The Impact of Highlighting Subjective Language on Perceived News Trustworthiness

Mohammad Shokri[1] **Vivek Sharma**[1] **Emily Klapper**[2]
Elena Filatova[3] **Shweta Jain**[4] **Sarah Ita Levitan**[2]

[1]Graduate Center, CUNY [2]Hunter College, CUNY
[3]City Tech, CUNY [4]John Jay College, CUNY

Abstract

The rise of misinformation and opinionated articles has made understanding how misleading or biased content influences readers an increasingly important problem. While most prior work focuses on detecting misinformation or deceptive language in real time, far less attention has been paid to how such content is perceived by readers, which is an essential component of misinformation's effectiveness. In this study, we examine whether highlighting subjective sentences in news articles affects perceived trustworthiness. Using a controlled user experiment and 1,334 article–participant ratings, we find that highlighting subjective content produces a modest yet statistically significant decrease in trust, with substantial variation across articles and participants. To explain this variation, we model trust change after highlighting subjective language as a function of article-level linguistic features and reader-level attitudes. Our findings suggest that readers' reactions to highlighted subjective language are driven primarily by characteristics of the text itself, and that making subjective language explicit plays a role in shaping perceived trustworthiness.

1 Introduction

Trust in media has reached an unprecedented low,[1] even as the internet enables news to circulate with unprecedented speed and reach. This combination has created fertile ground for misinformation and disinformation, which continue to produce serious societal and public health consequences (Vosoughi et al., 2018; Vasist et al., 2024; Ruohonen, 2024). The impact of misleading content is especially pronounced when individuals find it believable.

Believability is closely tied to perceptions of trust, including trust in the author, the outlet, or the narrative being presented. Prior work shows that readers often rely on textual cues to assess whether an article appears balanced, fair, or biased, and these trust judgments strongly shape whether they accept or reject the information (Rodrigo-Ginés et al., 2024; Gabriel et al., 2022).

Understanding how people respond to news as they encounter it is therefore a critical (Gabriel et al., 2022) and understudied challenge. Readers do not process all information uniformly: the same article may elicit skepticism, acceptance, or indifference depending on how it is framed and perceived. Capturing these responses has practical implications for how media platforms allocate fact-checking resources, prioritize interventions, and curate content for different audiences. In particular, expending effort on content that readers easily dismiss as untrustworthy may be less effective than addressing cases where subtle linguistic cues shape perceptions of credibility and persuasion (Babaei et al., 2021).

While much prior work has focused on computational approaches for identifying fake news or manipulative content, including propaganda (Sprenkamp et al., 2023; Sharma et al., 2025), deception (Velutharambath et al., 2024; Rani et al., 2025), and misleading claims in text (Pérez-Rosas et al., 2018; Volkova and Jang, 2018), these approaches largely operate independently of how content is perceived by readers. Yet, as trust judgments are formed during reading an article, understanding which linguistic signals influence these perceptions is critical for explaining why certain content appears credible or persuasive in the first place.

A central factor in these judgments is subjective language, which refers to the presence of opinions, attitudes, or emotionally charged language in the text (Wiebe et al., 2004). Subjective writing may signal author bias, but readers do not always notice such cues on their own. This creates an impor-

[1] https://news.gallup.com/poll/403166/americans-trust-media-remains-near-record-low.aspx

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 60–72
March 29, 2026 ©2026 Association for Computational Linguistics

tant intersection between misinformation, trust, and subjectivity, as subjective framing can increase the persuasive power of content even when the underlying claim is unsupported, provided that readers interpret it as trustworthy.

In this study, we focus not on determining whether news is factually accurate, but on understanding how readers perceive the trustworthiness of the text they encounter. While detecting subjectivity has been extensively studied in the literature (Wiebe et al., 2004; Antici et al., 2024; Shokri et al., 2024; Elbouanani et al., 2025), far less is known about how making subjectivity explicit affects readers' trust judgments. We address this gap by investigating whether explicitly highlighting subjective language alters perceived trustworthiness. Specifically, we conduct a controlled within-subjects human study in which subjective sentences in news articles are highlighted and compare participants' trust ratings before and after this intervention. Through quantitative analysis of the responses, we examine the extent to which subjectivity influences trust perceptions and identify the conditions under which a simple transparency cue can meaningfully affect readers' evaluations of news content.

2 Related Work

Trust in media has primarily been studied through qualitative approaches, such as interviews and surveys examining how audiences form trust and mistrust in news outlets (Tong, 2024; Garusi and Splendore, 2023; Wenzel et al., 2018). In the psychology literature, trust is commonly defined as a belief in the competence, benevolence, honesty, or predictability of another party, coupled with a willingness to rely on them despite uncertainty or risk (Mayer et al., 1995; McKnight et al., 1998). In the context of online settings and digital media, trust is often defined with respect to the perceived reliability and credibility of information sources and the technology platforms that mediate them. In particular, trust in online news involves not only evaluations of content quality, but also judgments about the integrity of the platform, the authenticity of content producers, and the transparency of information flows (Heuer and Breiter, 2018; Kolo et al., 2022; Lee and Lee, 2023; Cha, 2024).

Relatedly, misinformation and fake news detection have received substantial attention in the NLP community, with recent work focused on large language models as part of the solution (Ma et al., 2024; Han et al., 2025; Tong et al., 2025; Modzelewski et al., 2025). Earlier approaches examined linguistic differences between legitimate and deceptive news using supervised learning and crowd-sourced datasets (Pérez-Rosas et al., 2018; Rubin et al., 2016). Other methods incorporate fact-checking signals or metadata, such as speaker identity or political affiliation, to improve detection of deceptive or hyper-partisan content (Potthast et al., 2018; Wang, 2017; Long et al., 2017). While these efforts have significantly advanced computational techniques for identifying misinformation, they largely frame the problem as a classification task, offering limited insight into how linguistic cues are perceived by readers or how they shape trust judgments during consumption.

Only a limited number of computational studies have explicitly examined how readers form trust judgments about news, rather than focusing primarily on factual accuracy. In a recent study, Bohacek et al. (2023) introduces a dataset of Czech news articles annotated by humans into four trustworthiness categories: trustworthy, partially trustworthy, misleading, and manipulative. The authors fine-tune a range of transformer-based models on this dataset and report that RoBERTa (Liu et al., 2019) achieves the strongest performance, while overall results indicate substantial room for improvement. In a more reader-focused study, Gabriel et al. (2022) model how readers interpret news headlines, including reactions such as distrust and perceived reliability. A key component of their framework is the perceived label, which captures whether readers interpret a headline as real news or misinformation. Through a controlled user study, the authors show that machine-generated explanations such as inferred writer intent can influence readers' trust judgments. In particular, these explanations were found to increase trust in real news and decrease trust in misinformation, with significant correlations observed between gold labels and shifts in perceived trustworthiness for certain models.

In contrast to these approaches, we focus on how a specific and interpretable linguistic signal, namely subjectivity, influences perceived trustworthiness during reading. Rather than predicting trust labels or generating explanations, we study the causal effect of highlighting subjective sentences on readers' trust judgments through a controlled human-subjects experiment.

3 Dataset and Survey

This section describes the datasets used in our study, along with the data collection process, preprocessing steps, and the design of the human-subjects survey.

3.1 Article Selection and Preparation

We draw news articles from the CoAID dataset (Cui and Lee, 2020), a widely used benchmark for misinformation research that aggregates COVID-19–related content from both reliable news outlets and fact-checking sources. All articles used in this study are written in English. The dataset includes true articles collected from cross-verified authoritative outlets, as well as false or misleading articles identified through professional fact-checking organizations.

From the set of true articles, we randomly sampled 100 news articles related to COVID-19 vaccination in the United States. To ensure that articles contained a sufficient amount of subjective language for our intervention, we applied a finetuned RoBERTa-based (Liu et al., 2019) subjectivity classification model to each article and computed the proportion of sentences classified as subjective. We retained articles with a subjectivity ratio greater than 10%, resulting in a final set of 26 articles. Details of the finetuned model and threshold selection are provided in Appendix A.1.

To isolate the effect of textual content on trustworthiness perception, we presented articles in plain text format, removing potentially confounding factors such as author information, news outlet name, images, and advertisements. For each article, we created two versions: an original version and a highlighted version in which sentences annotated as subjective were visually marked using a yellow highlight.

3.2 Annotating Subjective Sentences

To reduce noise from automatic classification and ensure high-quality sentence-level labels, we manually annotated the filtered articles for subjective language. Three members of the research team independently annotated each sentence in the selected articles as either subjective or objective. During this process, annotators did not consistently agree that some articles contained any subjective sentences, and these articles were excluded. As a result, the final set was reduced from 26 to 17 articles. Inter-annotator agreement was assessed using Krippendorff's α (Krippendorff, 2018), which accounts for chance agreement among multiple annotators. The resulting Krippendorff's α agreement score was 0.73, indicating acceptable agreement. Final sentence labels were determined via majority vote among the three annotators.

3.3 Participant Recruitment

We recruited 150 participants through Prolific,[2] an online crowdsourcing platform. Eligibility criteria required participants to be: (1) native English speakers, (2) born in the United States, and (3) between 18 and 65 years of age. These criteria ensured participants had relevant cultural context for evaluating articles about U.S. COVID-19 policy and healthcare issues.

3.4 Survey Procedure

The survey consisted of three phases. First, participants answered a screening question about their stance on vaccines as a public health measure, with response options ranging from "Strongly Supportive" to "Strongly Opposed." This question captured participants' prior beliefs, which could influence their perceptions of article trustworthiness. Second, participants were provided with a definition of trust for the study:

> Trust Definition
>
> For this task, trust refers to how fair and unbiased you believe the author is in presenting the information.

Third, participants evaluated 10 randomly selected articles from the pool of 17. We employed a within-subjects design with counterbalanced presentation order: for each participant, 5 randomly selected articles were presented in original-first order (original version rated, then highlighted version), while the remaining 5 articles were presented in highlighted-first order (highlighted version rated, then original version). All ratings used a 4-point scale (1 = no trust, 2 = somewhat untrustworthy, 3 = somewhat trustworthy, 4 = fully trustworthy).

3.5 Data Quality Control

To ensure data quality, we implemented attention checks and excluded responses from participants who completed the survey in less than 6 minutes, which represented the bottom 5% of completion

[2] https://www.prolific.com

times. This threshold was established to filter out participants who may not have carefully read the articles. After applying this criterion, our final dataset consisted of 136 participants and 1334 paired article ratings.

4 Survey Results

This section examines the relationship between highlighting subjective sentences and perceived trustworthiness. We analyze data from 136 participants who evaluated 10 randomly selected news articles, yielding 1,334 paired observations where each participant rated both the original and highlighted versions of the articles.

4.1 Effect of Highlighting Subjectivity

To assess whether highlighting subjective sentences affects perceived trustworthiness, we compared participants' trust ratings for each article before and after highlighting. We conducted a paired-samples t-test and, as a robustness check, a Wilcoxon signed-rank test, which does not assume normally distributed paired differences (Wilcoxon, 1945). Table 1 presents the results. The paired t-test revealed a statistically significant negative effect of highlighting subjectivity ($t = -8.42$, $p < 0.001$), with participants rating highlighted articles as less trustworthy (μ = 2.47, σ = 1.05) compared to non-highlighted versions (μ = 2.68, σ = 1.02), a mean decrease of 0.21 points. The Wilcoxon signed-rank test corroborated this finding ($p < 0.001$). However, the effect size was small (Cohen's $d = -0.23$), and two-thirds of ratings remained unchanged, suggesting modest practical impact.

4.2 Order Effect Analysis

To verify that the observed reduction in trust was not attributable to presentation order, we conducted a 2×2 repeated measures ANOVA with version (original vs. highlighted) as a within-subjects factor and presentation order (original-first vs. highlighted-first) as a between-subjects factor. Results confirmed the main effect of highlighting (F(1, 134) = 21.16, p < 0.001, η_p^2 = 0.136), replicating our paired t-test findings. Critically, we found no significant main effect of presentation order (F(1, 134) = 0.42, p = 0.52, η_p^2 = 0.003) and no version × order interaction (F(1, 134) = 0.18, p = 0.67, η_p^2 = 0.001), confirming that the effect of highlighting was consistent regardless of whether participants saw the original or highlighted version first. Separate paired t-tests within each order condition further validated this pattern, with both original-first and highlighted-first conditions showing statistically significant decreases in trust. These results rule out order effects as an alternative explanation and support a causal interpretation of the highlighting intervention.

Measure	**Original**	**Highlighted**
Mean (SD)	2.68 (1.02)	2.47 (1.05)
Median	3.00	3.00
Statistical Tests		
Paired t-test	t = -8.42, $p < 0.001$	
Wilcoxon test	$p < 0.001$	
Cohen's d_z	-0.23	
Direction of Change		
Decreased	274 (20.5%)	
No change	879 (65.9%)	
Increased	181 (13.6%)	

Table 1: Effect of highlighting subjective sentences on perceived trustworthiness. Values reflect paired comparisons across N = 1,334 participant–article pairs. Cohen's d_z denotes the paired-samples effect size.

4.3 Variation Across Articles

While the overall effect of highlighting subjective sentences was negative, Figure 1 reveals substantial heterogeneity across individual articles. The effect ranged from -0.35 (Article 1) to +0.18 (Article 6), a span of 0.53 points on the 4-point scale. Exploratory paired t-tests for each article individually showed that 9 of 17 articles exhibited statistically significant effects at p < 0.05 (uncorrected), with 4 articles surviving Bonferroni correction for multiple comparisons (p < 0.003). Notably, 14 articles showed decreased trustworthiness, while 3 showed increased trustworthiness after highlighting. This variability suggests that the impact of highlighting subjective content is moderated by article-specific characteristics, such as topic, content, writing style, or the nature of the subjective statements themselves. The consistent direction of effects (predominantly negative) combined with the varying magnitude indicates that while highlighting generally reduces perceived trustworthiness, the strength of this effect depends on contextual factors that warrant further investigation.

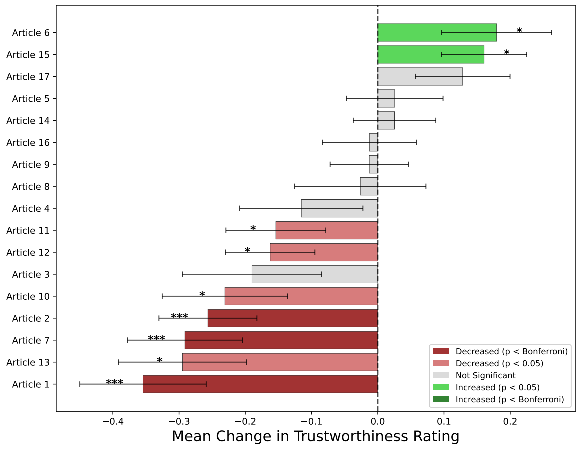

Figure 1: Mean change in trustworthiness ratings by article after highlighting subjective content. Error bars represent standard error of the mean. Asterisks indicate statistical significance from individual paired t-tests: * p < 0.05, *** p < 0.003 (Bonferroni-corrected). Negative values indicate decreased trustworthiness; positive values indicate increased trustworthiness.

4.4 Variation Across Raters

Beyond article-level effects of highlighting subjective content, we assess variability across participants. Some individuals may be more sensitive to the presence of highlighted subjectivity than others, leading to heterogeneous responses even when the overall trend is negative. Examining this participant-level variation helps clarify whether the observed effect reflects a broadly shared response or is driven by a subset of readers.

Analysis. To examine how consistently participants responded to highlighting, we computed the mean change in trustworthiness for each participant across all articles they rated (Δ*trust* = highlighted − original). Figure 2 shows the distribution of these mean changes, where negative values indicate a reduction in perceived trustworthiness after highlighting.

Results. Across 136 participants, the average change in perceived trustworthiness was small but consistently negative ($M = -0.21$, $SD = 0.34$), indicating that most participants rated highlighted articles slightly less trustworthy than their non-highlighted counterparts. However, there was substantial individual variability (range = $[-0.9, +1.0]$), as illustrated in Figure 2, suggesting that some participants were more sensitive to highlighting than others.

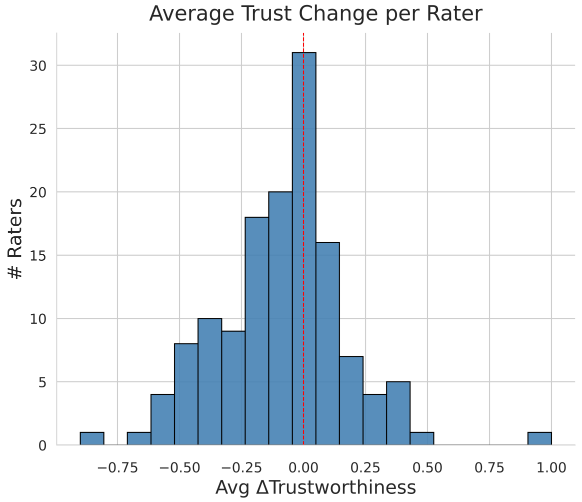

Figure 2: Distribution of mean change in trustworthiness ratings per participant (highlighted − original)

Moderation by Initial Vaccine Stance. To explore whether participants' prior attitudes toward vaccines moderated their response to highlighted subjectivity, we compared the mean change in trustworthiness ratings (Δ*trust*) across stance groups. Participants reported their general stance toward vaccines as a public health measure at the beginning of the survey. As shown in Figure 3, the direction of the effect was consistent across groups, with highlighting generally reducing perceived trustworthiness, though the magnitude of the decrease varied by stance. Participants who were somewhat supportive of vaccines exhibited the largest average decline ($M = -0.30$, $SD = 0.24$), followed by somewhat opposed ($M = -0.18$) and strongly opposed ($M = -0.05$) participants, while strongly supportive participants showed no change ($M = 0.00$).

A one-way ANOVA and Kruskal-Wallis test, which assess differences in central tendency across groups under parametric and non-parametric assumptions respectively, did not reveal a statistically significant difference across groups ($p > 0.05$). Nevertheless, the observed pattern is consistent with the possibility that highlighting subjective language has a stronger effect among participants with more moderate positions, while having little influence on those with firmly held views.

5 Features

To analyze factors associated with changes in perceived trustworthiness, we extracted interpretable features from both the articles and the participants

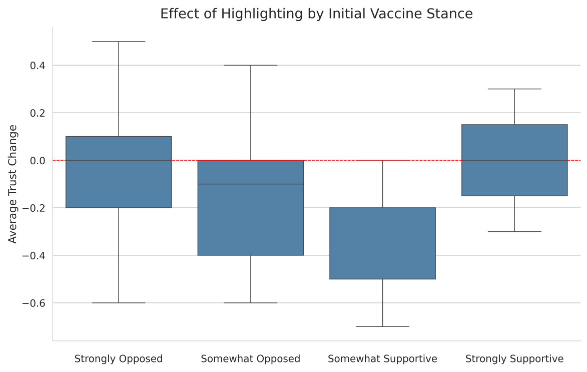

Figure 3: Mean change in perceived trustworthiness (highlighted − original) by participants' initial stance toward vaccines (Q59). Negative values indicate that highlighting subjective language reduced trust. Although the overall effect of highlighting was modestly negative across all groups, the magnitude of this reduction varied by stance, with the largest decline among participants who were somewhat supportive of vaccines. Error bars represent standard deviation within each stance group.

who rated them. Article-related features capture the linguistic, stylistic, and rhetorical properties of each article, while reader-related features describe participants' prior attitudes and behavioral tendencies that may form their sensitivity to highlighted subjective sentences.

5.1 Article Related Features

Subjectivity Density. This feature quantifies the proportion of sentences in an article classified as subjective, with values ranging from 0 to 1. It directly captures the extent of subjective language within an article. We hypothesize that articles with higher subjectivity density should elicit stronger trust reductions after highlighting, as the highlighted sentences occupy a greater share of the text.

Sentiment Polarity and Intensity. To capture the affective tone of each article, we computed two sentiment-based metrics: (1) the mean sentiment polarity, ranging from negative to positive, and (2) the average sentiment intensity, defined as the absolute magnitude of emotional valence regardless of sign. We derived these measures using the `cardiffnlp/twitter-roberta-base-sentiment` model (Loureiro et al., 2022), a transformer-based RoBERTa encoder that represents a widely adopted state-of-the-art approach for sentiment analysis in political and news discourse. For each sentence, the model outputs class probabilities for negative, neutral, and positive sentiment. We convert these probabilities to a continuous polarity score in $[-1, 1]$ by taking the expectation over class labels, and compute intensity as its absolute value. We then average these sentence-level scores to obtain article-level polarity and intensity and scale both features to the 0–1 range via min–max normalization for comparability across articles. Polarity reflects the direction of evaluative language, whereas intensity measures how emotionally charged the text is. We posit that articles with either highly positive or highly negative tone may be perceived as less objective, and thus may undergo greater reductions in perceived trustworthiness when subjective sentences are highlighted.

Persuasiveness. We quantify persuasiveness using a transformer-based regression model that predicts argument quality as a proxy for persuasive strength. Specifically, we apply the `webis/argument-quality-ibm-reproduced` model on Hugging Face, a recent reproduction of IBM's Project Debater argument-quality predictor (Zelch et al., 2025). The model estimates the perceived quality of written arguments on a continuous scale based on linguistic, syntactic, and discourse-level features learned from the large-scale Argument Quality dataset introduced by Gretz et al. (2020). For each article, we compute the model's predicted argument-quality score and normalize it to the 0–1 range to produce the persuasiveness score. Higher values indicate greater persuasive strength. This feature captures variation in an article's argumentative style that may be related to changes in perceived trustworthiness when subjective content is highlighted.

Readability. We measure textual readability using the Flesch–Kincaid Grade Level (Kincaid et al., 1975), a widely used readability index that estimates the years of education required to comprehend a text. The score is computed from two core components: the average number of words per sentence and the average number of syllables per word, following the formulation of Flesch (1948). For each article, we compute the Flesch–Kincaid score and then invert it so that higher values correspond to greater ease of reading. The resulting values are subsequently scaled to the 0–1 range via min–max normalization, where 0 indicates highly complex language and 1 corresponds to simple, accessible writing. Readability functions as a proxy for cognitive load: articles that are easier to read may render subjective statements more salient, whereas dense or syntactically intricate texts may obscure such cues and thus diminish the effect of highlighting.

Propaganda Techniques. To capture rhetorical strategies beyond sentiment and subjectivity, we annotated each article for the presence of specific propaganda techniques. We follow the taxonomy introduced in SemEval-2020 Task 11 on the Detection of Propaganda Techniques in News Articles (Da San Martino et al., 2020a), which defines a closed set of fine-grained rhetorical devices designed for span-level analysis in news text. This taxonomy includes commonly studied techniques such as Loaded Language, Name Calling or Labeling, Appeal to Fear or Prejudice, and Flag Waving. The definitions of all techniques are provided in Appendix A.2.

Because existing pretrained propaganda detectors are typically optimized for sentence-level span annotation rather than article-level presence, we employed a controlled, instruction-following large language model to perform document-level classification. Prior work has shown that large language models achieve performance comparable to state-of-the-art systems on propaganda detection tasks (Sprenkamp et al., 2023), motivating their use in this setting. Specifically, we prompted the OPENAI GPT-4.1-MINI model (OpenAI, 2024) with a standardized, closed-set annotation prompt requiring JSON-formatted outputs. For each article, the model identified which (if any) techniques appeared at least once, and we encoded these predictions as binary indicators. This approach provides an interpretable representation of rhetorical bias that complements the linguistic and affective features described above. The exact prompt used to obtain these annotations is shown below.

Propaganda Annotation Prompt

You are an expert annotator with extensive experience in analyzing political discourse and identifying rhetorical persuasion strategies. Your task is to determine whether an article contains any of a predefined set of propaganda techniques.

Follow these constraints carefully:

1. Use only the predefined propaganda technique labels.
2. A technique is present if it appears in at least one sentence.
3. Do not invent new labels and do not provide explanations.
4. Output a JSON dictionary of the form: `{"techniques_present": ["Loaded Language", "Red Herring", ...]}`

Examples:
`<FEW SHOT EXAMPLES>`

Article:
`<ARTICLE TEXT>`

5.2 Reader Related Features

Stance Toward Vaccines. Participants' initial attitudes toward vaccines were encoded as an ordinal variable ranging from strongly opposed (1) to strongly supportive (4). Prior stance plays an important role in how individuals interpret and spread misinformation, and may shape how readers respond to subjective language depending on whether it aligns with their beliefs (Derczynski et al., 2017; Hardalov et al., 2022). Stance values were normalized to the 0–1 range for use in the model.

6 Feature Analysis

Our analysis in Section 4 revealed that highlighting subjective language generally reduced perceived trustworthiness, though the magnitude of this effect varied across both articles and raters. To better understand the underlying drivers of this variability, we aim to identify the linguistic and attitudinal features that best explain the observed change in trustworthiness (Δ*trust*).

We hypothesize that both textual properties and individual differences (Verma et al., 2018) plausibly contribute to Δ*trust*, so we model the outcome as a weighted linear combination of article-related and reader-related features with separate scaling parameters for each group. Let $\mathbf{x}_j \in \mathbb{R}^p$ denote the article features for article j (e.g., subjectivity density, sentiment, readability, propaganda indicators) and $\mathbf{z}_i \in \mathbb{R}^q$ denote the rater features for participant i (e.g., stance, engagement). The model is expressed as

$$\Delta\text{trust}_{ij} = \alpha\,(\mathbf{w}^\top \mathbf{x}_j) + \beta\,(\mathbf{v}^\top \mathbf{z}_i) + b, \tag{1}$$

where α and β scale the relative influence of article- and reader-related predictors, $\mathbf{w}$ and $\mathbf{v}$ are the learned feature coefficients within each group, b is the intercept. Continuous predictors are normalized prior to estimation so that α and β are directly comparable in magnitude. This allows us to assess whether variation in trust change is driven more by article characteristics or by rater attributes.

We fit a linear regression model to estimate Δ*trust* using the feature set described above. Article-level predictors included subjectivity density, sentiment polarity and intensity, readability, and a set of one-hot propaganda indicators, while rater-level predictors consisted of each participant's vaccine stance. To ensure comparability across heterogeneous features, article and rater predictors were standardized separately before fitting the model. The regression jointly estimates the feature weights $\mathbf{w}$ and $\mathbf{v}$ as well as the group-level scaling parameters α and β, allowing us to directly compare the aggregate contribution of textual characteristics to that of rater attributes. The model was trained on all 1,334 participant-article pairs for which both original and highlighted trust ratings were available.

The model reveals a strong asymmetry in predictive influence between article- and participant-related features. The norm of the article-feature coefficients was substantially larger than that of the rater-feature coefficients ($\alpha = 0.34$ vs. $\beta = 0.039$), yielding an α/β ratio of **8.78**. This indicates that article properties account for nearly nine times more variance in Δ*trust* than individual differences in vaccine stance. Among article-level predictors, higher subjectivity density was associated with larger decreases in trust, while certain rhetorical devices such as Appeal to Authority and Black-and-White Fallacy showed notable associations with trust change. Participant stance exhibited a smaller but directionally consistent effect, with more pro-vaccine readers showing slightly larger trust reductions. Table 2 summarizes the standardized coefficients for all predictors. To assess the robustness of these estimates, we additionally report 95% bootstrap confidence intervals for the largest-magnitude coefficients in Appendix A.3.

Feature	**Coefficient**	**Type**
subj_ratio	-0.1968	Article
Black-and-White Fallacy	+0.1775	Article
polarity_raw	+0.1602	Article
Appeal to Authority	-0.0994	Article
Appeal to Fear or Prejudice	+0.0658	Article
Bandwagon	+0.0430	Article
Glittering Generalities	-0.0358	Article
vaccine_stance	-0.0388	Rater
intensity_raw	+0.0237	Article
fk_raw	+0.0236	Article
Loaded Language	+0.0207	Article
Name Calling or Labeling	+0.0147	Article
Exaggeration or Minimization	-0.0101	Article
readability	-0.0236	Article
Whataboutism	+0.0017	Article
Doubt	+0.0017	Article

Table 2: Standardized regression coefficients for the Δtrust prediction model, excluding features with zero-valued coefficients. Positive coefficients indicate that the feature is associated with increased trust in the highlighted version relative to the original, while negative coefficients indicate greater trust reduction. Article-level predictors dominate the model, with the aggregate coefficient magnitude of article features ($\alpha = 0.34$) substantially exceeding that of the rater feature ($\beta = 0.039$).

7 Conclusion

In this work, we examined how making subjective language explicit influences readers' perceptions of news trustworthiness. Through a controlled within-subjects survey, we showed that highlighting subjective sentences generally reduces perceived trust, though the magnitude of this effect varies across articles and individuals. Our feature-based analysis reveals that this variation is driven predominantly by article-level properties rather than readers' stance toward the topic. Among the article-related features, subjectivity density and specific rhetorical devices explain substantially more variance in trust change. These findings indicate that trust judgments are sensitive to how linguistic cues are surfaced during reading, even when factual content remains unchanged. This suggests that perception-focused cues may play a role in shaping reader evaluations alongside existing detection and verification efforts.

In light of these results, relying solely on real-time fact checking may be insufficient, since trust judgments often form before such interventions occur. Instead, our results point to the value of perception-focused interventions at the point of reading, complementing existing detection and verification efforts.

Limitations

Our study has several limitations that should be considered when interpreting the results. First, the number of news articles included in the experiment is relatively small. This choice reflects a deliberate trade-off driven by budgetary and statistical considerations. Because our primary objective is to measure how readers' trust judgments change in response to highlighting subjective language, reliable estimation of this effect requires a sufficiently large number of pre- and post-highlighting ratings for each article. Given a fixed budget, increasing the number of articles would necessarily reduce the number of evaluations per article. We prioritized depth of evaluation, resulting in approximately 75 ratings per article, at the cost of broader article coverage. Future work with larger budgets could extend this design to a wider range of articles while maintaining adequate per-article sampling.

Second, our feature analysis assumes a linear relationships between predictors and changes in perceived trustworthiness. While this choice supports interpretability and enables direct comparison between article-level and reader-level influences, it may bias the analysis toward simple relationships and overlook nonlinear effects or interactions between features. More flexible modeling approaches could capture richer patterns in how linguistic cues and reader characteristics jointly shape trust judgments.

Third, the set of features considered in this study is necessarily limited. Article-level predictors focus on a subset of interpretable linguistic and rhetorical properties, while reader-level features are restricted primarily to prior stance toward vaccines. Although this design facilitates transparent analysis, it does not exhaust the range of factors that may influence trust perceptions. Incorporating broader representations of discourse structure, narrative framing, or additional reader attributes may provide a more comprehensive account of trust formation.

References

Francesco Antici, Federico Ruggeri, Andrea Galassi, Katerina Korre, Arianna Muti, Alessandra Bardi, Alice Fedotova, and Alberto Barrón-Cedeño. 2024. A corpus for sentence-level subjectivity detection on english news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 273–285.

Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty, Elissa M Redmiles, Meeyoung Cha, and Krishna P Gummadi. 2021. Analyzing biases in perception of truth in news stories and their implications for fact checking. *IEEE Transactions on Computational Social Systems*, 9(3):839–850.

Matyas Bohacek, Michal Bravansky, Filip Trhlík, and Václav Moravec. 2023. Czech-ing the news: Article trustworthiness dataset for czech. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 96–109.

Jiyoung Cha. 2024. Predictors of the credibility of social media as a news outlet: An examination of the influences of social media contacts, source perceptions, and media use. *International Journal on Media Management*, 26(1-2):68–93.

Limeng Cui and Dongwon Lee. 2020. Coaid: Covid-19 healthcare misinformation dataset. *CoRR*.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 1377–1414.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Yu Seunghak, Roberto Di Pietro, Preslav Nakov, and 1 others. 2020b. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}*, pages 4826–4832. International Joint Conferences on Artificial Intelligence Organization.

Lingjia Deng and Janyce Wiebe. 2015. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1323–1328.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 69–76.

Akram Elbouanani, Evan Dufraisse, Aboubacar Tuo, and Adrian Popescu. 2025. Cea-list at checkthat! 2025: evaluating llms as detectors of bias and opinion in text. *arXiv preprint arXiv:2507.07539*.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.

Diego Garusi and Sergio Splendore. 2023. Advancing a qualitative turn in news media trust research. *Sociology compass*, 17(4):e13075.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15125–15140, Suzhou, China. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis-and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277.

Hendrik Heuer and Andreas Breiter. 2018. Trust in news on social media. In *Proceedings of the 10th Nordic conference on human-computer interaction*, pages 137–147.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Castulus Kolo, Joschka Mütterlein, and Sarah Anna Schmid. 2022. Believing journalists, ai, or fake news: The role of trust in media. In *HICSS*, pages 1–10.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4 edition. Sage Publications.

Seung Yeop Lee and Sang Woo Lee. 2023. Normative or effective? the role of news diversity and trust in news recommendation services. *International Journal of Human–Computer Interaction*, 39(6):1216–1229.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yunfei Long, Qin Lu, Rong Xiang, Minglei Li, and Chu-Ren Huang. 2017. Fake news detection through multi-perspective speaker profiles. In *Proceedings of the eighth international joint conference on natural language processing (volume 2: Short papers)*, pages 252–256.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Timelms: Diachronic language models from twitter. *arXiv preprint arXiv:2202.03829*.

Xiaoxiao Ma, Yuchen Zhang, Kaize Ding, Jian Yang, Jia Wu, and Hao Fan. 2024. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 508–521, Miami, Florida, USA. Association for Computational Linguistics.

Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

D Harrison McKnight, Larry L Cummings, and Norman L Chervany. 1998. Initial trust formation in new organizational relationships. *Academy of Management review*, 23(3):473–490.

Arkadiusz Modzelewski, Witold Sosnowski, Tiziano Labruna, Adam Wierzbicki, and Giovanni Da San Martino. 2025. PCoT: Persuasion-augmented chain of thought for detecting fake news and social media disinformation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. Technical report.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic detection of fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3391–3401.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 231–240.

Anku Rani, Dwip Dalal, Shreya Gautam, Pankaj Gupta, Vinija Jain, Aman Chadha, Amit Sheth, and Amitava Das. 2025. Sepsis: I can catch your lies–a new paradigm for deception detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 97–128.

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

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Jukka Ruohonen. 2024. A comparative study of online disinformation and offline protests. *SN Social Sciences*, 4(12):232.

Elena Savinova and Fermin Moscoso Del Prado. 2023. Analyzing subjectivity using a transformer-based regressor trained on naïve speakers' judgements. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 305–314.

Vivek Sharma, Mohammad Mahdi Shokri, Sarah Ita Levitan, Elena Filatova, and Shweta Jain. 2025. Analysis of propaganda in tweets from politically biased sources. *arXiv preprint arXiv:2507.08169*.

Mohammad Shokri, Vivek Sharma, Elena Filatova, Shweta Jain, and Sarah Levitan. 2024. Subjectivity detection in English news using large language models. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 215–226, Bangkok, Thailand. Association for Computational Linguistics.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large language models for propaganda detection. *arXiv preprint arXiv:2310.06422*.

Jingrong Tong. 2024. From content to context: A qualitative case study of factors influencing audience perception of the trustworthiness of covid-19 data visualisations in uk newspaper coverage. *Journalism*, 25(7):1481–1499.

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290.

Pramukh Nanjundaswamy Vasist, Debashis Chatterjee, and Satish Krishnan. 2024. The polarizing impact of political disinformation and hate speech: A cross-country configural narrative. *Information Systems Frontiers*, 26(2):663–688.

Aswathy Velutharambath, Amelie Wührl, and Roman Klinger. 2024. How entangled is factuality and deception in german? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9538–9554.

Nitin Verma, Kenneth R Fleischmann, and Kolina S Koltai. 2018. Demographic factors and trust in different news sources. *Proceedings of the Association for Information Science and Technology*, 55(1):524–533.

Svitlana Volkova and Jin Yea Jang. 2018. Misleading or falsification: Inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the The Web Conference 2018*, pages 575–583.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Andrea Wenzel, Anthony Nadler, Melissa Valle, and Marc Lamont Hill. 2018. Listening is not enough: Mistrust and local news in urban and suburban philly. *Columbia Journalism Review*.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational linguistics*, 30(3):277–308.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Ines Zelch, Matthias Hagen, Benno Stein, and Johannes Kiesel. 2025. Reproducing the argument quality prediction of project debater. In *Proceedings of the 12th Argument mining Workshop*, pages 181–188.

A Appendix

A.1 Training a Subjectivity Classifier

To filter articles with a subjective sentence ratio greater than 10%, we trained a sentence-level subjectivity classifier based on RoBERTa (Liu et al., 2019). The model was trained on two subjectivity-annotated datasets: the MPQA opinion corpus (Deng and Wiebe, 2015) and a recently released dataset containing sentence-level subjectivity annotations for COVID-19 and crime-related news (Savinova and Del Prado, 2023).

MPQA. The MPQA (Multi-Perspective Question Answering) dataset is a widely used resource in sentiment analysis and opinion mining. It is designed to capture the multifaceted nature of subjective language and contains text segments annotated for sentiment polarity and subjectivity. The dataset includes content from multiple sources, such as news articles, product reviews, discussion forums, and social media, reflecting the varied contexts in which subjective expressions occur. To control for genre-related variation, we restrict our experiments to MPQA sentences drawn from news articles. We use version three of the MPQA opinion corpus.

News Dataset. The second dataset consists of sentences from news articles and Facebook posts related to *crime* and *COVID-19*, published by four major UK news outlets, for a total of 7,751 sentences (Savinova and Del Prado, 2023). We exclude Facebook posts, as they are shorter and may not be authored by journalists. Consequently, all experiments in this paper use only news sentences, yielding a dataset of 2,973 sentences, including 1,013 subjective and 1,960 objective sentences.

Training Details. We fine-tuned a RoBERTa-base model for binary sentence-level subjectivity classification using the Adam optimizer with a learning rate of 1×10^{-5}. Training was performed for 10 epochs with standard cross-entropy loss. Across the two training datasets, the classifier achieved an average F1 score of 0.88. The resulting model was used to estimate the proportion of subjective sentences in each article for the purpose of article filtering.

A.2 Propaganda Techniques

Below is a list of the definitions for propaganda techniques we used for propaganda detection in news articles (Da San Martino et al., 2020b):

- Loaded language: Using words/phrases with strong emotional implications (positive or negative) to influence an audience.
- Name calling or labeling: Labeling the object of the propaganda campaign as either something the target audience fears, hates, finds undesirable or otherwise loves or praises.
- Repetition: Repeating the same message over and over again, so that the audience will eventually accept it.
- Appeal to fear: support an idea by instilling fear against other alternatives
- Exaggeration or minimization: Either representing something in an excessive manner: making things larger, better, worse.
- Doubt: Questioning the credibility of someone or something. - appeal to fear/prejudice: Seeking to build support for an idea by instilling anxiety and/or panic in the population towards an alternative, possibly based on preconceived judgments.
- Flag-waving: Playing on strong national feeling (or with respect to a group, e.g., race, gender, political preference) to justify or promote an action or idea.
- Causal oversimplification: Assuming one cause when there are multiple causes behind an issue. We include scapegoating as well which is defined as the transfer of the blame to one person or group of people without investigating the complexities of an issue.
- Slogans: A brief and striking phrase that may include labeling and stereotyping.
- Appeal to authority: Stating that a claim is true simply because a valid authority/expert on the issue supports it, without any other supporting evidence. Include the special case where the reference is not an authority/expert,although it is referred to as testimonial in the literature.
- Black-and-white fallacy: Presenting two alternative options as the only possibilities, when in fact more possibilities exist, eliminating any other possible choice. and as an extreme telling the audience exactly what actions to take, which is also called as dictatorship.

- Thought-terminating cliche: Words or phrases that discourage critical thought and meaningful discussion about a given topic. They are typically short, generic sentences that offer seemingly simple answers to complex questions or that distract attention away from other lines of thought.

- Whataboutism: Discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

- Reductio ad hitlerum: Persuading an audience to disapprove an action or idea by suggesting that the idea is popular with groups hated in contempt by the target audience. It can refer to any person or concept with a negative connotation.

- Red herring: Introducing irrelevant material to the issue being discussed, so that everyone's attention is diverted away from the points made.

- Bandwagon: Attempting to persuade the target audience to join in and take the course of action because "everyone else is taking the same action".

- Obfuscation/intentional vagueness/confusion: Using deliberately unclear words, so that the audience may have its own interpretation."

- Straw men: Refuting arguments that were not presented.

A.3 Confidence Intervals for Feature Coefficients

To quantify uncertainty in the feature-level effects reported in Section 6, we estimate 95% confidence intervals for the regression coefficients using non-parametric bootstrap resampling. Specifically, we repeatedly resample the participant-article pairs with replacement and refit the linear model described in Equation (1) on each resampled dataset. For each feature, we compute percentile-based confidence intervals from the resulting empirical distribution of coefficients.

Figure 4 shows the bootstrap means and 95% confidence intervals for the features with the largest absolute coefficients. Subjectivity density exhibits a consistently negative association with changes in perceived trustworthiness, with confidence intervals that exclude zero, indicating a stable effect across bootstrap samples. In contrast, many individual rhetorical features display wider confidence intervals that often include zero, reflecting substantial variability across articles and contexts. Reader-level stance remains small and tightly bounded around zero, reinforcing the conclusion that variation in trust change is driven primarily by article-level properties rather than individual differences.

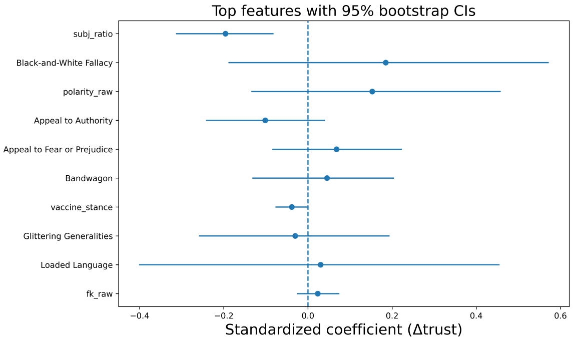

Figure 4: Bootstrap means and 95% confidence intervals for the largest-magnitude regression coefficients predicting changes in perceived trustworthiness (Δtrust). Error bars indicate percentile-based confidence intervals estimated via 1,000 bootstrap resamples. Positive values indicate increased trust in the highlighted version relative to the original, while negative values indicate greater trust reduction.