delivering Multiple Aspects of News to Mitigate Media Bias. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 443–452.

Pennycook, G.; and Rand, D. G. 2022. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nature Communications*, 13(1).

Powers, D. M. W. 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.*, 2.

Pryzant, R.; Diehl Martinez, R.; Dass, N.; Kurohashi, S.; Jurafsky, D.; and Yang, D. 2020. Automatically Neutralizing Subjective Bias in Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01): 480–489.

Raykar, V. C.; and Yu, S. 2011. Ranking Annotators for Crowdsourced Labeling Tasks. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, 1809–1817. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781618395993.

Raza, S.; Reji, D. J.; and Ding, C. 2022. Dbias: detecting biases and ensuring fairness in news articles. *International Journal of Data Science and Analytics*.

Recasens, M.; Danescu-Niculescu-Mizil, C.; and Jurafsky, D. 2013. Linguistic models for analyzing and detecting biased language. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1650–1659.

Ribeiro, F. N.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. P. 2018. Media Bias Monitor : Quantifying Biases of Social Media News Outlets at Large-Scale. In *Twelfth International AAAI*

Conference on Web and Social Media, 290–299. AAAI Press. ISBN 978-1-57735-798-8.

Shaw, A. D.; Horton, J. J.; and Chen, D. L. 2011. Designing Incentives for Inexpert Human Raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, CSCW '11, 275–284. New York, NY, USA: Association for Computing Machinery. ISBN 9781450305563.

Sheng, V. S.; and Zhang, J. 2019. Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 9837–9843.

Soliman, W.; and Tuunainen, V. K. 2015. Understanding Continued Use of Crowdsourcing Systems: An Interpretive Study. *Journal of Theoretical and Applied Electronic Commerce Research*, 10(1): 1–18.

Spinde, T.; Hamborg, F.; Donnay, K.; Becerra, A.; and Gipp, B. 2020. Enabling News Consumers to View and Understand Biased News Coverage: A Study on the Perception and Visualization of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, 389–392. Association for Computing Machinery. ISBN 978-1-4503-7585-6.

Spinde, T.; Hamborg, F.; and Gipp, B. 2020. An Integrated Approach to Detect Media Bias in German News Articles. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, 505–506. Association for Computing Machinery. ISBN 978-1-4503-7585-6.

Spinde, T.; Hinterreiter, S.; Haak, F.; Ruas, T.; Giese, H.; Meuschke, N.; and Gipp, B. 2024. The Media Bias Taxonomy: A Systematic Literature Review on the Forms and Automated Detection of Media Bias. arXiv:2312.16148.

Spinde, T.; Jeggle, C.; Haupt, M.; Gaissmaier, W.; and Giese, H. 2022. How Do We Raise Media Bias Awareness Effectively? Effects of Visualizations to Communicate Bias. *PLOS ONE*, 17(4): 1–14.

Spinde, T.; Kreuter, C.; Gaissmaier, W.; Hamborg, F.; Gipp, B.; and Giese, H. 2021a. Do You Think It's Biased? How To Ask For The Perception Of Media Bias. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Spinde, T.; Plank, M.; Krieger, J.-D.; Ruas, T.; Gipp, B.; and Aizawa, A. 2021b. Neural Media Bias Detection Using Distant Supervision With BABE - Bias Annotations By Experts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Dominican Republic.

Spinde, T.; Richter, E.; Wessel, M.; Kulshrestha, J.; and Donnay, K. 2023. What do Twitter comments tell about news article bias? Assessing the impact of news article bias on its perception on Twitter. *Online Social Networks and Media*, 37-38: 100264.

Spinde, T.; Rudnitckaia, L.; Kanishka, S.; Hamborg, F.; Bela; Gipp; and Donnay, K. 2021c. MBIC – A Media Bias Annotation Dataset Including Annotator Characteristics. *Proceedings of the iConference 2021*.

Spinde, T.; Sinha, K.; Meuschke, N.; and Gipp, B. 2021d. TASSY - A Text Annotation Survey System. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

Stumpf, S.; Rajaram, V.; Li, L.; Burnett, M.; Dietterich, T.; Sullivan, E.; Drummond, R.; and Herlocker, J. 2007. Toward Harnessing User Feedback for Machine Learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces*, IUI '07, 82–91. New York, NY, USA: Association for Computing Machinery. ISBN 1595934812.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. *arxiv*.

Vaccaro, M.; and Waldo, J. 2019. The effects of mixing machine learning and human judgment. *Communications of the ACM*, 62(11): 104–110.

Vraga, E. K.; and Tully, M. 2015. Media Literacy Messages and Hostile Media Perceptions: Processing of Nonpartisan versus Partisan Political Information. *Mass Communication and Society*, 18(4): 422–448.

Wallace, E.; Rodriguez, P.; Feng, S.; Yamada, I.; and Boyd-Graber, J. 2019. Trick Me If You Can: Human-in-the-Loop Generation of Adversarial Examples for Question Answering. *Transactions of the Association for Computational Linguistics*, 7: 387–401.

Weil, A. M.; and Wolfe, C. R. 2022. Individual differences in risk perception and misperception of COVID-19 in the context of political ideology. *Applied cognitive psychology*, 36(1): 19–31.

Wessel, M.; Horych, T.; Ruas, T.; Aizawa, A.; Gipp, B.; and Spinde, T. 2023. Introducing MBIB - the First Media Bias Identification Benchmark Task and Dataset Collection. In *Proceedings of 46th International ACM SIGIR Conference (SIGIR 23)*. ACM.

Wiethof, C.; Roocks, T.; and Bittner, E. A. C. 2022. Gamifying the Human-in-the-Loop: Toward Increased Motivation for Training AI in Customer Service. In *Artificial Intelligence in HCI*, 100–117. Cham: Springer International Publishing. ISBN 978-3-031-05643-7.

Workshop, B.; Scao, T. L.; Fan, A.; Akiki, C.; Pavlick, E.; and Ilić, S. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arxiv*.

Xintong, G.; Hongzhi, W.; Song, Y.; and Hong, G. 2014. Brief survey of crowdsourcing for data mining. *Expert Systems with Applications*, 41(17): 7987–7994.

Xu, J.; and Diab, M. 2024. A Note on Bias to Complete.

Yaqub, W.; Kakhidze, O.; Brockman, M. L.; Memon, N.; and Patil, S. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. ACM. ISBN 978-1-4503-6708-0.

Zeng, Z.; Tang, J.; and Wang, T. 2017. Motivation mechanism of gamification in crowdsourcing projects. *International Journal of Crowd Science*, 1(1): 71–82.

Ethics Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, in Section 4.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, in Section 6.
   (e) Did you describe the limitations of your work? Yes, in Section 6.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in Section 6.
   (g) Did you discuss any potential misuse of your work? Yes, in Section 6.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, in Section 4 and Section 6.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? Yes.
   (b) Have you provided justifications for all theoretical results? Yes.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N.A., as to the best of our knowledge, this is the first work done for HITL for media bias in written news articles.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? Yes, in Section 6.
   (e) Did you address potential biases or limitations in your theoretical framework? Yes, in Section 6.
   (f) Have you related your theoretical results to the existing literature in social science? Yes, in Section 1 and Section 2.
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Yes, in Section 1 and Section 6.
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? Yes, in Section 4.
   (b) Did you include complete proofs of all theoretical results? Yes, in Section 5.
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. Full code will be submitted upon acceptance to guarantee anonymity. However, the collected data is available on https://doi.org/10.5281/zenodo.8344891.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, in Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? N.A.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? No.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, especially as this is the first iteration of the system.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, in Section 6.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes.
   (b) Did you mention the license of the assets? N.A.
   (c) Did you include any new assets in the supplemental material or as a URL? Yes, and additional material in the code on acceptance.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, see Appendix A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. In the pre-study, no personally identifiable information is stored. In the data collection phase, no data besides location data feedback, and website interactions on NewsUnfold is stored.
   (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes, in Section 6.
   (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? Yes, according to the guidelines.
6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...
   (a) Did you include the full text of instructions given to participants and screenshots? Yes, see Appendix A and Appendix C.
   (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? N.A., no participant risks were identified as our application focuses on a regular news consumption process similar to most news platforms.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, in Section 5.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, in Section 4 and Section 4.

A Appendix: Feedback Mechanism Study Texts

Data Processing Agreement

Who are we and how do we use the data we collect from you through this survey? This research study is being conducted by the Media Bias Research Group. We are a group of researchers from various disciplines with the goal of developing systems and data sets to uncover media bias or unbalanced coverage in articles. This study is anonymous. That means that we will not record any information about you that could identify you personally or be associated with you. On the basis of the collected data, we aim to publish scientific papers on presentations of articles that help to detect biased language, but these publications do not allow any inference to you as an individual. Once the study is published, the anonymized data might be made available in a public data repository. Your rights to access, change, or move your information are limited insofar as the data may no longer be modified after the data has been published in anonymized form. The reason for this is that we need to manage your information in specific ways for the research to be reliable and accurate. Once anonymized, we will not be able to delete your data. The study itself is not hosted on Prolific, but on a dedicated external server. Once the survey is complete, you will be shown a unique code that you can enter in the Prolific form. Participation in this study is voluntary. You may choose not to participate and you may withdraw at any time during the study without any penalty to you. If you have any questions about the study or study procedures, you may contact the Media Bias Research Group, info@media-bias-research.org.

- I agree to the processing of my personal data in accordance with the information provided herein.(Checkbox)

Demographic Survey

1. What gender do you identify with? (Female, Male, Other, Prefer not to say)
2. What is your age? (Input field for number)
3. What is the highest level of education you have completed? (8th grade, Some high school, High school graduate, Vocational or technical school, Some college, Associate degree, Bachelor's degree, Graduate work, Ph.D., I prefer not to say)
4. What is the level of your English proficiency? (Proficient, Independent, Basic)
5. Do you consider yourself to be liberal, conservative, or somewhere in between? Please slide to record your response. (Very liberal to Very conservative, -10 to 10 point slider)
6. How often on average do you check the news? (Never, Very rarely, Several times per month, Several times per week, Every day, Several times per day)

Info on Media Bias

Before you can start we will now provide you with a few examples that should help you to understand possible media bias instances better. For each example, a sentence with a biased word (blue colored) is shown first followed by its impartial representation (green colored). Please note that bias is different from negative sentiment. Bias is ambiguous and subtle, it can be positive, negative, or not even have a particular sentiment but it still can imply or intensify the opinion/emotion.

Subjective Intensifiers:

Schnabel himself did the fantastic reproductions of Basquiat's work.

Schnabel himself did the accurate reproductions of Basquiat's work.

Strong labels:

'The people want the Truth!': Trump gloats over the loss of American media jobs.

'The people want the Truth!': Trump tweets over the loss of American media jobs.

One-sided terms:

Concerned Women for America's major areas of political activity have consisted of opposition to gay causes, pro-life law...

Concerned Women for America's major areas of political activity have consisted of opposition to gay causes,

anti-abortion law...

Attention Check on Bias

How is bias connected to sentiment? Based on the information that was provided to you earlier, please select the correct option.

- Bias is the same as negative sentiment.
- Bias can be both positive, negative or even not have particular sentiment. (correct answer)
- Bias is the same as positive sentiment.
- Bias is not connected to sentiment at all.

Trust Check

Can we trust your data for scientific research? For example, if you failed to pay attention to some questions, please answer 'No'. Please answer honestly, you will receive full payment regardless of your answer. Please select one option. (Yes, you can trust my data for scientific research. No, you may not want to trust my data for scientific research.)

B Appendix: Detailed UX Survey Results

How did you like NewsUnfold? (10 responses) Six participants expressed a strongly positive sentiment, stating, for instance, that they found it innovative. Two expressed that the bias detection might need some calibration, one found it "okay", and one remained unsure.

Ease of Use: How easy was NewsUnfold to use? (13 responses) Participants were asked to rate the ease of use of NewsUnfold on a 10-point scale, with 10 indicating high ease of use. The average rating for ease of use was 8.46, with a median score of 9, implying that users found NewsUnfold user-friendly and intuitive.

How did NewsUnfold impact your reading? (12 responses) 6 participants stated it made them read more carefully, critically, and slowly instead of skimming. Two stated bias was easier to recognize because they had to think twice. One said they did more active thinking about what bias is. One didn't feel much impact in unbiased articles. One wanted it in all of their browsing. One said it made them argue with the AI instead of skimming the article.

How did you feel about giving feedback on the sentences? (11 responses) Three participants found it easy to give feedback, while two reported it felt either difficult because it disrupted their reading flow or because it felt like a chore. Two participants felt unsure, with one skipping the tutorial. Two reported only doing it when they would have more time. One stated only to give feedback when disagreeing with the classification. One participant appreciated sharing their reasoning in the free-text field 3.

How do you feel about the highlights in the text? (10 responses) Nine participants liked the highlights and found them helpful, one calling it their favorite part. However, one participant found it distracting and raised concerns about highlighting quotes as biased.

Net Promoter Score (NPS): How likely would you recommend NewsUnfold to a friend, family, or colleague? (13 responses) The calculated mean NPS was 6.23. This score indicates participants were neutral to slightly in favor of recommending NewsUnfold.

How do you like the User Interface of NewsUnfold? (11 responses) 9 participants found it easy and clean, with one stating the "look is sleek and appropriate for a modern website". One participant experienced a bug using Firefox on mobile and described it as a "bit sluggish." One participant found the UI "a bit bland."

What irritated you? Did you encounter any problems? (10 responses) Bugs and irritations included the character limit in the free-text field 3, the multiple steps in the feedback window on mobile devices, overlays blocking the text on Firefox mobile, out-of-line tooltips, and encountering jumping buttons. One person expressed a slight annoyance in instances they disagreed with the classifier. One person was confused because they skipped the tutorial. Two didn't encounter problems.

Anything else you want to share with us? (1 response) One participant suggested that it might be interesting to add a note indicating that direct quotes are more likely to be biased and may not necessarily reflect the opinions of the authors.

C Appendix: Material Bias, Demographics, and Additional Screenshots

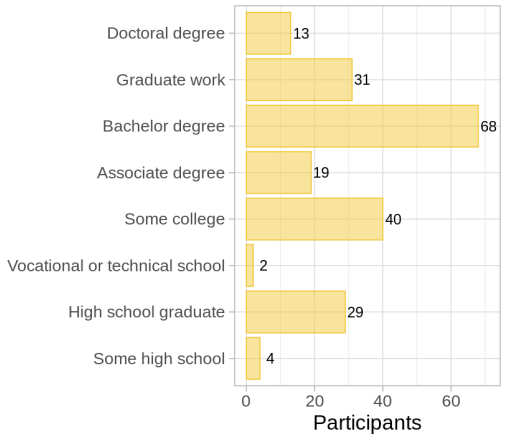

Figure 7: Education of participants in the feedback mechanism study.

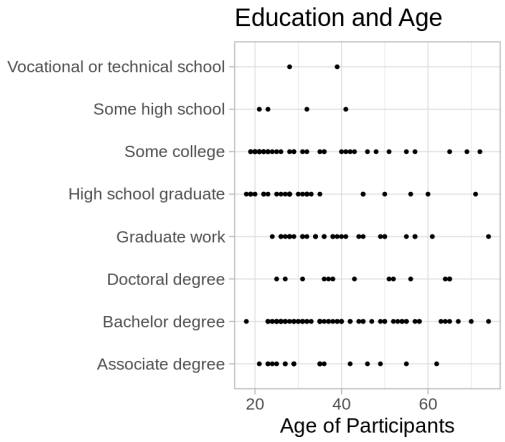

Figure 8: Education of participants mapped to their age in the feedback mechanism study.

Dep. Variable	R-squared	Adj. R-squared	F-statistic	Prob (F-statistic)	Log-Likelihood	No. Observations	AIC	BIC	Df Residuals	Df Model	Covariance Type	Const coef	x1 coef
y	0.009	-0.002	0.8424	0.361	240.50	100	-477.0	-471.8	98	1	nonrobust	0.0553	.000004

Table 4: OLS Regression Results for F1 Score of the NewsUnfold Feedback

	Experts biased	Experts not biased	Classifier biased	Classifier not biased
Left article	16	21	8	29
Right article	24	21	12	33

Table 5: Bias rating of sentences in feedback mechanism study articles by classifier and experts.

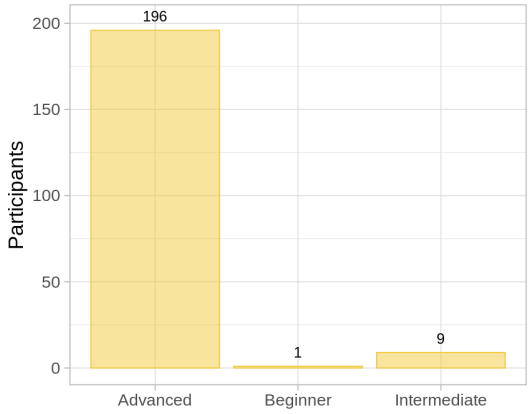

Figure 9: English proficiency of participants in the feedback mechanism study.

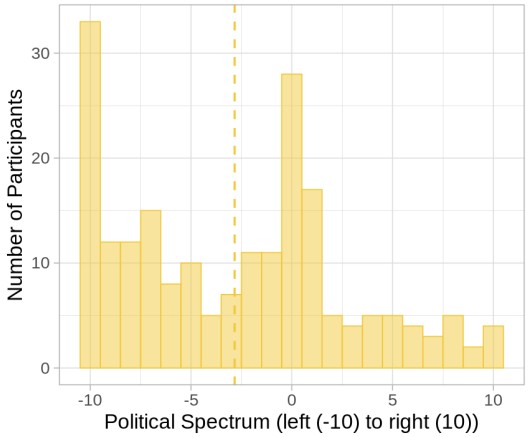

Figure 11: Political orientation of participants in the feedback mechanism study.

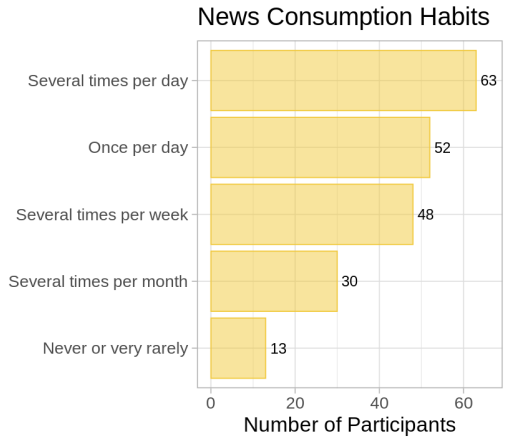

Figure 10: News consumption habits of participants in the feedback mechanism study.

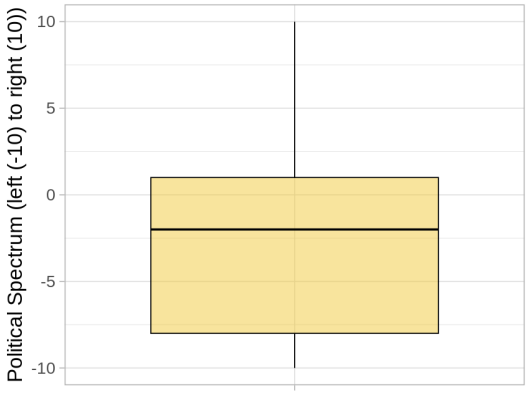

Figure 12: Average political orientation of participants in the feedback mechanism study.

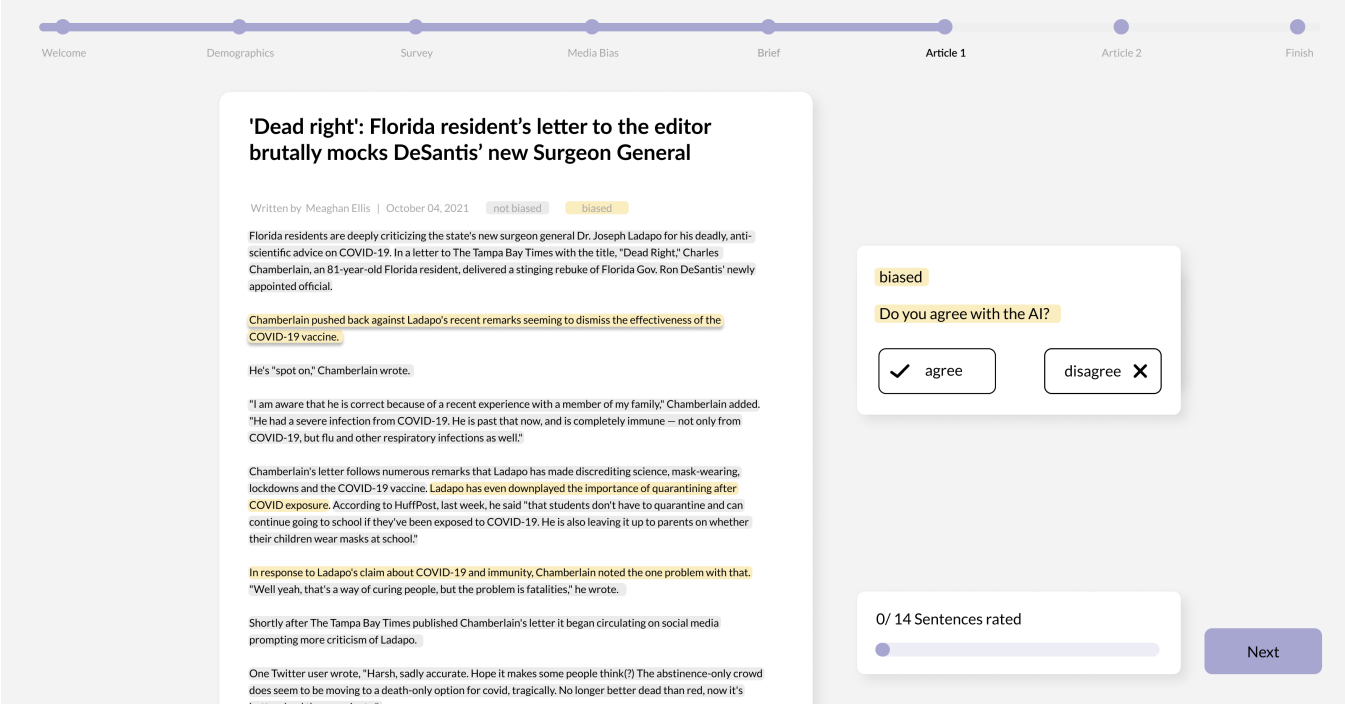

Figure 13: Screenshot of the highlight mechanism on the study platform for the preliminary feedback mechanism study.

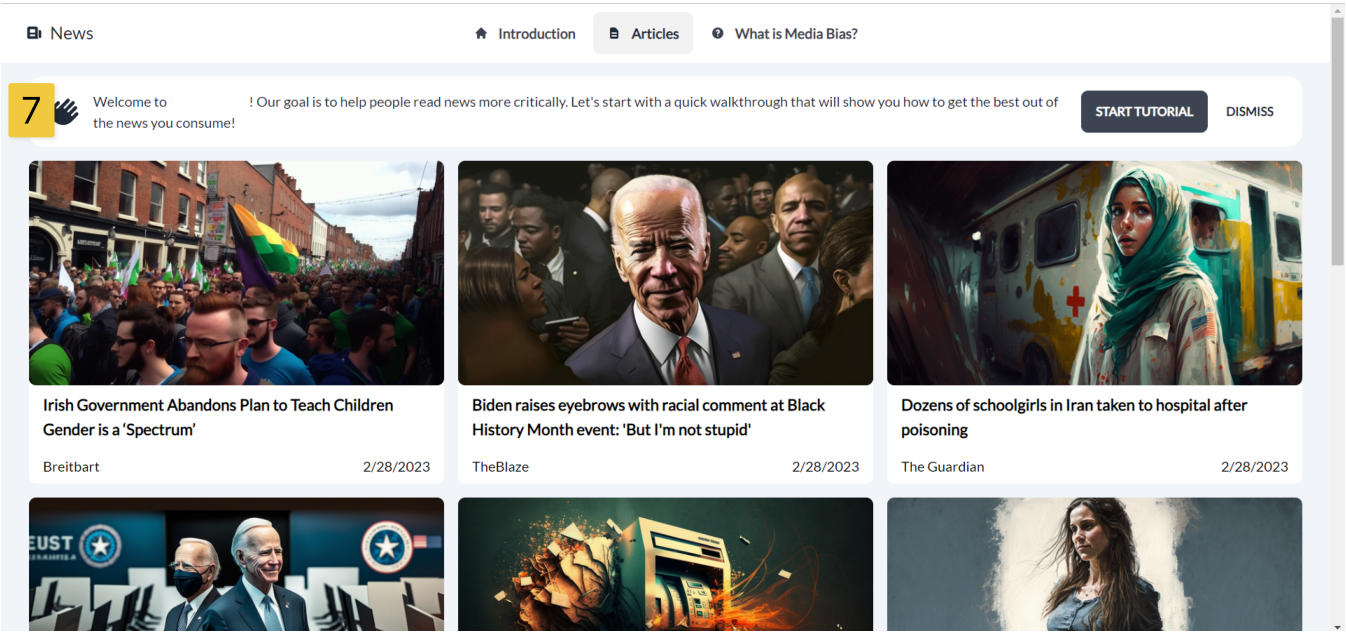

Figure 14: Screenshot of Articles Overview Page in NewsUnfold with the option to start the tutorial. Table 2 explains the elements with yellow numbers.

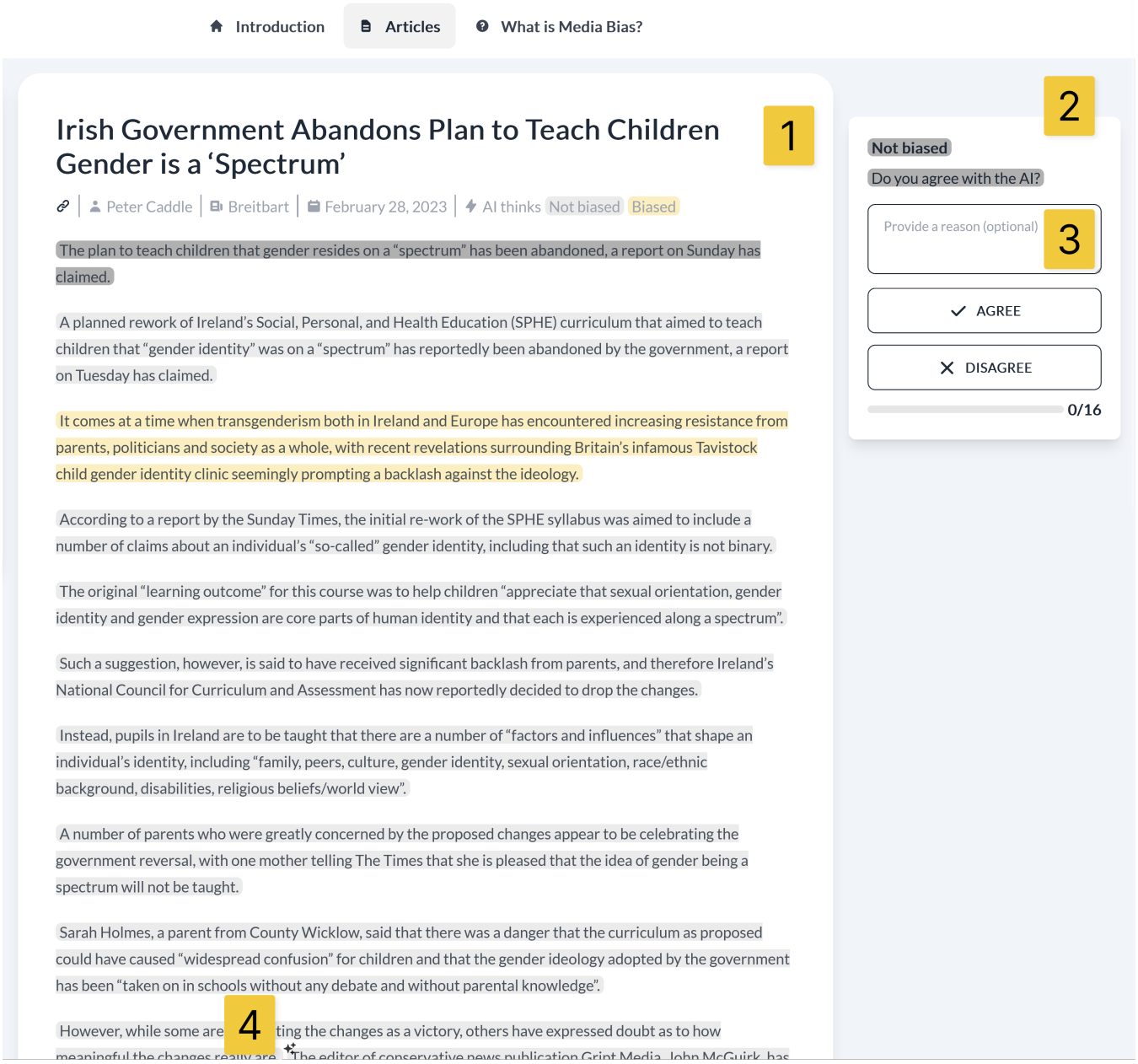

Figure 15: NewsUnfold Article View. Table 2 explains the elements with yellow numbers.

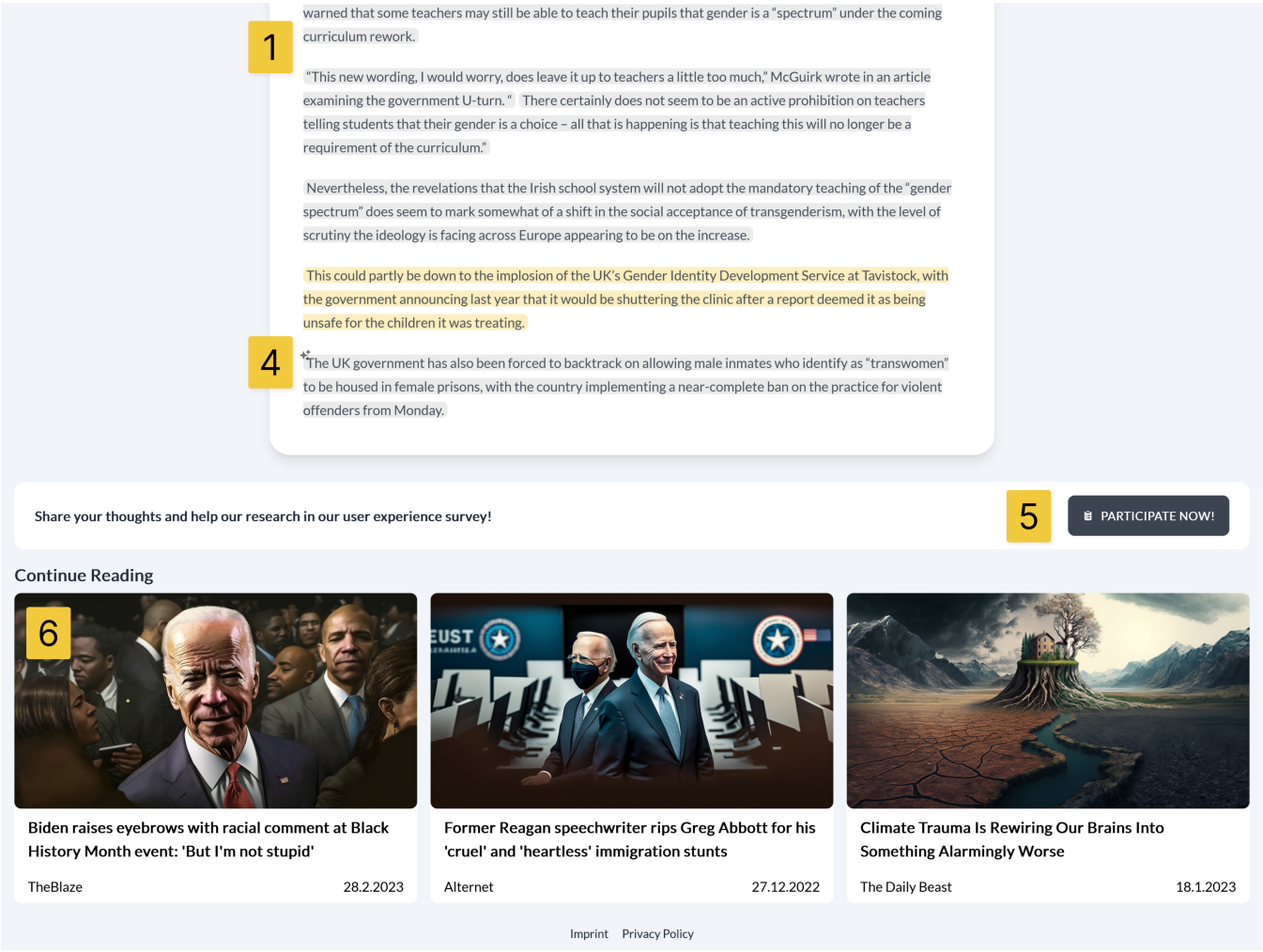

Figure 16: Screenshot of recommended articles and button to the UX survey on NewsUnfold. Table 2 explains the elements with yellow numbers.