NewsComp: Facilitating Diverse News Reading through Comparative Annotation

Md Momen Bhuiyan*
Virginia Tech
Blacksburg, USA
momen@vt.edu

Sang Won Lee
Virginia Tech
Blacksburg, USA
sangwonlee@vt.edu

Nitesh Goyal
Google Research
New York, USA
niteshgoyal@acm.org

Tanushree Mitra
University of Washington
Seattle, USA
tmitra@uw.edu

ABSTRACT

To support efficient, balanced news consumption, merging articles from diverse sources into one, potentially through crowdsourcing, could alleviate some hurdles. However, the merging process could also impact annotators' attitudes towards the content. To test this theory, we propose comparative news annotation; that is, annotating similarities and differences between a pair of articles. By developing and deploying *NewsComp*—a prototype system—we conducted a between-subjects experiment (N = 109) to examine how users' annotations compare to experts', and how comparative annotation affects users' perceptions of article credibility and quality. We found that comparative annotation can marginally impact users' credibility perceptions in certain cases; it did not impact perceptions of quality. While users' annotations were not on par with experts', they showed greater precision in finding similarities than in identifying disparate important statements. The comparison process also led users to notice differences in information placement and depth, degree of factuality/opinion, and empathetic/inflammatory language use. We discuss implications for the design of future comparative annotation tasks.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; *Empirical studies in HCI*.

KEYWORDS

News Reading; Annotation; Comparison; Design

ACM Reference Format:
Md Momen Bhuiyan, Sang Won Lee, Nitesh Goyal, and Tanushree Mitra. 2023. NewsComp: Facilitating Diverse News Reading through Comparative Annotation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3544548.3581244

*The author conducted a portion of the work while interning at the University of Washington.

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581244

1 INTRODUCTION

News media often produces content that is significantly biased in favor of a particular ideology, especially on contentious topics [31, 51, 72], and news consumers are affected by such biases [21]. Therefore, developing an informed opinion on a subject requires critically consuming news content from multiple sources. While the internet gives users access to news from multiple sources, when given choices, people tend to choose content that aligns with their viewpoints due to confirmation-seeking tendencies [14, 55, 55, 70, 73]. Furthermore, the task of engaging with multiple perspectives is not easy and probably not performed equitably by all users [34, 66]. One potential solution to this problem could be to use experts (i.e., journalists) to combine news items on an event from varying sources into a single story. However, a limited number of experts would likely find it difficult to manage the volume of news stories generated by news outlets from around the world. On the other hand, studies have shown that crowdworkers' output can be significantly correlated with experts' in some annotation tasks [2, 3, 9, 11, 74]. Building on such results, this work explores whether crowdsourcing could be a viable approach to combining news articles from varying sources. For a lay user, such a crowdsourcing task can be broken down into two aspects of comparative annotation: (i) finding similarities and (ii) finding important disparities. These annotations can be useful to both news consumers and fact-checkers, whether professional or crowdsourced (e.g., BirdWatch[1]). For everyday news consumers, merged articles can provide balanced perspectives on news events. Second, fact-checkers can use similarity/dissimilarity annotations to validate claims through multiple sources or trace the origin of specific statements. Besides, a by-product of any annotation task is that performing the task could also affect the annotators attitude towards the content, in our case, the news articles or the issue at hand. In this work, we ask:

RQ1. *How well do users perform comparative annotation?*

RQ2. *How does comparative news annotation affect users' perceptions of credibility and news quality?*

Here, we use a simplified notion of comparative annotation: statements that are similar and statements that are dissimilar but important. Using the concept of comparative news annotation, we developed and tested *NewsComp* (see the interface in Figure 2): a prototype that allows readers to compare and annotate similar and contrasting statements between only two news articles. *NewsComp* has two components: (i) a comparative or side-by-side view of two articles from different sources, and (ii) an annotation tool. Specifically, the annotation tool allows performing the two annotation

[1]https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation

tasks: (i) identifying similar statements across a pair of articles and connecting them with lines (3 in Figure 2), and (ii) identifying disparate statements (statements with no similarity) from each article that are important and should be included in the other article in the pair (6 in Figure 2). To design the tool, we conducted a series of think-aloud formative studies with Google Drawings to observe the annotation process. During those interviews, we noticed users considering different criteria for annotation. For example, some considered only the content in each statement, while others considered the underlying themes behind the statements. Informed by the think-aloud sessions, we ask annotators to provide the reasoning behind their annotation (e.g., 2 in Figure 2).

To answer our research questions, we conducted a between-subjects experiment with *NewsComp* in a controlled environment. We recruited 109 participants using Facebook advertising, which allowed us to recruit users from a large and diverse pool. Participants were randomly assigned to either the treatment or the control group. For the study, we used two pairs of articles on two contentious topics: immigration and abortion. To generate gold standards for the sake of comparison, we recruited two experts from the university's Department of Communication (one of whom had five years experience as a journalist) and asked them to perform annotation and rate the articles. Our experts found different degrees of similarity between the two pairs of articles; specifically, the pair of immigration articles had high dissimilarity (high contrast), while the pair of abortion articles pair was highly similar (low contrast). During the study, users in the treatment read and annotated a pair of articles on the same topic, and then responded to a questionnaire designed to address our research questions (related to perceptions of credibility and quality). Meanwhile, users in the control group read a pair of articles on separate events without adding annotations and responded to the same questionnaire. We analyzed the extent to which news consumers' annotations matched experts' annotations, the impact of article topic and users' news expertise (knowledge of current events, perceived value of media literacy) on annotation quality, the reasons behind annotations, and how the treatment group's article perception compared to the control.

Regarding RQ1, we found that users performed poorly on both annotation tasks. However, they had higher precision in finding similarities than in identifying important statements among the disparate statements. We also found that filtering out annotations based on the number of users who annotated an item can rule out some false positives in finding similar statements, thus improving their collective F1 score. In our study, ruling out annotations made by fewer than five or six users produced the highest F1 score. Users with low current event knowledge made more annotations and had higher recall. We also found that while annotating statement similarity, users provided different types of criteria, such as seeing connections when two statements discuss the same person, location, date, quote, or other information. Among statements with no similarity, when annotating if a statement is important and should be included in the other article, users sometimes marked a statement important if it provided clarification or elaboration on other statements or if it provided a missing perspective. Furthermore, we found that both generic words (e.g., "quote" and "similar") and article-specific words (e.g., "lawsuit") mentioned in the rationales can differentiate incorrect annotations from correct ones. Perhaps such generic words in rationales can be used to filter out false positives in annotations on articles on different topics. Comparing the articles, annotators also saw differences in perspectives presented, information placement, depth of detail, amount of factual/opinion statements, empathetic presentation, and use of inflammatory language. Perceptions of *NewsComp* itself were mixed, though skewed more towards positive than negative. Regarding RQ2, we found that the treatment group's credibility ratings were significantly different compared to the control group's for high-contrast articles. For low-contrast articles, users in both groups performed similarly. There were no significant effects on perceptions of quality. Overall, this study indicates that we can leverage the comparative annotation mechanism to engage users in reading multiple perspectives. However, since users produce annotations with high error rates, creating tools to assist in annotation could help reduce errors. We discuss applications for annotated data, such as developing a holistic view of an event from multiple news sources, teaching machines to discern article quality, training machine learning algorithms to generate better annotations, and assisting fact-checkers in their work. We conclude with implications for the design of future comparative annotation tasks, such as modularizing into subtasks, providing supporting features to reduce load, and supporting co-annotation by multiple users.

2 BACKGROUND AND RELATED WORKS

In this section, we begin by providing some background on media bias and multi-perspective online news consumption. Thereafter, we discuss related research on designing annotation tools for making sense of information and the effects of such annotations.

2.1 The Need for Multiperspective News Consumption

While news articles should ideally follow established journalistic practices, various forms of biases and inaccuracies are injected into articles during the content production process. This begins in the information gathering stage, where journalists must select events and related facts from sources. In doing so, news publishers can influence which topics readers perceive to be relevant by selectively reporting on topics of their choosing [65]. Next, journalists include and exclude information from sources (e.g., press releases, other news articles, and studies), shaping the perspective on the event. In the writing phase, journalists make stylistic choices which may reflect their view of the news item, thereby producing biased coverage. For instance, journalists may introduce bias through the use of labeling ("a senator" vs. "a Republican senator") and word choice ("illegal alien" vs. "undocumented immigrant"). Such methods allow journalists to promote a particular interpretation of a topic [24].

Research suggests that a majority of news consumers are affected by media bias [21, 44, 54] in different ways [23, 65]. Such bias can influence voting or election outcomes [20, 23, 54]. Furthermore, media bias promotes polarization in public opinion, especially on contentious topics [72]. Some scholars argue that media bias challenges the pillars of American democracy [37, 83]. Overall, these works point to the need to consume news from diverse perspectives to deal with biases in the media.

2.2 Barriers to Multiperspective News Consumption Online

Lazarsfeld et al. introduced the two-step flow model of communication, referring to the two gatekeeping stages that occur before an individual forms an opinion on a subject: first by news organizations, and then by opinion leaders in the individual's social circle [45]. Even though the internet has democratized access to information, including news, news consumption in the internet age still seems to follow the two-step flow model in communication in at least two ways: news selection and consumption [16, 45, 67]. First, personalization algorithms act as filters for content selection; thus, they perform a gatekeeping function similar to that of opinion leaders in the pre-internet age [57, 67, 71]. Second, pervasive, echo chamber–esque news comment sections tend to promote opinions from opinion leaders with views aligned with users' own beliefs [16, 35]. One problematic aspect of this internet-based, two-step communication is that users may not be aware of the second gatekeeping stage, given that algorithmic effects are often hidden, and partisan biases of opinion leaders in comment sections may also be obscured by anonymity [17, 67]. Even when readers become aware of content with a political slant opposed to their own, they may lack the motivation to consume that content that due to political polarization and confirmation-seeking tendencies [6, 13, 19, 27, 43, 55, 73]. Indeed, some research has found that while people might read more content when using diverse content selection tools, this leads primarily to an increase in the amount of content consumed, not the diversity of the content [14]. One reason for this outcome could be individual differences between diversity-seeking and challenge-averse people; challenge-averse people may tend not to consume diverse content [53]. To address this bias, some prior works developed mechanisms to promote diverse news selection through design tools, such as NewsCube [58–61], or through nudges to read alternative viewpoints [52]. Though these prior works demonstrated improved exposure—that is, clicks or visits to news sites with diverse political slants—there is a gap in our understanding of whether design tools can encourage critical engagement and whether such engagement affects users' perceptions of the news. Furthermore, these tools do not ensure that people read articles on the same events from politically diverse sources. We aim to bridge this gap by bundling pairs of articles on the same event from differing perspectives in a comparative annotation interface to test engagement and its effects.

2.3 Designing for Information Consumption through Comparison

Scholarship on reasoning, comprehension, and learning outlines different mechanisms in understanding information, whether users learn from data or the structure of information [4, 38, 78]. Sometimes, reading multiple sources alone can help change a reader's mental model of a subject [10, 69]. Comparison can further help people recognize common features shared across items or identify features that distinguish them [15, 38, 77, 78]. Some suggest that a comparison mechanism allows users to create broad concept categories by grouping similar concepts in either a bottom-up approach (clustering) or a top-down approach (assigning existing categories) [84]. In HCI research, creating design elements or affordances for easier information consumption is not new. For example, interactive elements allow users to choose where to go or what to read next [26]. Such design elements can assist readers in constructing a cognitive model to support a thorough understanding of a news event. This construction of a news schema is supported by providing signals—layouts, visual elements, and textual structures—to news readers that meet their expectations for news [40]. For example, a newspaper reader's understanding of certain affordances (e.g., section labels, such as "opinion") may assist them in contextualizing and understanding the information [75]. It may even boost recall significantly [62]. Building on a similar idea, we design a comparative interface where pairs of articles are displayed side by side to facilitate the comparison process for users.

2.4 Annotating Using the Crowd and its Effect

In educational settings, annotation has long been used to boost reading comprehension, critical thinking and meta-cognitive skill improvement. Many online annotation tools have been developed over the last decade, including Gibeo [7] and HyLighter [46]. Much of the research on annotation focuses on effects in classroom settings and often takes the form of collaborative annotation [56]. On annotation tasks, prior research also suggests that users' performance may vary with demographic characteristics, political biases, task complexity, and subject matter [5, 9, 30, 48, 50]. Some research indicates that annotation technology could improve users' effectiveness and efficiency in information-related tasks, such as search tasks [39]. Informed by such outcomes, we explore the effectiveness of comparative annotation in identifying content quality in a news consumption setting.

3 FORMATIVE STUDY FOR DESIGNING *NEWSCOMP*: THINK-ALOUD INTERVIEWS

To design an interface where users can simultaneously compare news articles, we began with a set of think-aloud interviews [2]. We used two types of prototypes during this phase: a high-fidelity interface powered by algorithmic similarity metrics and a Google Drawings board. By high-fidelity, we mean an interactive prototype with a working front-/back-end. During this phase, we conducted a total of 10 think-aloud interviews with four members of our research groups (none of whom are authors of this paper) and six undergraduate students with different majors (communication, political science, and computer science) and levels of news consumption expertise. We used a separate set of participants for interviews with each prototype. These interviews revealed several aspects for consideration in our design. Below, we discuss how our interview process evolved and summarize the insights we gathered.

3.1 Inaccurate Algorithmic Annotation

This phase consisted of six interviews conducted with an interface we designed to support comparison through similarity scores obtained from a state-of-the-art sentence transformer and its semantic sentence matcher[3]. All six participants mentioned that the algorithm's similarity annotations were inaccurate. This effect may

[2] All of our studies were approved by the institutional review board at our university
[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

result from differences in how users and algorithms compute similarities. Whereas a human can take a statement, event, surrounding sentences and other contextual aspects into consideration while finding similarity, algorithms are likely to prioritize word similarity.

3.2 Annotating on Google Drawings

Next, we moved towards asking users to perform the annotation task using a Google Drawings board. Here, we laid out a pair of news articles side by side on the drawing board by segmenting them into sentences (see Figure 1). Then, we asked our participants to perform two annotation tasks, one after another: (i) find similar statements between the two articles and draw lines between them, and (ii) revisit statements without corresponding, similar statements to see if they convey important information that should be included in the other article. Since users reported inaccuracies in algorithmic statement matching, we refrained from providing machine-generated annotations as suggestions. Figure 1 shows a screenshot of the interface and the annotations provided by one of the participants. As in the prior interviews, we recorded participants' actions. These tasks took longer compared to the previous interviews, as users iterated over each statement multiple times to add annotations. After completing the task, we asked semi-structured interview questions to clarify the participants' actions. Our observations and the participants' answers helped us identify several considerations for our study, outlined below.

3.2.1 Criteria for Finding Similarities: Content and Underlying Theme. After the annotation task sessions, we asked participants to elaborate on the criteria they used to find similarities. From their responses, we found two similarity criteria: content and underlying theme. Though one participant mentioned structural position (e.g., the lede in a news article) as a criterion, none of the other participants mentioned it. In our final deployment, we asked users to provide rationales for why statements were similar.

3.2.2 Considerations for Finding Important Information Present in Only One Article. When asked about how participants chose statements conveying important information that was worthy of inclusion in the other article, they mentioned two considerations. The first of these considerations was whether a statement fit the narrative of the other article. Participants suggested that a statement in article A should only be labeled "important" if it fits the narrative in article B and provides important context missing from article B. Such missing information might include statements detailing what happened after an event or how something happened. Even when a statement did not fit within the narrative of the other article, some suggested that such a statement should still be included (*"This task is difficult, because the two articles are focusing on different narratives ... the other article does a bad job at portraying them as such [smuggled people being seen as inhuman]. Therefore, I think that bringing the human cost displayed here into the other article would be helpful." - P3*). Participants mentioned that any information among the dissimilar statements that could be inferred from other statements or the context of an article did not need to be included. In light of these nuances in the reasoning behind answering the questions, the result suggests that participants were more critically engaged in reading and comparing the two articles during this exercise than they were during the exercise that presented ML-recommended results. Therefore, in our final design, we asked users to provide rationales behind annotations when finding something important to be included in the other article.

3.2.3 Readers Have Varying Expertise in Identifying Similarities. During these interviews, we noticed that readers' different levels of expertise in reading news and knowledge on the topic led to different annotations. During two interviews with participants who had less news expertise, we presented another participant's thematic connection annotations and asked if the interviewees could understand the original annotator's intentions. Neither participant was able to explain the annotator's intentions. These findings led us to one of our research questions; specifically, how user characteristics relating to news expertise affect comparative annotation.

Overall, both studies revealed to us that a comparative annotation task could strengthen users' engagement with news content, and we implemented such a task in our final design for *NewsComp*.

3.3 The *NewsComp* Interface and How It Works

Based on the findings from our two formative studies, Figure 2 shows the final *NewsComp* interface we implemented. At the top of the page, instructions for the tasks are laid out (1). The task asks the user to read the pair of articles and perform two steps: (i) find similar statements within the articles and create links between them, and (ii) check if a statement with no corresponding similar statement in one article is important and worthy of inclusion in the other article. To help users understand how to perform the two steps, there is a button below the task description that opens a video/GIF showing a tutorial of both steps. In our deployment, we made a point of reminding users to watch the tutorial before proceeding. A toolbar below the taks description (2) helps users perform the first step. Specifically, when users link two statements, the toolbar shows the statements that are highlighted (in yellow) and provides a text box where they can supply a rationale for the connection right before finalizing the annotation. Below the toolbar, two news articles are presented side by side (as in our initial interface) with the article title at the top, followed by statements segmented exactly as in the original article. To limit preexisting biases, there are no links to the canonical source, nor any reference to the authorship. The interface also hides any nontextual components (i.e., videos and images). To begin connecting statements, users click the two statements to select them. Selected statements are highlighted with a yellow background. To mark a statement as worthy of inclusion in the other article, each statement contains a checkbox (6). There is also a text box below this checkbox for users to provide a rationale for marking a statement as important. When a user connects two statements, the checkbox and text box for the statement are programatically disabled and grayed out. After selecting a statement from each article, a dashed arrow representing a pending connection appears (4). When a user finalizes a connection by filling in a rationale and clicking "connect," the dashed arrow (⇢) changes to a solid arrow (→) to represent a confirmed connection (3). In addition, a list of connections appears to the right of the articles (5) to allow users to delete connections they have created. Users can delete a connection by clicking the cross button next to it. After finishing both tasks, users scroll to the bottom of the page and click a button to confirm they have finished the annotation tasks. This system was built with a React front end with Bootstrap CSS, and a Flask back-end server

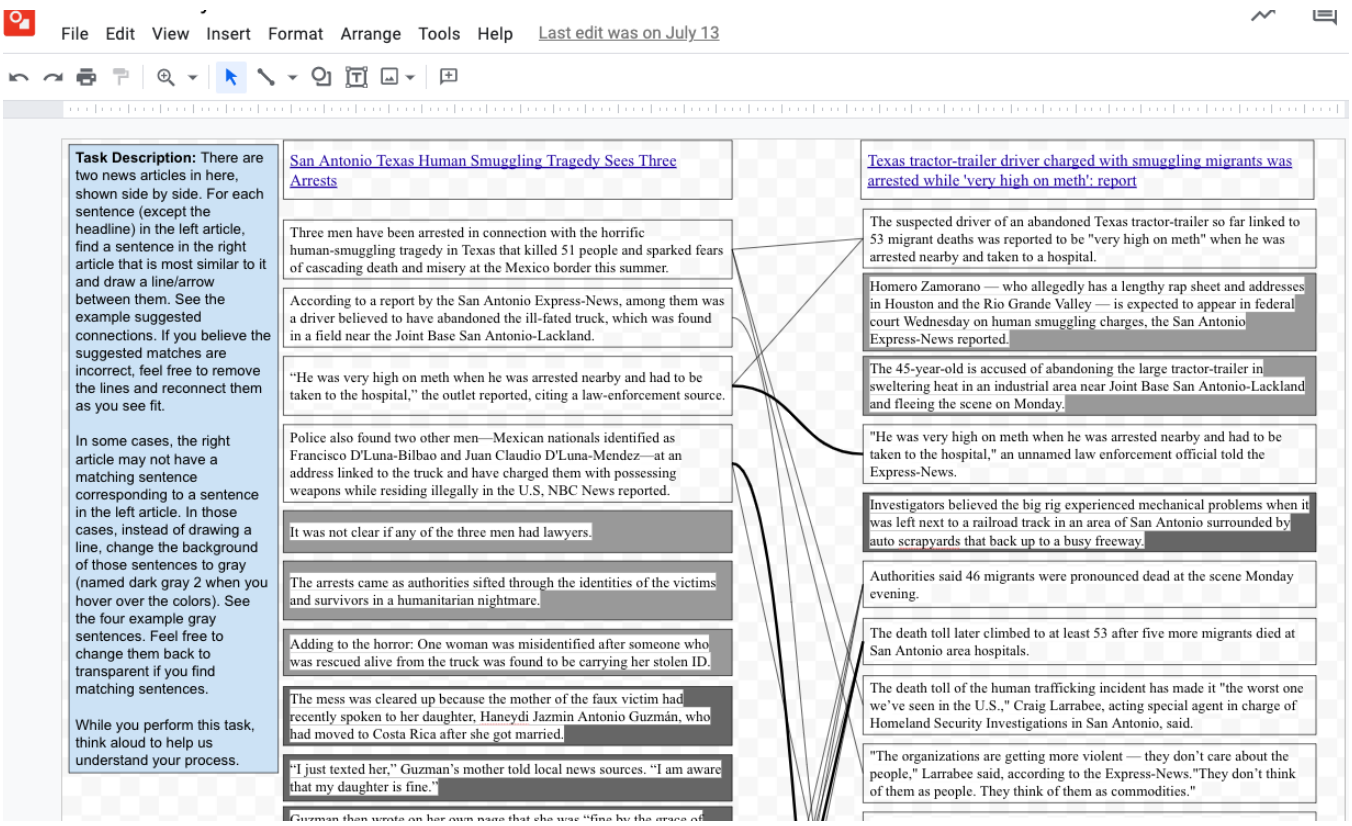

Figure 1: A Google Drawings board used for think-aloud interviews. Similar to the high-fidelity interface, two articles are presented side by side here. Users can use all the available tools to link similar statements or highlight dissimilar statements that contain important information which should be included in the other article.

with a MySQL database. To draw the connection lines, we used the Leader-Line[4]. To scrape news articles, we used NewsPaper[5]. For the formative study, we used Sentence Transformer[6] to generate sentence embedding and calculate sentence pair similarity score.

4 EVALUATION STUDY

Using *NewsComp*, we examine two research questions:

RQ1. *How well do users perform comparative annotation?*

RQ2. *How does comparative news annotation affect users' perceptions of credibility and news quality?*

To answer, we conducted a between-subjects experiment in a controlled environment using two pairs of news articles. We created a separate interface for the control users. In the control interface, only one article is shown at a time. Figure 3 shows the study design for our experiment with four experimental groups: two treatment and two control groups. Each group read two news articles. While the treatment group was able to view two articles on the same topic simultaneously, the control users read articles on different topics sequentially to account for any learning effects from recall and comparison. All four groups read stories from two sources with different political leanings. We randomized article location (left or right) for the treatment group and article order (first or second) for the control group to account for any ordering effect.

4.1 Article Selection

For our study, we picked two politically contentious topics (immigration and abortion), where reading content from diverse perspectives can be beneficial. The topics were chosen from recent news coverage at the time of the user studies. For each topic, we chose articles published at least two weeks prior to deployment to limit possible recall effects. Pairs were selected by finding two articles from politically opposed sources under the same story bundle on Google News. When choosing article pairs, we picked pairs with different levels of similarity and difference. Since the article pair on abortion(E_2) had more similarities than differences, we categorized the pair into the *low-contrast* category. On the other hand, the pair on immigration (E_1) had more apparent differences than similarities, so we categorized it into the *high-contrast* category. This categorization was confirmed by our experts' gold standard annotations (see 4.6), which identified more than 50% of the article text as similar in the low-contrast pair while identifying less than 25% of the text as similar in the high-contrast pair. The selected articles (E_1L, E_1R, E_2L, E_2R) are reproduced in Appendix B.

[4]https://github.com/anseki/leader-line
[5]https://newspaper.readthedocs.io
[6]https://huggingface.co/sentence-transformers

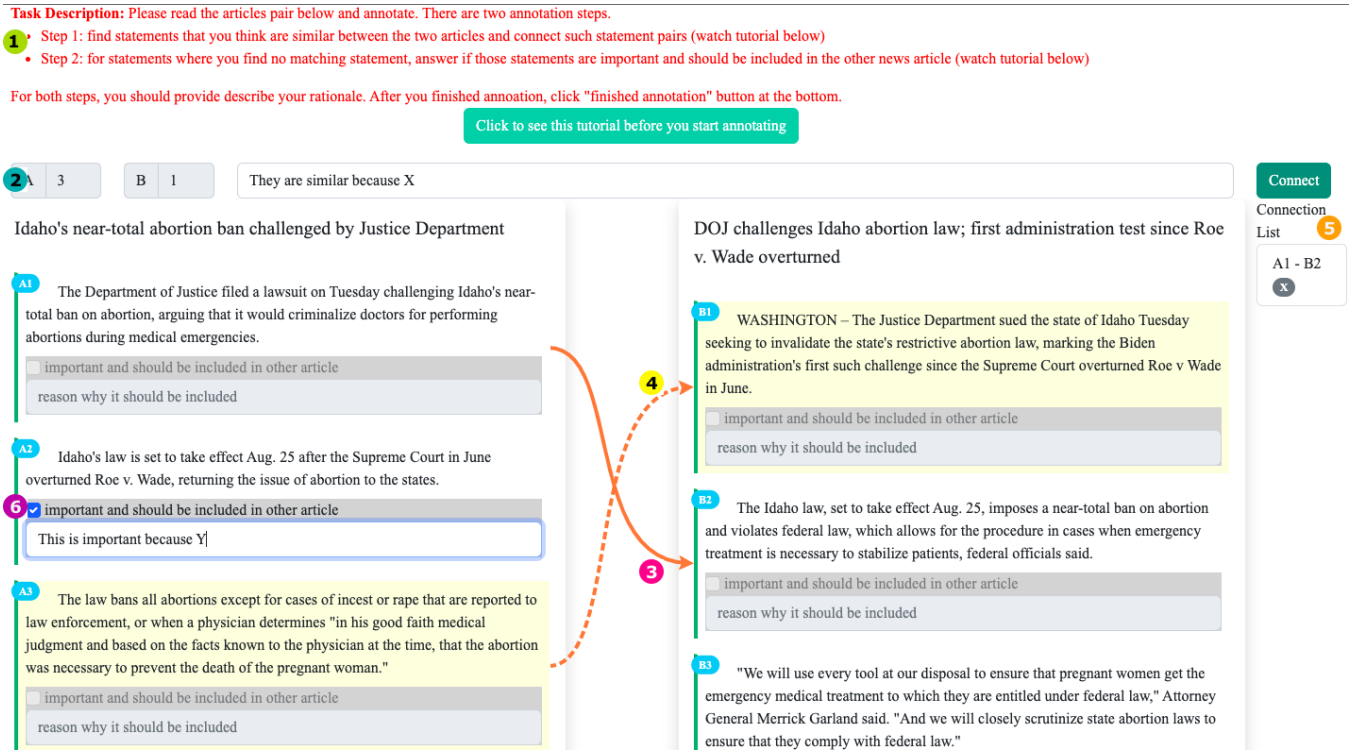

Figure 2: *NewsComp* Interface showcasing features with random annotations. 1 Annotation instructions in two steps: find and connect similar statements, and answer if a statement with no corresponding, similar statement is important to include in the other article. 2 Toolbar to finalize a connection by providing a rationale 3 A solid arrow representing a connection already created 4 A dashed arrow indicating that the connection creation tool is active 5 A list of connections including deletion buttons 6 The importance question in step 2.

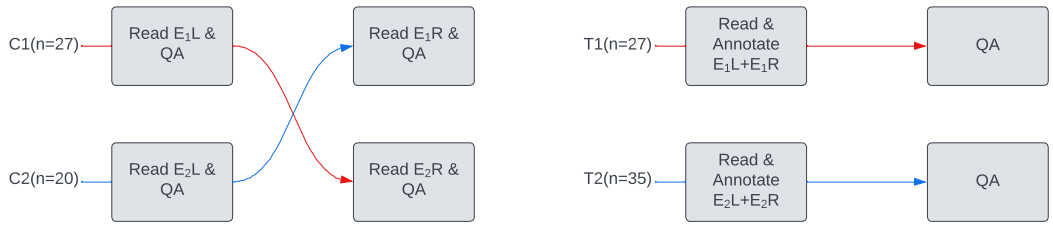

Figure 3: Study design showing the experimental conditions for each of the four participant groups. Here, C and T respectively represent control and treatment groups; the number of participants is given in parentheses. Because we used four articles, we had two control groups (C1–2) and two treatment groups (T1–2). Article E_XP represents an article about event X from a source with political leaning P (L for left, R for Right). Articles with $X = 1$ were about immigration, while those with $X = 2$ were about abortion. For example, E_2R indicates a news article about abortion from a right-leaning source. The E_1 pair had high contrast, while the E_2 pair had low contrast. In the study, we randomized the order/position of the articles for each participant.

4.2 Measuring Credibility, Quality, Current Event Knowledge, Media Literacy

To address RQ1, we measured two expertise metrics: *current event knowledge* and *value of media literacy*. Here, the value of media literacy differentiates users' general media literacy from their expertise on topics related to our study. To answer RQ2, we use perceptions of article *credibility* and *quality*, and we compare the treatment groups' assessments with the control groups'. Below, we discuss how we measured each metric.

4.2.1 Current Event Knowledge (CEK) and Value of Media Literacy (VML). To capture users' news-related knowledge, we adapted the Current Event Knowledge measure created by Maksl et. al. [47]. Here, we included questions relevant to the two chosen article topics and some other timely topics (see Table 1). To measure users'

Credibility [49]	i) It is biased (I) ii) It is not fair (I) iii) It doesn't tell the whole story (I) iv) It is not accurate (I) v) It cannot be trusted (I)
Quality [76]	i) It shows multiple viewpoints ii) It has information on causes and consequences iii) It provides balanced viewpoints
Current Event Knowledge (CEK) [47]	i) Who is/was Kamala Harris? (a) President (b) **Vice President** (c) **Senator from California** (d) UN Ambassador ii) What does the recent Supreme Court ruling overturning Roe v. Wade entail? (a) **Abortion is not a constitutionally protected right** (b) In Missouri, abortion is legal before 24 weeks (c) All US states allow abortion for rape and incest (d) **There is confusion about abortion rights relating to miscarriage and ectopic pregnancy** iii) In California (a) **everyone, including undocumented individuals, has the right to access their crime report** (b) there are no immigrant detention facilities (c) **state and local police officers cannot inquire about an individual's immigration status during a routine check** iv) How is the Fed responding to the high inflationary economic condition? (a) **Raising the interest rate** (b) Lowering the interest rate (c) Keeping the interest rate the same
Value of Media Literacy (VML) [79]	i) Two people might see the same news story and get different information from it ii) People are influenced by news whether they realize it or not iii) News is designed to attract an audience's attention iv)Writing techniques can be used to influence a viewer's perception v) People should accept information from the news on face value (I) vi) It is the job of citizens to overcome their own biases in consuming news vii) People need to critically engage with news content viii) The main purpose of the news should be to entertain viewers (I)

Table 1: Questionnaires used in the study. Credibility and quality questions were asked after reading or annotating. (I) means these items were inverted for analysis. The correct responses appear in boldface. The CEK questionnaire contains multiple-choice questions, while the VML, credibility, and quality questions are 5-point Likert items. The VML and CEK items were presented in the pre-survey.

perceptions of media literacy, we used a prior scale created by Vraga et. al. [79]. To calculate CEK scores, we added 1 point for each right answer and deducted 1 point for each wrong answer. In our study, CEK ranged from -1 to 7, with 4 being the median value. For VML, we average the responses across items. The score for VML ranged from 1 to 8, with 6 being the median. Finally, for both measurements, we use the median score to create a binary response variable with values "low" and "high." For example, users scoring less than 4 in CEK were categorized as low-CEK users and vice versa.

4.2.2 Credibility & Quality. We used a five-item questionnaire by Meyer et al. [49] to measure users' perceptions of credibility for every news item (see Table 1). In our study, we found that this measure had high internal consistency (Cronbach's $\alpha = 0.85$), close to the result in Meyer et al. For news quality detection, we use a modified version of the questionnaire suggested by Urban et al. [76] (see Table 1). Similarly to the credibility questionnaire, participants' responses to these questions showed high internal consistency (Cronbach's $\alpha = 0.89$). We measured all of these items on a 5-point Likert scale, from "Strongly Disagree" (1) to "Strongly Agree" (5). Note that scores for the credibility items are inverted for analysis.

4.3 Recruitment

To recruit participants for our final *NewsComp* interface, we used Facebook advertising for two weeks in August 2022. This method allowed us to organically recruit diverse participants from a large pool. We also did limited advertising on news subreddits (such as, `r/politics`, `r/moderatepolitics`, `r/news`, `r/neoliberal`, and `r/conservative`) through private messages from our research group's Reddit account, reaching about 40 users. Two users responded to these messages. Since the article topics are US-centric, our ads targeted people living in the US with interest in news-related pages. Thus, our study result may not be generalizable beyond the context of the US. The advertisement led users to a pre-survey to sign up for the study. In the pre-survey, we screened users with the following study eligibility criteria: (i) I am 18 years old or over, (ii) I reside in the United States, (iii) I read at least one news article online every day, (iv) My primary language for news consumption is English, and (v) I use a laptop or a desktop for online news reading. Besides these criteria, we also screened out users who failed attention checks, had an IP address outside of the US, or spent very little time (less than half of the median time, which was two minutes) in the pre-survey. Overall, 685 users clicked on the survey, out of which 238 passed the screening criteria. We invited all of these participants to the study in multiple batches. Ultimately, 109 participants completed the study. Participants who completed the study were compensated with $7.50 gift cards for the 30-minute study, in line with the state's minimum wage.

4.4 Procedure

Users who met the screening criteria in the pre-survey filled out the rest of the survey, which contained questions about demography, including gender, age, race, education, and political affiliation, and

the two news expertise measures. Within three days of submitting the survey, we invited eligible users to participate in the study via email. In the email, we provided the consent document, instructions for using the interface, and a link to the study website. When users clicked the link to access the study website, they were randomly assigned to one of four groups (two treatment and two control) to ensure a balanced sampling design. Since some people who clicked the link ultimately did not complete the study, the final group sizes are not exactly equal. Recall that after visiting the study website, treatment users were asked to view a tutorial on how to add annotations before reading and annotating articles. After finishing the annotation process, users responded to the credibility and quality questionnaire. The annotation interface, including the articles and annotations, was still visible at this time. In the control condition, there was no annotation task, and participants read only one article at a time. The control users additionally responded to the credibility and quality questionnaire after reading each article (see Figure 3).

4.5 Participant Pool

Due to screening and self-selection bias, our study participants were not equally distributed in certain demographic dimensions, such as age. Additionally, since one of our aims was to identify how users with different demographic characteristics compare in their annotations in RQ2, we invited more users in the treatment condition. Though our pre-survey had a large number of categories for different demographic characteristics, we merged groups with small numbers of respondents for more meaningful differentiation. Figure 4 shows this distribution. Here, we grouped two consecutive age groups, merged participants from nonwhite races together, and divided respondents by education into those with any university degree versus those with no degree. Generally, participants were skewed towards younger age groups, male, white, college-educated, and politically left-leaning.

4.6 Gold Standard Generation

To compare annotation quality, we used expert-produced gold standards. To obtain gold standards for both the annotations and the perception metrics, we recruited two senior PhD students from the university's Department of Communication for an interview session. Both had past experience in conducting news content analysis research and were familiar with both topics used in our study. One of the experts also worked as a journalist for more than five years. To generate credibility and quality perception scores, both were given the original links to the news articles and asked to rate the RQ1 questions[7]. They were also allowed to do any outside research they wished. After rating their perceptions, we provided the article pairs in a Google Drawings board and asked them to add annotations, much like the process from our think-aloud interviews. After adding annotations independently, the expert annotators met with each other to resolve any conflicts. Through this method, we built consensus around our gold standard annotations.

[7]This task was performed before generating annotations. We did not ask them to work within the comparative interface, with the assumption that they would rate the articles accurately irrespective of any comparison.

5 RESULTS

To answer both research questions, we compared users' annotation and perception responses against the expert-produced gold standards. For this purpose, we performed a series of analyses involving mean testing, analysis of variance, and regression. For free-form text responses (specifically, the annotation rationales), three authors performed thematic coding (see supplemental document for data and codes). Below, we outline the results.

5.1 RQ1: How well do users perform comparative annotation?

5.1.1 Performance on Connection-Making. Figure 5 shows the distribution of total connections users made, correct connections made, their recall, and precision relative to the gold standard. The median number of connections between articles fell below the gold standard for both article pairs, as shown in Figure 5(a). Between the two article topics (immigration and abortion), users on average made more connections—correct or otherwise—between the abortion articles (the low-contrast article pair). Furthermore, users' precision was significantly better on the abortion articles than on the immigration articles (Mann-Whitney U = 136.5, $p < 0.001$). However, we did not find any significant difference in recall.

5.1.2 Performance on Importance Detection. Figure 6 shows the distribution of total importance annotations users made, correct importance annotations, recall, and precision relative to the gold standard. As shown in Figure 6(a), the median number of importance annotations was consistently above the gold standard. Between the two article pairs, users on average annotated more items as important—correctly or otherwise—in the immigration articles (the high-contrast article pair). Though users' recall was high due to the large numbers of importance annotations added, their precision was low, with the median per article being less than or equal to 0.25. Comparatively, for connection-making annotation task, users' median precision and recall are higher than these median for importance annotations.

5.1.3 Annotation Agreement. Next, we examined how users agreed on annotations among themselves by plotting the count of users annotating each item. Figure 7 shows the distribution of this analysis. Here, we differentiated between agreement on correct and incorrect annotations. For the connection-making task (Figure 7(a)), we found that the annotation count for correct items was significantly higher than the count for incorrect items (Mann-Whitney U = 517.0, $p < 0.01$). However, for importance detection (Figure 7(b)), the corresponding counts did not differ significantly. Furthermore, we also observed some outliers (high agreement in some annotations) in the connection-making annotation task not made by the experts. In Figures 7(c) and (d), we examine how annotation performance changes by filtering annotations by the number of concurring users. Overall, a threshold of five users produces the highest F1 score (55%) for the connection task, while peak performance (41%) occurs at a threshold of six users for the importance detection task.

5.1.4 Effect of News Expertise. We investigated whether levels of news expertise affect users' annotation performance with *NewsComp* by performing Mann-Whitney U tests on precision and recall

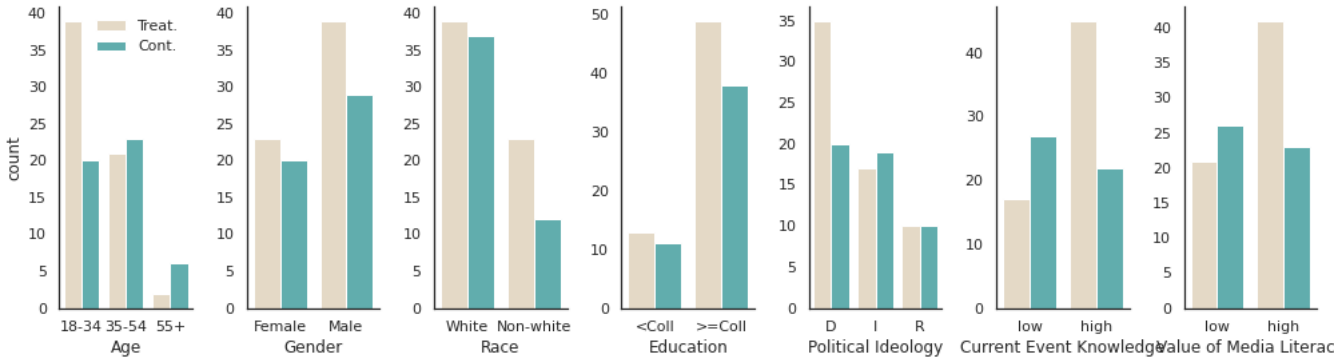

Figure 4: Graphs showing the distribution of participant demographics across the treatment and control groups.

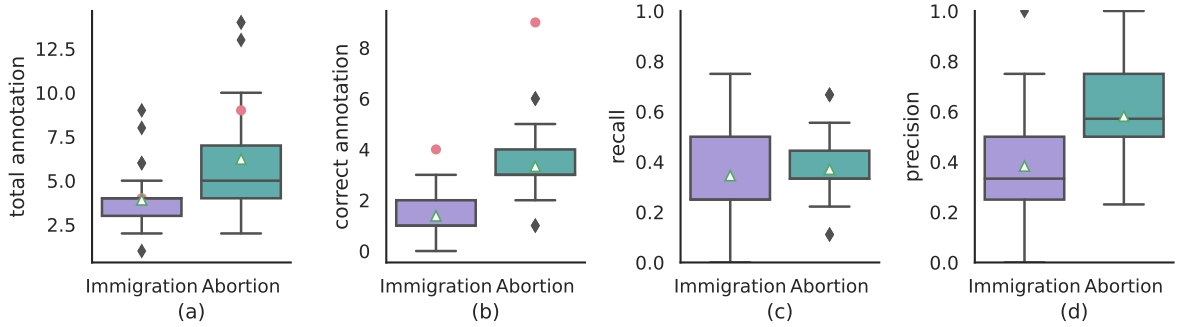

Figure 5: Distribution of connection making by users. White and red dots respectively represent the average and experts' annotation.

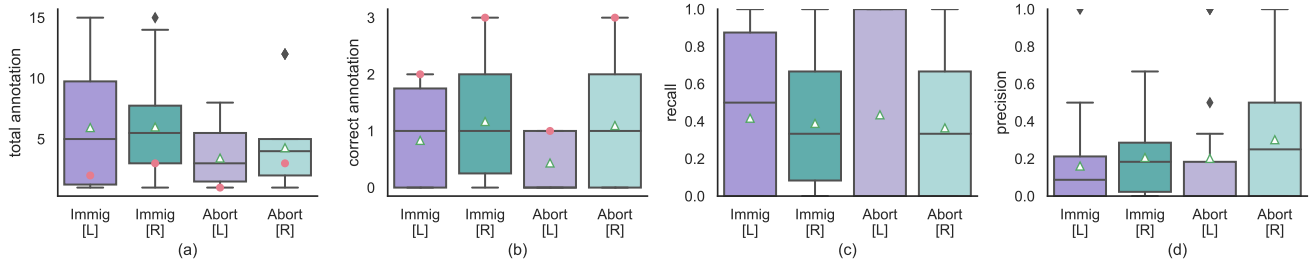

Figure 6: Distribution of importance detection by users. White and red dots respectively represent the average and experts' annotation.

scores (for connection-making and importance detection). Figure 8 shows the distribution of those scores divided by two news expertise criteria, Current Event Knowledge (CEK) and Value of Media Literacy (VML) perception. Here, only for recall scores on the importance detection task (Figure 8 (c)), we found significant differences in values between low and high CEK (Mann-Whitney U = 810.5, $p < 0.05$). None of the other tests detected a statistically significant difference. Furthermore, we modeled these variables against user characteristics with a series of linear models (M1–M4 in Table 4 in Appendix C). These models were similarly significant ($\beta = 0.35, p < 0.01$ in M1). However, since the model effect sizes (R^2) are low (0.10), there may be confounding variables not accounted for in these models affecting the outcome. It is therefore difficult to make any strong claims in this regard, and we instead leave this to future experiments.

5.1.5 Reasons behind the Annotations. Three of the authors thematically coded the rationales provided by the participants during annotation. Each author performed initial coding and discussed the results with the others to agree upon a code book. Then, the first

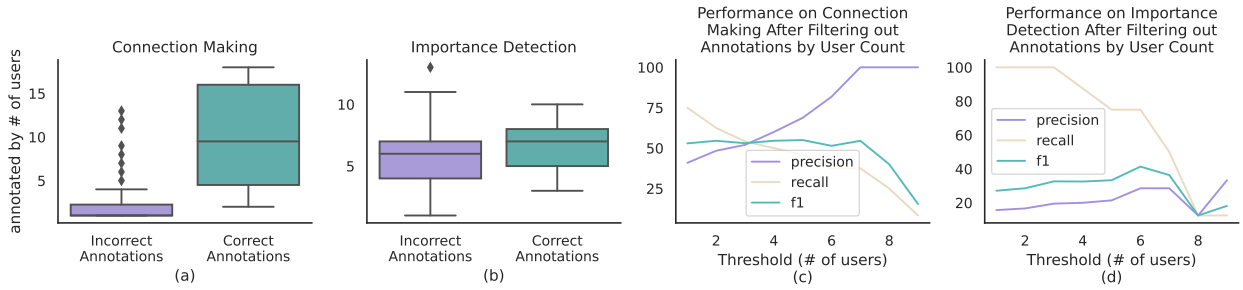

Figure 7: (a & b) User agreements on incorrect and correct annotations. (c & d) We filtered annotations by the number of concurring users to see how annotation performance changes as the threshold moves. Here, for connection making and importance detection, the F1 scores peak at five (55%) and six (41%) users, respectively.

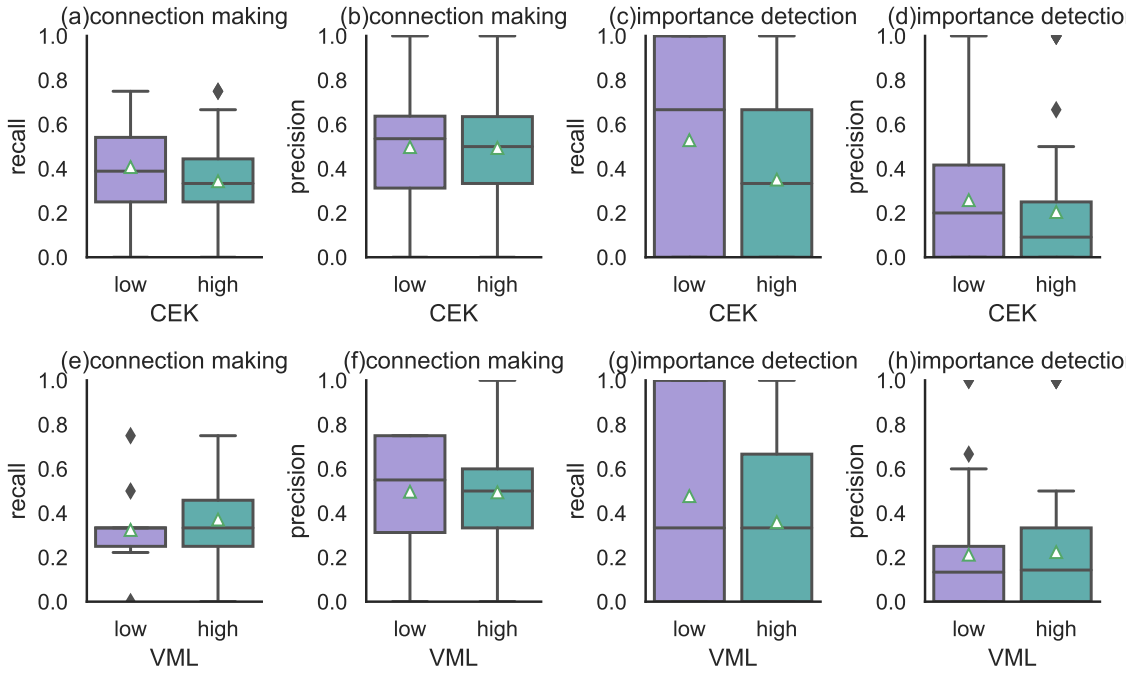

Figure 8: Distribution of recall and precision for connection-making and importance detection divided into low/high CEK users (top), and low/high VML users (bottom).

author coded the responses accordingly and the others checked the final codes. Participants annotated 250 connections and 305 important statements. Table 2 shows the coding scheme we developed for each annotation task with sample responses matching the code. There are six codes for connection-making and five codes for importance detection (including one in each for empty responses). Here, the codes in the connection-making task offer potential answers to the 5W1H questions (Who, What, Where, When, Why, and How). For example, the code "Person" answers the *who* aspect of the event, while the code "Information" answers a combination of what, why, and how questions. For importance detection, users sometimes claimed that a statement was important without explaining the reason for this assessment. In other cases, users mentioned that a statement was important because it clarified or elaborated on existing statements, or because it provided an account from a missing perspective. Figure 9 shows the count of each rationale in terms of the developed codes, divided into correct and incorrect annotations. For connection-making (Figure 9(a)), we found that a majority of users identified similarities when the same information was presented in both articles, followed by mentions of the same person. For importance detection(Figure 9(b)), many responses were coded into the clarification and elaboration categories, followed by the

	Code	Definition	Example Response
Connection	Empty (20%)	Empty or N/A response	
	Person (19%)	Mentions that both statements refer to one or more persons involved in an event (not including quotes)	*They are similar because they both mention* ***the owners of the truck***
	Location (1%)	Mentions that both statements refer to the same location where an event occurred	*This excerpt shows where the truck was found and both gave* ***an identical location*** *...*
	Date (2%)	Mentions that both statements refer to a single date when an event occurred or will occur	*Both statements note that the ban will take effect on* ***August 25th***
	Quote (9%)	Mentions that both statements reference either the same quote or different quotes from the same person	*They are similar because both highlight* ***a quote from Becerra (HHS Secretary)*** *insisting ...*
	Information (48%)	Mentions that both statements contain the same information describing the what, why, or how of the event	*Similar because [both] discuss* ***Medical Treatment and Labor Act***
Importance	Empty (29%)	Empty or N/A response	
	Important (15%)	Mentions that a statement is important without providing a reason	*[because it is an]* ***important part*** *of the news*
	Clarification (43%)	Mentions that a statement clarifies or elaborates on other statements	*The* ***statement in the other article*** *from Becerra (HHS Secretary) is* ***confusing****.*
	Missing (8%)	Mentions that a statement presents a perspective missing from the other article	***No statement from Lawrence*** *in the other article*
	Factual (4%)	Mentions that a statement is factual and not an opinion	***Facts here.*** *It isn't opinion being interjected into a news story.*

Table 2: Coding scheme for annotation rationales.

empty response category. Figure 9 suggests that the rationales are not distributed proportionally for correct and incorrect annotations. Notably, users appear more likely to make errors in certain cases. For example, for importance detection (Figure 9(b)), the ratio of correct and incorrect "important" annotations shows that these annotations are more likely to be mistaken than others. Although we performed regression on the codes to differentiate correct and incorrect annotations, the model effect sizes were very low for both models ($R^2 < 0.05$).

Since differentiating correct and incorrect annotations by codes did not work well, we fit two models to predict incorrect annotations (false positives) for both the connection-making and importance detection tasks using the top 50 TF-IDF [8] text features from users' responses on the rationals. Figure 10 shows the text features with significance for this analysis. Examining the words with significant coefficients, we can see that some words are generic (e.g., "quote", "similar", "context"), while others are article-specific (e.g., "Garland" (the current attorney general), "lawsuit", "smuggling"). These differences suggest that such generic words in rationales can be used across articles to differentiate false positives from true positives, while specific words may not be usable. We discuss these results further in section 6.1.

5.1.6 Comparative Perception between Article Pairs. After the annotation task, in addition to asking about credibility and quality perception, we asked users what they noticed when comparing the two articles ("Comparing the two articles, what else did you notice about how each portrayed the issue?"). Analyzing the responses, we found five themes, summarized in Table 3. Notably, more than one fourth of the participants remarked on informational placement or depth (16/62), perspectives and biases (22/62) and factuality/opinions (16/22). Some also noticed empathetic news reporting (5/62) and the use of inflammatory language (3/62).

5.1.7 Perception of the Tool. Apart from the task-specific questions, we also asked annotators about their perceptions of the *NewsComp* tool. Overall, perceptions of the tool were split among positive (28/62), neutral (21/62), and negative (15/62) sentiment. The reasons behind negative sentiment included the lengthy nature of the task (3/15), difficulty in performing annotation (8/15), and confusion regarding the instructions (4/15). While it may have been hard in the beginning, users quickly learned how to use the tool (*"It was a bit confusing to learn how to use the tool, but it was easy to use once I played around with it." - U1*). Improvements to the tool design could potentially address these issues. For instance, during connection-making, a search tool could assist users with finding similar statements quickly. Users also suggested improvements such as allowing them to set the weight of the connected lines, change the colors of lines, and see how others annotated a statement.

5.2 RQ2: How does comparative news annotation affect users' perceptions of credibility and news quality?

To answer this question, we performed a two-way ANOVA on the response variables, credibility scores, and quality scores. To calculate each score, we first summed item scores for the questionnaire on each score (credibility and quality) and standardized them on a [0, 1] interval. Then, we performed the two-way ANOVA on the group and article interaction. Figure 11 shows the result of this analysis. We did not find any significant difference in quality perceptions. We also performed a one-sample t-test on users' quality perception responses against those of the experts. Though we did find some similarity for the abortion articles (Figure 11(b)), in the case of the immigration articles (11(a)); that is, the high-contrast pair, the difference between the users' and experts' ratings was

[8] TF-IDF stands for term frequency–inverse document frequency, a statistic representing how important a word is to a document in a collection of documents

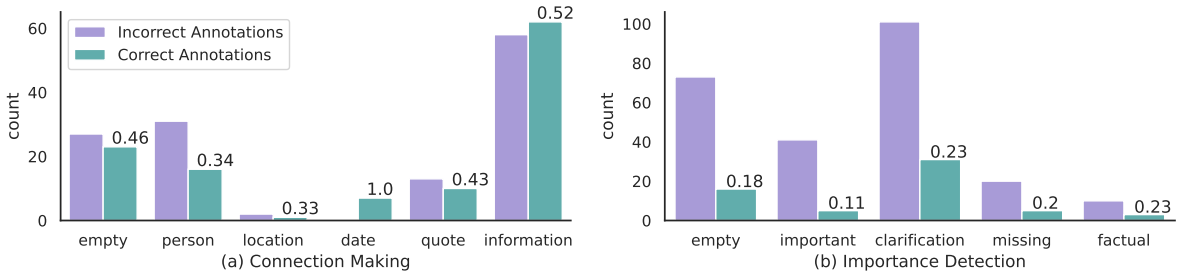

Figure 9: Annotation counts by coded rationales divided into correct and incorrect annotations. The numbers over the bars represent the ratio of correct to incorrect annotations within each code.

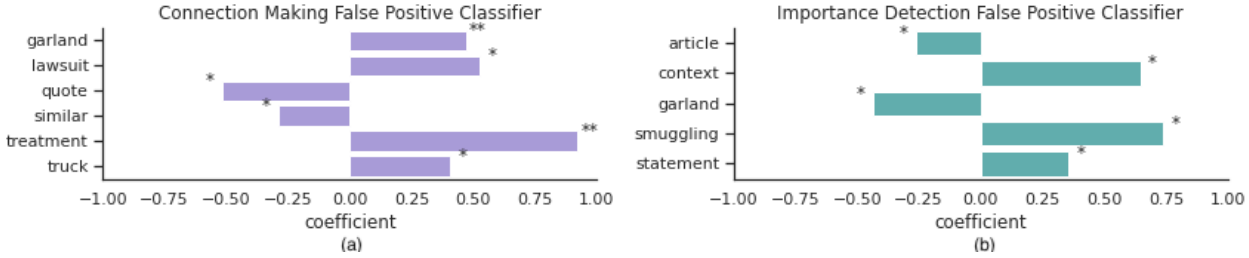

Figure 10: False positive detection with OLS using the top 50 TF-IDF words in users' responses. Here, we listed only words with significant coefficients. For example, when users' mentioned "quote" in a rationale, the annotation was less likely to be erroneous. On the other hand, when users mentioned the general nature of the event ("lawsuit" in this example), the annotation was more likely to be erroneous. The model effect sizes (R^2) were 0.34 and 0.22, respectively.

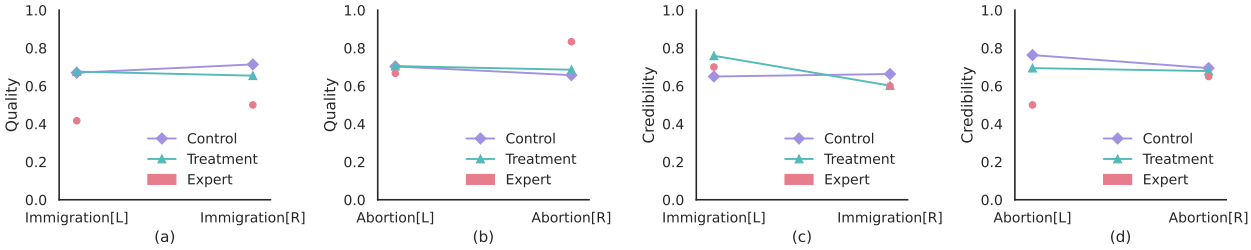

Figure 11: Interaction effects of groups and articles. We only found a marginal interaction effect (p=0.052) for credibility score on articles regarding immigration (c).

Theme(n)	Example Response
Perspectives and Biases (22)	They were taking **different sides** of the equation and putting forward different thought processes
Information Placement or Depth (14)	The right article was **less descriptive** and focused more on the restrictions and not the case. It was definitely telling the story from one point of view. The article on the left was very informative and unbiased
Factuality or Opinions (16)	Article B provided responses from the Idaho government, whereas Article A did not include **commentary** from Idaho, but instead from Texas which did not seem relevant
Empathetic Reporting (5)	There were more **humane aspect** in the article on the left
Inflammatory Language (3)	Article B seemed to be making the issue out to be more **controversial** by going back and forth between perspectives more frequently

Table 3: Themes in users' responses to a question asking what they noticed about the two articles overall. Note that while an example response may belong to multiple themes, only the portion relevant to the listed theme is presented in bold.

significant. This result suggests that neither group performed well on the quality question, especially for the high-contrast article pair.

Next, we analyzed the interaction of credibility with the articles. In Figure 11(c), our analysis found that the interaction between the experimental group and the article is marginally significant for the articles on immigration with a close to moderate effect size ($F(1) = 3.83$, $p = 0.052$, Cohen's $f = 0.22$). We did not find any significant effect for articles on abortion. Note that when we fit a mixed-effects regression model on the same data additionally considering repeated measure, we found significant interaction effects on the high-contrast articles (see Appendix D). Moreover, when we look at the articles on abortion, despite the experts not performing comparative annotation, their perceptions of one article (from a left-leaning source) were significantly lower than those of the other (from a right-leaning source). During our discussion, we found that experts used certain criteria to arrive at this assessment [9]. On the other hand, the comparative annotation task may have given users the impression that both articles are highly similar (leading to similar ratings) and discouraged them from examining other differences too closely. This result suggests that there is room for improvement in task designs for comparative annotation. We discuss further implications in section 6.2.

6 DISCUSSION

Through our experiment, we found that users generally perform poorly in annotation tasks for finding similar statements and identifying important statements among statements with no similarity between two news articles. However, they have better precision in finding similarities than in identifying important statements. We found that when a high number of users find two statements similar, such annotations have a high chance of coinciding with experts' annotations. Furthermore, we found that users with low current event knowledge may perform better annotations. Analyzing users' rationales behind the annotations, we found several reasons for finding similarities (e.g., mentions of the same person or information) and identifying important statements among statements with no similarity (e.g., statements clarifying something or providing a missing perspective). Furthermore, we found that certain words have significant power in differentiating false and true positives. After annotation, users also mentioned noticing differences in how article pairs represented things, such as perspectives, information placement, information depth, and facts/opinions. In RQ2, we found that annotation tasks may have limited effects on users' perception of news credibility for high-contrast news articles. Below, we discuss the implications of these results.

6.1 RQ1: Annotation Performance

Our results indicate that although users perform poorly in general, their performance varies across annotation task types and article pairs, depending on the degree of contrast between the articles in the pair. Compared to the experts, average users seem to find fewer connections and consider more statements worthy of inclusion in the other article in a pair. This difference shows how experts and users differ when reading two news articles. This difference could stem from analytical capabilities; that is, perhaps finding similarities and differences is a task that requires particular expertise relating to news. Or, perhaps low knowledge users are more attentive to the articles. Another reason behind the difference could be users' preconceived biases regarding the media in general [36]. Such perceptions may have influenced users to see fewer connections and more differences. Therefore, one direction for future research may be to examine the rationales behind identified differences by varying the task complexity and user characteristics. For example, we can ask readers to perform small subtasks, such as identifying sources or labeling word choices [29] to test if these influence their perceptions of bias.

Another noteworthy aspect here is that our participant group came from Facebook advertising, not from platforms like Amazon Mechanical Turk or Upwork typically used for crowdsourcing. This suggests that users outside of crowd work platforms can also perform effectively on crowdsourced tasks. In the future, research could look into how well workers from crowdsourcing platforms and other sources compare in terms of performance.

Our result indicates that crowd annotation in subjective tasks is to a little extent affected by users' backgrounds—in our case, their news expertise, aligning with prior works [22, 33, 64]. Therefore, we can train users using their news expertise as a targeting criterion. Since user performance also varies by task, helping users improve quality on a particular task area might also help. Furthermore, designers can support users in annotation tasks through various interventions. For example, since our TF-IDF models identified some generic words that can distinguish false positives, such data could also be used to provide users with feedback or warnings to improve annotation quality.

We also discovered the effects of comparative annotation on users' overall impressions, leading to differences in perceptions of viewpoints, information attributes (placement, depth, and factuality/opinion), and emotional attributes (empathetic vs. inflammatory language). These differences could impact users' attitudes towards an article. For example, between informational and emotional attributes, understanding which differences impact perceptions of trustworthiness could be one future avenue of work.

To improve users' performance, one option could be through collaboration—learning from each other through social annotation [32]. Indeed, prior research shows that when people see others' annotations, it can persuade them to take certain actions, such as changing ratings when faced with opposing social opinions [18]. Furthermore, research suggests that displaying social information about the annotator, such as their level of expertise, can persuade and build trust [28]. Incorporating social information on other annotators during collaboration may improve learning. In a collaborative environment, we still need to handle annotator bias, since bias from a small group of users could propagate to a larger pool of users and cause unexpected effects. Therefore, examining such collaborative annotations and their impact on user performance is another potential direction of research.

[9] These articles talk about the DOJ's challenge against an abortion restriction law in Idaho. Here, the experts mentioned that the article from the left-leaning source did not mention an opposing perspective, such as that of Idaho's government. Note, however, that the article did cite the state attorney general of Texas, who supports abortion restrictions.

6.2 RQ2: The Effect of Engaging through Annotation

While it may not be true in all cases, our results indicate that in cases where there is significant contrast between a pair of news articles, users might be somewhat influenced by comparative annotation tasks. Our work can inform related future works on improving engagement with plural viewpoints through annotations [82]. Compared to works that show visualizing biases alone does not improve perception of bias [68], our work suggests that additional engagement could be helpful. The effect we see may stem from complex information processing that occurs when users engage with competing messages [8]. Since our result did not reveal any universally significant effect, it does point towards the idea that only certain perceptions are affected. Therefore, one direction for future research could include looking into different perception paradigms to further identify the limits of such effects.

Even though we found limited effects on perceptions of credibility, this does not necessarily limit the applicability of comparative annotation. As we saw in section 5.1.6, a user's understanding of the differences between articles could have other impacts. Besides, repeatedly annotating two sources can create certain impressions in the long run. For example, seeing repeated differences in the use of factual statements or depth in reporting could affect users' perceptions of credibility. Furthermore, we can ask whether crowdworkers from such platforms as Mechanical Turk would also remain unaffected by the annotation task. In any case, *NewsComp* could be purposefully deployed to crowdworkers while also providing general users the option to perform annotation. In such a case, users desiring more ways to engage and community fact-checkers might be more attracted to it. Regardless, there are further uses for the annotated data.

6.3 Applications of Annotated Data

Our annotated data could be used in various ways. For instance, it could be incorporated into a system that combines information from multiple sources to provide a holistic view of an event. It is not uncommon in online spaces for information overload to make it harder for people to efficiently consume information [1, 25]. A holistic view could particularly be useful to users in such a scenario, especially for sensemaking purposes [80, 81]. Such a system would mimic strategies humans typically employ to consume information efficiently, such as organizing information by tagging, sorting, and indexing [12]. We could further build upon this by introducing mechanisms for peer-curated information [63] A second potential use of the annotated data would be in training algorithmic models to generate better annotations, which could in turn be used to better curate information for readers. As we saw in our initial think-aloud interviews, people find the accuracy of existing SOTA ML systems insufficient for finding semantic similarities and differences. The annotated data could help to improve such algorithms. A third use of the data is for fact-checkers. Fact-checkers can use annotated information to validate claims through the use of linked statements from multiple sources. They can also use such links to trace the origins of statements. Perhaps a portion of these fact-checking tasks could be delegated to automatic fact-checking algorithms. Furthermore, even crowd fact-checkers (e.g., from Twitter's BirdWatch) could use the annotated data to validate claims.

6.4 Merging Articles Into One and Testing Effects

One of the goals of this research on comparative annotation was to combine diverse perspectives into one. With our annotated data, crowd tasks can be designed to accomplish such merging of perspectives. However, there are some considerations for task design in this process. Take similar statements as an example—if two statements are very similar, a task could ask workers to choose one or the other. On the other hand, if selecting one statement necessarily results in the omission of important information from the other, then the crowd task may also require editing. In the case of merging important disparate statements, as noted in our think-aloud interviews, one important consideration is checking whether a statement fits the narrative of the current article. We can either include this criteria or discard it, which would lead to differences in the outcome, (i.e, the merged article). Taking a step further, this merging process can be extended from article pairs to larger groups of articles. Merging larger groups of articles would require a multistep selection, voting, and reconciliation process. Finally, while we found that the effect of annotation on perception was limited, could merged articles affect users' perceptions of an event differently than articles from a single source? Future research answering such a question would generate new knowledge regarding the utility of comparative annotation.

6.5 Implications for Comparative Annotation Task Design

Motivated by users' perceptions of *NewsComp*, we identified two major issues in the comparative annotation task: the lengthy nature of the task and difficulty in performing the task. Since one of our research questions focused on the impact of performing annotation, our experiment was designed so that users performed a complete annotation task on two articles before responding to the questions. If the annotation impact is not of interest, both of these issues can be resolved. First, we can modularize the tasks by breaking them into small pieces (e.g., making connections between two paragraphs instead of two entire articles), in line with prior research on devising microtasks for complex work [41, 42]. However, could such modularization cause a backfire effect? For example, if an annotator is assigned two dissimilar paragraphs from a pair of broadly very similar articles, could that skew their perception? This is one potential consideration for designing small, modular tasks.

Second, even if the task is not divided into smaller components, there are other options for improvement. For instance, finding similarities can be made easier through the addition of such features as automatic suggestion and filtering. Here, algorithms can provide automatic suggestions and users can search by keyword to limit the options to choose from.

Third, tasks can be divided for co-annotation to reduce difficulty. For example, one annotator might suggest connections while another annotator votes on the suggestions. In addition, as a tutorial, displaying example annotations from other users could also help

resolve some concerns. However, the examples need to be generic enough not to significantly impact users' own future annotations.

Fourth, apart from issues related to task difficulty, there is another issue that will need attention in the future. In the think-aloud interviews and the deployment of *NewsComp*, we discovered some disconnects in the rationales provided for annotations. Particularly, for connection-making, we did not see use of thematic similarity during deployment. Perhaps regular users may need nudges to identify high-level thematic similarity. Overall, there are ample opportunities for improving the tasks in *NewsComp*.

6.6 Limitations

Our work is not without limitations. First, our study was conducted in a controlled environment which may differ from that of a user's typical news reading sessions. Therefore, some of the observed effects could have been products of the environment. However, we emulated a typical news consumption environment as best we could, from content selection to the design of the interface. Therefore, our results offer some validity that future works can build on. Second, since our study procedure involved signing up for the study and voluntary completion criteria, some self-selection bias exists, similar to other research in this domain. However, we did advertise on Facebook to find users organically instead of recruiting users from crowd survey platforms, which provided some benefits to the selection process. Third, we conducted the study within a US-centric context, limiting its generalizability. Future research could resolve such issues by conducting similar research with a larger country pool. Finally, the task in the study was a bit lengthy (20 min) relative to tasks that crowd workers typically perform. Though the articles in the study were not excessively long (11-16 sentences), this could still have affected task quality. Future work can further examine how performance varies by task complexity. Overall, our work has certain merits that require further exploration in the future.

7 CONCLUSION

In this work, we examined how well users perform on a comparative news annotation task featuring a pair of news articles, and how the annotation task affects users' perceptions of the articles. Comparing our users' annotations against those of experts, we found that users generally performed very poorly on the annotation task. However, certain information, such as the number of users who made a given annotation and users' rationales behind annotations, can be used to detect incorrect annotations. Furthermore, we found some marginal changes in users' credibility perceptions for certain news articles after completing the annotation process. Our work has implications for designing future comparative annotation systems.

ACKNOWLEDGMENTS

This paper would not be possible without our study participants. We also appreciate the valuable feedback we received from the anonymous reviewers, the members of the EchoLab at Virginia Tech, and the members of the SCALE Lab at the University of Washington, and Natasha Noy from Google Research. Bhuiyan and Mitra were generously supported by National Science Foundation grant #2128642.

REFERENCES

[1] Linda Aldoory and Mark A Van Dyke. 2006. The roles of perceived "shared" involvement and information overload in understanding how audiences make meaning of news about bioterrorism. *Journalism & Mass Communication Quarterly* 83, 2 (2006), 346–361.
[2] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances* 7, 36 (2021), eabf4393.
[3] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
[4] W Brian Arthur. 1994. Inductive reasoning and bounded rationality. *The American economic review* 84, 2 (1994), 406–411.
[5] Mevan Babakar. 2018. Crowdsourced Factchecking.
[6] Joseph O Baker and Amy E Edmonds. 2021. Immigration, presidential politics, and partisan polarization among the American public, 1992–2018. *Sociological Spectrum* 41, 4 (2021), 287–303.
[7] Scott Bateman, Rosta Farzan, Peter Brusilovsky, and Gord McCalla. 2006. OATS: The open annotation and tagging system. *Proceedings of I2LOR* (2006).
[8] W Lance Bennett. 1981. Perception and cognition. In *The handbook of political behavior*. Springer, 69–193.
[9] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.
[10] Ivar Bråten and Helge I Strømsø. 2011. Measuring strategic processing when students read multiple texts. *Metacognition and Learning* 6, 2 (2011), 111–130.
[11] David V Budescu and Eva Chen. 2015. Identifying expertise to extract the wisdom of crowds. *Management Science* 61, 2 (2015), 267–280.
[12] Liz Carver and Murray Turoff. 2007. Human-computer interaction: the human and computer as a team in emergency management information systems. *Commun. ACM* 50, 3 (2007), 33–38.
[13] Pew Reseach Center. 2014. Political polarization in the american public. *Ann Rev Polit Sci* (2014).
[14] Sidharth Chhabra and Paul Resnick. 2013. Does clustered presentation lead readers to diverse selections? In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1689–1694.
[15] Michelene Chi. 1992. Conceptual change within and across ontological categories: Examples from learning and discovery in science. (1992).
[16] Sujin Choi. 2015. The two-step flow of communication in Twitter-based public forums. *Social science computer review* 33, 6 (2015), 696–711.
[17] Denzil Correa, Leandro Araújo Silva, Mainack Mondal, Fabrício Benevenuto, and Krishna P Gummadi. 2015. The many shades of anonymity: Characterizing anonymous social media content. In *Ninth International AAAI Conference on Web and Social Media*.
[18] Dan Cosley, Shyong K Lam, Istvan Albert, Joseph A Konstan, and John Riedl. 2003. Is seeing believing? How recommender system interfaces affect users' opinions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 585–592.
[19] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Scientific reports* 7, 1 (2017), 1–9.
[20] Stefano DellaVigna and Ethan Kaplan. 2007. The Fox News effect: Media bias and voting. *The Quarterly Journal of Economics* 122, 3 (2007), 1187–1234.
[21] Peter M DeMarzo, Dimitri Vayanos, and Jeffrey Zwiebel. 2003. Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics* 118, 3 (2003), 909–968.
[22] Tim Draws, David La Barbera, Michael Soprano, Kevin Roitero, Davide Ceolin, Alessandro Checco, and Stefano Mizzaro. 2022. The Effects of Crowd Worker Biases in Fact-Checking Tasks. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 2114–2124.
[23] James N Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *The Journal of Politics* 67, 4 (2005), 1030–1049.
[24] Robert M Entman. 2007. Framing bias: Media in the distribution of power. *Journal of communication* 57, 1 (2007), 163–173.
[25] Martin J Eppler and Jeanne Mengis. 2008. The concept of information overload-a review of literature from organization science, accounting, marketing, mis, and related disciplines (2004). *Kommunikationsmanagement im Wandel* (2008), 271–305.
[26] William P Eveland Jr and Sharon Dunwoody. 2001. User control and structural isomorphism or disorientation and cognitive load? Learning from the Web versus print. *Communication research* 28, 1 (2001), 48–78.
[27] Morris P Fiorina, Samuel J Abrams, et al. 2008. Political polarization in the American public. *ANNUAL REVIEW OF POLITICAL SCIENCE-PALO ALTO-* 11 (2008), 563.

[28] Jennifer Golbeck and Kenneth R Fleischmann. 2010. Trust in social Q&A: the impact of text and photo cues of expertise. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–10.

[29] 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, 4 (2019), 391–415.

[30] Naeemul Hassan, Mohammad Yousuf, Mahfuzul Haque, Javier A Suarez Rivas, and Md Khadimul Islam. 2017. Towards A Sustainable Model for Fact-checking Platforms: Examining the Roles of Automation, Crowds and Professionals. https://doi.org/10.1145/3308560.3316734

[31] Edward S Herman and Noam Chomsky. 2010. *Manufacturing consent: The political economy of the mass media.* Random House.

[32] Janette R. Hill, Liyan Song, and Richard E. West. 2009. Social learning theory and web-based learning environments: A review of research and discussion of implications. *International Journal of Phytoremediation* 21, 1 (2009), 88–103.

[33] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[34] Reuters Institute. 2022. Overview and key findings of the 2022 Digital News Report | Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2022/dnr-executive-summary. (Accessed on 09/11/2022).

[35] Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment.* Oxford University Press.

[36] Mark Jurkowitz, Amy Mitchell, Elisa Shearer, and Mason Walker. 2020. U.S. Media Polarization and the 2020 Election: A Nation Divided | Pew Research Center. https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/. (Accessed on 09/15/2022).

[37] Daniel Kahneman and Amos Tversky. 2013. Choices, values, and frames. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 269–278.

[38] Kenneth A Kavale. 1980. The reasoning abilities of normal and learning disabled readers on measures of reading comprehension. *Learning Disability Quarterly* 3, 4 (1980), 34–45.

[39] Ricardo Kawase, Eelco Herder, and Wolfgang Nejdl. 2009. A comparison of paper-based and online annotations in the workplace. In *European Conference on Technology Enhanced Learning*. Springer, 240–253.

[40] Damon Kiesow, Shuhua Zhou, and Lei Guo. 2021. Affordances for Sense-Making: Exploring Their Availability for Users of Online News Sites. *Digital Journalism* 0, 0 (2021), 1–20. https://doi.org/10.1080/21670811.2021.1989316

[41] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[42] Aniket Kittur, Boris Smus, Susheel Khamkar, and Robert E Kraut. 2011. Crowdforge: Crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 43–52.

[43] Joshua Klayman. 1995. Varieties of confirmation bias. *Psychology of learning and motivation* 32 (1995), 385–418.

[44] Steven Kull, Clay Ramsay, and Evan Lewis. 2003. Misperceptions, the media, and the Iraq war. *Political science quarterly* 118, 4 (2003), 569–598.

[45] Paul F Lazarsfeld, Bernard Berelson, and Hazel Gaudet. 1968. The people's choice. In *The people's choice*. Columbia University Press.

[46] David G Lebow and Dale W Lick. 2005. HyLighter: An effective interactive annotation innovation for distance education. In *20th Annual Conference on Distance Teaching and Learning*. 1–5.

[47] Adam Maksl, Seth Ashley, and Stephanie Craft. 2015. Measuring news media literacy. *Journal of Media Literacy Education* 6, 3 (2015), 29–45.

[48] Miriam J Metzger, Ethan H Hartsell, and Andrew J Flanagin. 2015. Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. *Communication Research* (2015), 0093650215613136.

[49] Philip Meyer. 1988. Defining and measuring credibility of newspapers: Developing an index. *Journalism quarterly* 65, 3 (1988), 567–574.

[50] Tanushree Mitra and Eric Gilbert. 2015. CREDBANK: A Large-Scale Social Media Corpus with Associated Credibility Annotations. In *Proc. ICWSM'15*.

[51] Sendhil Mullainathan and Andrei Shleifer. 2002. Media bias.

[52] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of The International AAAI Conference on Web and Social Media*, Vol. 7. 419–428.

[53] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.

[54] Joseph Napolitan. 1972. *The election game and how to win it.* Doubleday.

[55] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.

[56] Elena Novak, Rim Razzouk, and Tristan E. Johnson. 2012. The educational use of social annotation tools in higher education: A literature review. *Internet and Higher Education* 15, 1 (2012), 39–49. https://doi.org/10.1016/j.iheduc.2011.09.002

[57] Eli Pariser. 2011. *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin.

[58] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 443–452.

[59] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2012. A computational framework for media bias mitigation. *ACM Transactions on Interactive Intelligent Systems* 2, 2 (2012). https://doi.org/10.1145/2209310.2209311

[60] Souneil Park, Minsam Ko, Jungwoo Kim, Ho-jin Choi, and Junehwa Song. 2011. NewsCube2 . 0 : An Exploratory Design of a Social News Website for Media Bias Mitigation. *Workshop on Social Recommender Systems* (2011), 1–5. https://pdfs.semanticscholar.org/b87b/f0986b2e9fe34a22ed0c19cfd32ed06857d0.pdf

[61] Souneil Park, Kyung Soon Lee, and Junehwa Song. 2011. Contrasting opposing views of news articles on contentious issues. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 1 (2011), 340–349.

[62] Val Pipps, Heather Walter, Kathleen Endres, and Patrick Tabatcher. 2009. Information recall of Internet news: Does design make a difference? A pilot study. *Journal of Magazine Media* 11, 1 (2009), 1–20.

[63] Odette Pollar. 2003. *Surviving information overload: how to find, filter, and focus on what's important.* Thomson Crisp Learning.

[64] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? The Effects of Judgment Scale and Assessor's Background. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 439–448.

[65] Dietram A Scheufele. 2000. Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. *Mass communication & society* 3, 2-3 (2000), 297–316.

[66] Slate. 2013. How people read online: Why you won't finish this article. https://slate.com/technology/2013/06/how-people-read-online-why-you-wont-finish-this-article.html. (Accessed on 09/11/2022).

[67] Oren Soffer. 2021. Algorithmic personalization and the two-step flow of communication. *Communication Theory* 31, 3 (2021), 297–315.

[68] Timo Spinde, Felix Hamborg, Karsten Donnay, Angelica Becerra, and Bela Gipp. 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*. 389–392.

[69] Steven A Stahl, Cynthia R Hynd, Bruce K Britton, Mary M McNish, and Dennis Bosquet. 1996. What happens when students read multiple source documents in history? *Reading Research Quarterly* 31, 4 (1996), 430–456.

[70] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of communication* 60, 3 (2010), 556–576.

[71] Natalie Jomini Stroud. 2011. *Niche news: The politics of news choice.* Oxford University Press on Demand.

[72] Cass R Sunstein. 1999. The law of group polarization. *University of Chicago Law School, John M. Olin Law & Economics Working Paper* 91 (1999).

[73] Cass R Sunstein. 2009. http://Republic.com 2.0.

[74] James Surowiecki. 2005. *The wisdom of crowds.* Anchor.

[75] David Tewksbury and Scott L Althaus. 2000. Differences in knowledge acquisition among readers of the paper and online versions of a national newspaper. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 457–479.

[76] Juliane Urban and Wolfgang Schweiger. 2014. News quality from the recipients' perspective: Investigating recipients' ability to judge the normative quality of news. *Journalism Studies* 15, 6 (2014), 821–840.

[77] Stella Vosniadou and William F Brewer. 1987. Theories of knowledge restructuring in development. *Review of educational research* 57, 1 (1987), 51–67.

[78] Stella Vosniadou and Andrew Ortony. 1989. *Similarity and analogical reasoning.* Cambridge University Press.

[79] Emily Vraga, Melissa Tully, John E Kotcher, Anne-Bennett Smithson, and Melissa Broeckelman-Post. 2015. A Multi-Dimensional Approach to Measuring News Media Literacy. *Journal of Media Literacy Education* 7, 3 (2015), 41–53.

[80] Karl E Weick. 1995. *Sensemaking in organizations.* Vol. 3. Sage.

[81] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421.

[82] Gavin Wood, Kiel Long, Tom Feltwell, Scarlett Rowland, Phillip Brooker, Jamie Mahoney, John Vines, Julie Barnett, and Shaun Lawson. 2018. Rethinking engagement with online news through social and visual co-annotation. *Conference on Human Factors in Computing Systems - Proceedings* 2018-April (2018), 1–12. https://doi.org/10.1145/3173574.3174150

[83] John R Zaller et al. 1992. *The nature and origins of mass opinion.* Cambridge university press.

[84] Pengyi Zhang and Dagobert Soergel. 2020. Cognitive mechanisms in sensemaking: A qualitative user study. *Journal of the Association for Information Science and Technology* 71, 2 (2020), 158–171. https://doi.org/10.1002/asi.24221

	Imp. Recall (M1)		Imp. Precision (M2)		Conn. Recall(M3)		Conn. Precision(M4)	
	β	std. err.	β	std. err.	β	std. err.	β	std. err.
(Intercept)	0.32***	0.08	0.21***	0.04	0.34***	0.05	0.37***	0.05
CEK[Low]	0.35**	0.13	0.05	0.07	0.07	0.06	0.02	0.07
VML[Low]	0.18	0.12	-0.02	0.06	-0.04	0.06	0.02	0.07
	R^2=0.12		R^2=0.01		R^2=0.04		R^2=0.19	
N_{obs}=62							*p<0.05, **p<0.01, ***p<0.001	

Table 4: Linear models of recall and precision for connection-making and importance detection with user characteristics as predictors.

	Qual(M5)		*Cred*(M6)	
	β	std. err.	β	std. err.
(Intercept)	0.67***	0.05	0.65***	0.05
Group[Treat.]	0.01	0.07	0.11	0.07
Article[Abortion(R)]	0.04	0.07	0.01	0.07
Article[Immigration(L)]	0.03	0.07	0.11	0.07
Article[Immigration(R)]	-0.01	0.05	0.04	0.05
Group [Treat.] * Article[Abortion(R)]	-0.07	0.09	-0.17*	0.08
Group [Treat.] * [Immigration(L)]	0.00	0.09	-0.18	0.09
Group [Treat.] * Article[Immigration(R)]	0.02	0.08	-0.12	0.08
	R^2 = 0.43		R^2 = 0.48	
N_{user} = 109, $N_{article}$ = 4, N_{obs} = 218			*p<0.05, **p<0.01, ***p<0.001	

Table 5: Mixed-effects regression on quality and credibility score using the interaction of experimental condition and articles.

A THINKALOUD INTERVIEWS QUESTIONNAIRES

- What are the viewpoints expressed in each article? How would you compare the viewpoints between the articles?
- How would you compare the numbers of actors reported in each article?
- How would you compare each article providing complete information about what happened/where/when/who was involved?
- How would you compare the analytical quality in each article? Do they provide information on causes, consequences, evaluations and claims of/from the event?
- How transparent are the authors for each article about their sources (e.g,name, function, circumstances of quote)?
- How would you compare the comprehensibility of the article pair (e.g., simplicity in terms/phasing, conciseness, coherence)?
- How would you compare impartiality in content presentation between the articles (balanced viewpoints and actors, article
- author personally evaluating/judging the reported situation)?
- How would you compare ethical standards (e.g., discriminating any party involved, neutral phrasing) between the reports?
- Whose perspective this article represents more than others? Is there any particular group/party/side that the article focus to represent compared to the other article?

B ARTICLES USED IN THE DEPLOYMENT

- E_1L: Immigration (Left)
- E_1R: Immigration (Right)
- E_2L: Abortion (Left)
- E_2R: Abortion (Right)

C EFFECT OF USER CHARACTERISTICS

We modeled user characteristics to predict precision and recall in annotation tasks, shown in table 4. We accounted for several factors in these models, including users' demographic characteristics (age, gender, education, and political affiliation) and news expertise metrics (CEK and VML).

D RQ2: MIXED-EFFECTS MODELS

Besides ANOVA, we also performed a series of mixed-effects regression model on users' quality and credibility perception using experimental variables, in Table 5. Similar to ANOVA results, we found significant interaction effect on credibility only for high contrast article.