Reducing Bias in a Misinformation Classification Task with Value-Adaptive Instruction

Nicholas Diana[1][0000−0002−8187−3692] and John Stamper[2][0000−0002−2291−1468]

[1] Colgate University ndiana@colgate.edu
nickdiana.com
[2] Carnegie Mellon University jstamper@cmu.edu
dev.stamper.org

Abstract. Instructional technology that supports the development of media literacy skills has garnered increased attention in the wake of recent misinformation campaigns. While critical, this work often ignores the role of myside bias in the acceptance and propagation of misinformation. Here we present results from an alternative approach that uses natural language processing to model the dynamic relationship between the user and the content they are consuming. This model powers a debiasing intervention in the context of a "fake news detection" task. Information about the user- and content-values was used to predict when the user may be prone to myside bias. The intervention resulted in significantly better performance on the misinformation classification task. These results support the development of content-general and embedded debiasing systems that could encourage informal learning and bias reduction in real-world contexts.

Keywords: Misinformation · Myside Bias · Confirmation Bias · Personalization · Media Literacy · Civic Technology · Civics Education

1 Introduction

Modern digital media has novel features that set it apart from traditional media, including a lack of editorial oversight, the democratization of media sources, and the rapid propagation of stories (particularly stories that are emotionally charged). Many of these features provide an infrastructure that allows misinformation to flourish in ways that would be difficult in a pre-digital age [13]. In response to these new and pressing challenges, media literacy education has increasingly emphasized misinformation classification (i.e., the ability to accurately identify misinformation) as a critical civic skill. Recent successful misinformation campaigns illustrate both the public's susceptibility to believing so-called "fake news" [13] as well as the dire consequences that result from a failure to teach and exercise this fundamental media literacy skill.

While there has been an increased interest in the development of instructional tools designed to improve misinformation classification [9, 11], few solutions account for the impact of bias. In this paper we present a contrasting approach

that aims to model the dynamic relationship between the user and the content they are consuming. We used this model to predict when the user may be most susceptible to bias, and to provide adaptive recommendations in those moments.

We hypothesized that an intervention that leverages the *Alignment* between user and content values will reduce the impact of myside bias on ratings of plausibility. We expect that user ratings after seeing the value-adaptive intervention will be more accurate, and that, generally, the intervention will encourage any change in ratings to be a change in the right direction (i.e., towards the correct answer). This work may inform the development of tools for reducing bias when evaluating the veracity of information in real-world contexts.

Background The specific skill isolated in the current experiment is one's ability to accurately estimate the plausibility of events, specifically in the realm of United States politics. These estimations are based on what we know about various political actors and how we believe they might behave. As such, these estimations can be honed with experience, by comparing what we believed was plausible with what was actually true. In the real world, a number of potential factors may play a role in determining plausibility. We attempt to control for these factors to isolate impact of bias, specifically *Myside Bias*, or one's tendency to evaluate claims or evidence more favorably if the claim or evidence supports one's own beliefs or worldview [15]. We expect myside bias to cause users to overestimate the plausibility of headlines that support their own beliefs, ultimately impacting the user's accuracy on the misinformation classification task.

We estimated user values using Moral Foundations Theory [7], which argues that moral judgements are driven by the importance we ascribe to a small set of *moral foundations*. These moral foundations have been empirically shown to be highly predictive of both general voting behavior [4] as well as specific political beliefs [10]. The output of the Moral Foundations Questionnaire (MFQ) is a vector of five scores, representing the degree to which the student values each of the five foundations when making moral judgments.

2 Method

Based on a power analysis, eighty-three (83) participants were recruited using the participant recruitment platform Prolific. Participants were required to be 18 years of age or older, U.S. citizens, and not have participated in any of our research group's prior studies. The estimated completion time was 28 minutes, and participants were paid $3.15 ($6.75/hour) for participating. Participants who failed reading-checks (n=2) were excluded from analyses. The remaining 81 participants (36 female, 42 male, and 3 "Other/Prefer not to say") ranged in age from 18-68 years old (M=34.44). These participants were drawn from a politically diverse population as evidenced by their scores on the MFQ.

The entirety of the experiment was conducted through an online web interface. After viewing the consent form, participants were directed to a set of instructions that described the main task of the experiment (misinformation

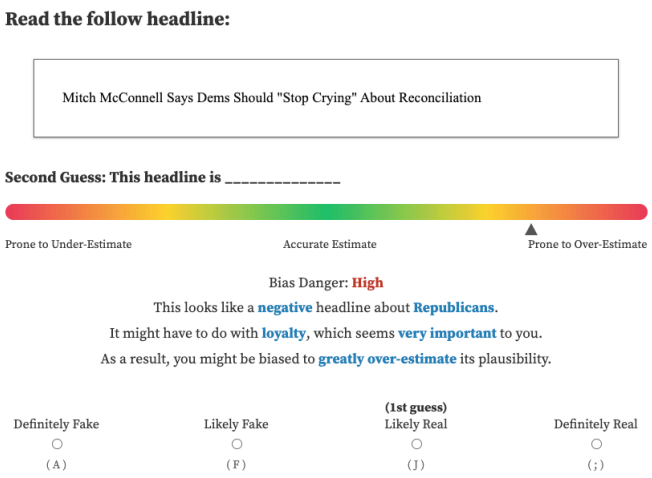

Fig. 1. A screenshot of the online interface including the value-adaptive intervention. The model-driven components of the adaptive intervention are displayed in bolded blue text. User performance improved significantly after seeing the intervention ($p < .05$).

classification). Following the instructions, participants were given a pre-study survey that included the MFQ [8]. Participants were then directed to the misinformation classification task: a series of 52 news headlines taken from or based on Politifact headlines in the FakeNewsNet news misinformation dataset [12]. Headlines had two relevant features: authenticity (authentic or fabricated) and veracity (real or fake). *Authentic Real* and *Authentic Fake* headlines were actual news headlines classified as either "real" or "fake" (respectively) by Politifact. *Fabricated Real* and *Fabricated Fake* headlines were exact copies of authentic headlines, except that the subject of the headline was changed to a subject from the opposing side of the political spectrum. For example, the *authentic fake* headline *"BREAKING: Federal Judge Grants Permission To Subpoena Trump"* would be changed to the *fabricated fake* headline *"BREAKING: Federal Judge Grants Permission To Subpoena Obama"*.

For each item, users provided an initial rating of plausibility, then were shown additional (non-correctness) feedback (i.e., the debiasing intervention), and finally were asked to provide a second rating of plausibility in light of this information. After providing a second rating, users were given correctness feedback. In this way, the task closely resembles the Judge Advisor System [14] often employed in decision-making research. Following the classification task, participants were asked to complete a short post-study questionnaire that included questions about demographic information.

Alignment is a metric designed to estimate the extent to which the user's values (as measured by the MFQ) align with the values present in the text of the headline the user is reading. In this experiment, alignment was computed using Distributed Dictionary Representations (DDR) [6] an NLP method. This

method allows for the modeling of abstract psychological constructs, such as the foundations in Moral Foundations Theory. The output of this process is a vector of five scores, representing the degree to which the text was semantically similar to each of the five foundations (see [2] for a more complete discussion of this process). *Alignment* is computed simply by computing the cosine similarity of the result of the user's Moral Foundations Questionnaire (a vector of five values, one per foundation) and the result of the DDR analysis (a similar vector of five values, one per foundation). Previous work has shown that alignment is a reliable predictor of bias on argument evaluation tasks [3].

After providing an initial rating, users were given model-driven feedback about their predicted susceptibility to bias. If above a threshold (50%), then the user was shown additional model-driven information including estimations of the headline's predominate value, political affiliation, valence, and the relationship to a user's values. Figure 1 shows a screenshot of the intervention. The blue text in the figure indicate the model-generated and/or user-adaptive feedback components. The generation of each of the elements of the intervention is detailed below.

The intervention consisted of two stages. First, a logistic regression model was used to predict the likelihood of a correct answer given alignment, the number of prior opportunities, and item-level effects (i.e., average difficulty). The beta values used in this model were derived from the results of a pilot study. This information is conveyed to users in the form of text reading "Bias Danger: [Level], where [Level] is low ($> 75\%$ likelihood), moderate ($50 - 75\%$) or high ($< 50\%$).

If the likelihood of a correct response was less than 75%, the user was shown an elaborated intervention that included information from two additional predictive models. Predictions about the text's valence, subject, and most relevant foundation were generated through the use of a SimCSE (Simple Contrastive Learning of Sentence Embeddings) model trained on the Stanford Natural Language Inference Corpus [5]. This model assessed the similarity of headline text to a set of archetypal sentences based on the topics explored in the Moral Foundations Vignettes [1]. The SimCSE model chose the archetypal sentence that most closely matched each news headline (e.g., the (fake) news headline *"BREAKING: Federal Judge Grants Permission To Subpoena Obama"* was most similar to the negative archetypal sentence *"A Democrat breaks the law."*).

Users were also given a prediction about their likelihood to over- or underestimate plausibility due to bias. This second likelihood prediction was derived from a second logistic regression model that predicts the likelihood of overestimation (coded as 1) or underestimation (coded as 0) given the user's alignment score and the number of previous opportunities. Again, the beta values used in this model were derived from the results of a pilot study. Additionally, a qualifier of either "slightly" or "greatly" was given to the prediction based on the likelihood score (either between 25% and 75% or outside of that range, respectively).

3 Results and Discussion

We hypothesized that user ratings after seeing the debiasing intervention would be more accurate than their initial ratings. A paired samples t-test was used to compare the outcomes of first and second attempts. There was a slight but significant improvement in outcomes between first ($M = .69, SD = .45$) and second attempts ($M = .71, SD = .45$) $t(80) = -2.56, p = .01$). Changes in ratings from one class to another were relatively rare. To get a more nuanced measure of the impact of the intervention, the direction of movement between the initial and second rating was also analyzed. That is, we assessed whether or not the intervention encouraged movement toward the correct answer – even if the user ultimately provided an incorrect class. We found that, of the instances in which a user changed their score between the first and second ratings, users were significantly more likely to "move" their ratings in the correct direction ($X^2(2, 2616) = 119.87, p < .001$) after seeing the intervention. Users were also asked to provide feedback about the quality and effectiveness of the AI assistance. Most users found the AI assistant's feedback to be helpful and mostly accurate.

These results provide additional evidence for the importance of *user-content alignment* in misinformation classification. Users more accurately classified misinformation after seeing the value-adaptive feedback, and the intervention encouraged movement towards the correct response. Taken together, these results suggest that the intervention resulted in both incremental and meaningful changes in user responses, a finding mirrored in the qualitative user feedback. This work has implications for the development of future media literacy instructional technologies, suggesting that accurate models of user learning in this require the consideration of bias. Future work will aim to provide similar value-adaptive debiasing interventions in real-world contexts. Integrating this *just-in-time* intervention into real-world settings where users encounter misinformation will shed light on the impact of this intervention in the presence of the numerous other factors that may play a role in a user's determination of plausibility.

The nature of the intervention's presentation in this study was limited by the fact that it leveraged item-level information (based off of previous experiments) in the initial outcome-based prediction stage. Including this item-level information provides greater accuracy, as it likely captures important baseline plausibility information. While alignment is included in this outcome prediction model, the practical result of this prediction is that users are seeing the elaborated intervention on items that are, on average, more difficult to classify. This is perhaps unavoidable as the specific context of an individual headline (i.e., the actors and their behavior) may always be the primary factor in determining plausibility. Nevertheless, to isolate the impact of bias, the second, bias-based prediction stage did not leverage any item-level information.

4 Conclusion

Identifying misinformation is a key media literacy skill, and one that may depend on the interaction between the user and the content they are consuming.

We found that a value-adaptive debiasing intervention improved performance on a misinformation classification task. These results provide evidence for the importance of the dynamic relationship between user- and content-values, particularly in the media literacy domain.

References

1. Clifford, S., Iyengar, V., Cabeza, R., Sinnott-Armstrong, W.: Moral foundations vignettes: A standardized stimulus database of scenarios based on moral foundations theory. Behavior research methods **47**(4), 1178–1198 (2015)
2. Diana, N., Stamper, J., Koedinger, K.: Towards value-adaptive instruction: A data-driven method for addressing bias in argument evaluation tasks. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–11 (2020)
3. Diana, N., Stamper, J.C., Koedinger, K.: Predicting bias in the evaluation of unlabeled political arguments. In: CogSci. pp. 1640–1646 (2019)
4. Franks, A.S., Scherr, K.C.: Using moral foundations to predict voting behavior: Regression models from the 2012 us presidential election. Analyses of Social Issues and Public Policy **15**(1), 213–232 (2015)
5. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821 (2021)
6. Garten, J., Hoover, J., Johnson, K.M., Boghrati, R., Iskiwitch, C., Dehghani, M.: Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. Behavior research methods **50**(1), 344–361 (2018)
7. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral foundations theory: The pragmatic validity of moral pluralism. In: Advances in experimental social psychology, vol. 47, pp. 55–130. Elsevier (2013)
8. Graham, J., Nosek, B.A., Haidt, J., Iyer, R., Spassena, K., Ditto, P.H.: Moral foundations questionnaire. Journal of Personality and Social Psychology (2008)
9. Hone, B., Rice, J., Brown, C., Farley, M.: Factitious (2018), factitious.augamestudio.com
10. Koleva, S.P., Graham, J., Iyer, R., Ditto, P.H., Haidt, J.: Tracing the threads: How five moral concerns (especially purity) help explain culture war attitudes. Journal of Research in Personality **46**(2), 184–194 (2012)
11. Literat, I., Chang, Y.K., Eisman, J., Gardner, J.: Lamboozled!: The design and development of a game-based approach to news literacy education. Journal of Media Literacy Education **13**(1), 56–66 (2021)
12. Shu, K., Mahudeswaran, D., Wang, S., Lee, D., Liu, H.: Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data **8**(3), 171–188 (2020)
13. Silverman, C.: This analysis shows how viral fake election news stories outperformed real news on facebook (Nov 2016), https://www.buzzfeednews.com/craigsilverman/viral-fake-election-news- outperformed-real-news-on-facebook
14. Sniezek, J.A., Buckley, T.: Cueing and cognitive conflict in judge-advisor decision making. Organizational behavior and human decision processes **62**(2), 159–174 (1995)
15. Stanovich, K.E., West, R.F., Toplak, M.E.: Myside bias, rational thinking, and intelligence. Current Directions in Psychological Science **22**(4), 259–264 (2013)