PairWise: Mitigating Political Bias in Crowdsourced Content Moderation

Jacob Thebault-Spieker[1], **Sukrit Venkatagiri**[1], **David Mitchell**[2], **Chris Hurt**[1], **Kurt Luther**[1]
[1]Virginia Tech, [2]University of Illinois at Urbana–Champaign
{jthebaultspieker, sukrit, chris526, kluther}@vt.edu, davidgm2@illinois.edu

Abstract

Crowdsourced labeling of political social media content is an area of increasing interest, due to the contextual nature of political content. However, there are substantial risks of human biases causing data to be labelled incorrectly, possibly advantaging certain political groups over others. Inspired by the social computing theory of social translucence and findings from social psychology, we built PairWise, a system designed to facilitate interpersonal accountability and help mitigate biases in political content labelling.

Introduction

Labelling political content is increasingly challenging for large social media companies. For instance, when Mitt Romney spoke about "binders full of women" during a 2012 US presidential debate, "Twitter needed . . . to figure out, in real-time, why such an obtuse phrase so quickly became such a popular hashtag and whether it was an appropriate thing to post to its trending topics" (Gray and Suri 2017). Further, Facebook recently (Business 2018) enacted a policy requiring that political accounts (a) go through an authorization process and (b) label any political content they post. Such tasks require companies to label political content.

Given the importance and massive scale of these tasks, companies have invested substantial efforts in automating content moderation. But even as capabilities improve, the need for human effort — often crowd workers — to handle difficult cases remains (Gray and Suri 2017). One major reason for including humans in the labelling process is the recognition that content labeling is subjective and contextual, and therefore requires human intelligence. However, there is a risk of systematic biases advantaging some political views over others, due to the vagaries in human judgment and the subjective nature of labeling tasks. Concerns about political biases in content moderation are so widespread that the US government recently set up a process that allowed people to report perceived instances of bias (Stack 2019).

To address these problems, we built PairWise, a system that brings together pairs of crowd workers to synchronously moderate political content. PairWise seeks to mitigate political bias by leveraging design cues from both social psychology and the social computing theory.

Design Rationale

Research in judgment and decision-making psychology (Lerner and Tetlock 2002) suggests that it may be possible to attenuate political bias in labelling tasks by making crowd workers (a) accountable to another person, where they (b) know who that person is and (c) believe that person has legitimate claim to hold them accountable. This echoes social translucence theory, in which Erickson and Kellogg (2000) demonstrated Babble, a system that created *accountability* through *visibility*, and *awareness* among coworkers by visualizing who was participating and who was not, and making members of the team aware of that pattern. The social psychology literature parallels these ideas, emphasizing the importance of knowing who one is accountable to, and believing they are legitimate in that role.

Both social psychology research and social translucence theory also suggest that for accountability to occur, it is important to be able to *understand social context that influences work decisions and processes*. Erickson and Kellogg's Babble made visible online the social context that already existed between coworkers. However, such context does not currently exist in crowdsourcing, where crowd workers commonly work independently, but in parallel on related tasks. They almost certainly do not *know* one another, may be entirely unaware of one another, and cannot see each other's work.

Therefore, with the goal of using social accountability to attenuate political bias in crowdsourced labeling decisions, we designed PairWise with three core features (Figure 1): (1) *shared work environment*, (2) *visible work output*, and (3) *contextual information about their coworker*. By creating a socially translucent work environment and the necessary conditions for accountability in human judgment and decision making, PairWise integrates social computing theory and social psychology findings, towards mitigating biases in political content labelling.

The PairWise System

We built PairWise as a real-time web application, where workers are paired off into their own work environments, to complete a set of political labelling tasks alongside one another. PairWise is built using the Python Flask framework,

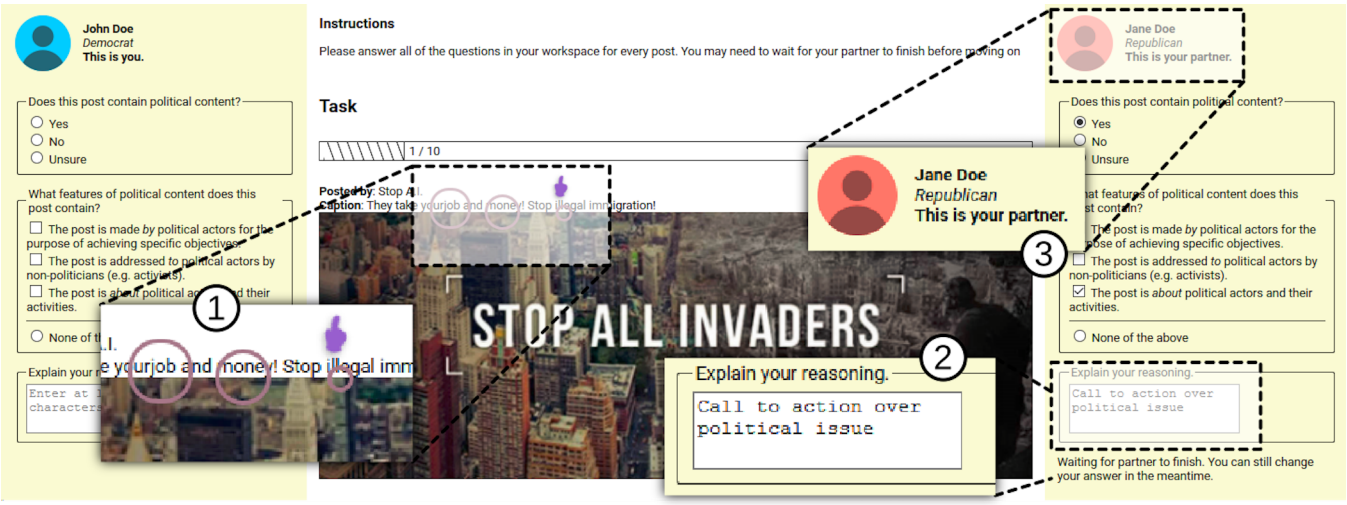

Figure 1: An example interface screenshot, showing the three core features of PairWise. (1) shows a *shared work environment*, indicating working synchronously with another person. (2) shows the other person's *work output*. (3) shows *contextual information* about the other person.

and showing workers' presence to one another is supported through the JavaScript library TogetherJS.[1]

The PairWise interface creates this shared work environment with the three core features described above. In the *shared work environment*, workers doing the same task in parallel can see one another doing the task — PairWise mirrors their coworker's mouse movements on their own screen. Second, to make *visible work output* for each coworker, PairWise also shows each worker their coworker's answers to the labelling task in real time, lending credibility to their legitimacy as a coworker. Third, PairWise provides *politically relevant information* about each worker to the other (e.g., their political orientation) through their profile. Each of these features is designed to help move an individual labelling task interface towards an environment that provides a rich context between people working towards the same goal.

As a crowd worker joins PairWise, they arrive at a page where they wait to be paired up with a coworker. Once that coworker has joined, both workers are redirected to their *shared work environment* which consists of a central panel that shows shared task information (a social media post that needs to be labelled), and two work panels, on either side of the task area (one work panel for each user). Each work panel asks three questions about the post: (1) is the post political, non-political, or unclear? (2) which political communication guidelines (McNair 2017) apply to your decision? and (3) provide written rationale for your assessment. Throughout this process, workers can see both their own answers and cursor, a well as those of their coworker. Each work panel also shows a profile card at the top of the screen, which provides relevant contextual information about each worker (e.g. their political affiliation). Once both coworkers have responded, then they both move on to the next post.

Traditionally, workers would complete these tasks in parallel, but entirely independently. By contrast, PairWise seeks to create accountability in the political labelling task by providing workers with visibility into who their coworker is, and by providing evidence of whether that person has legitimate claim to hold the worker accountable. By creating a shared context for their interactions, their understanding of one another, and their work, PairWise is designed for both participants to be aware that their coworker is doing the same work they are, in order to create the conditions for both workers to attenuate their own biases.

Evaluation (In Progress)

We are conducting a controlled experimental study focused on the efficacy of our interventions in mitigating bias. This experiment will address three research questions. First, does bias manifest in political labelling tasks? Second, do socially and contextually transparent signals help mitigate bias? And third, how do socially and contextually transparent signals compare to other, non-social bias mitigation strategies?

To measure the effects on bias, our study compares crowd workers' answers in different conditions to a ground truth dataset (McNair 2017). Formally, we are running a $1 + 4 \times 2$ between-subjects experiment, wherein we vary both our bias mitigation strategies and the amount of information a worker receives about their coworker. With regard to our bias mitigation strategies, we have one non-social bias mitigation *baseline* condition informed by prior work (Hube, Fetahu, and Gadiraju 2019), and three *social intervention* bias mitigation conditions. With regard to the amount of information a worker receives about their coworker, workers will either both be anonymous, or they will be shown each other's political orientation. For an authentic stimulus set, we are using the Facebook ads made public as a part of the US House Intelligence Committee investigation into Russia's role in the 2016 US presidential election [2], which has been labelled based on (McNair 2017) to serve as ground truth.

[1] https://togetherjs.com/

[2] https://intelligence.house.gov/social-media-content/

References

Business, F. 2018. The Authorization Process for US Advertisers to Run Political Ads on Facebook is Now Open.

Erickson, T., and Kellogg, W. A. 2000. Social translucence: an approach to designing systems that support social processes. *Transactions on Computer-Human Interaction (TOCHI* 7(1):59–83.

Gray, M. L., and Suri, S. 2017. The Humans Working Behind the AI Curtain. *Harvard Business Review*.

Hube, C.; Fetahu, B.; and Gadiraju, U. 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*, CHI '19, 407:1–407:12. New York, NY, USA: ACM. event-place: Glasgow, Scotland Uk.

Lerner, J., and Tetlock, P. 2002. Bridging individual, interpersonal and institutional approaches to judgment and choice: The impact of accountability on cognitive bias. *Emerging perspectives in judgment and decision making, ed. S. Schneider and J. Shanteau* 431–57.

McNair, B. 2017. *An introduction to political communication*. Routledge.

Stack, L. 2019. Trump Wants Your Tales of Social Media Censorship. And Your Contact Info. *The New York Times*.