Crowd-O-Meter: Predicting if a Person is Vulnerable to Believe Political Claims

Mehrnoosh Sameki[*], Tianyi Zhang[+], Linli Ding[+], Margrit Betke[*], Danna Gurari[+]
[*]Boston University, [+]University of Texas at Austin

Abstract

Social media platforms have been criticized for promoting false information during the 2016 U.S. presidential election campaign. Our work is motivated by the idea that a platform could reduce the circulation of false information if it could estimate whether its users are vulnerable to believing political claims. We here explore whether such a vulnerability could be measured in a crowdsourcing setting. We propose Crowd-O-Meter, a framework that automatically predicts if a crowd worker will be consistent in his/her beliefs about political claims; i.e., consistently believes the claims are true or consistently believes the claims are not true. Crowd-O-Meter is a user-centered approach which interprets a combination of cues characterizing the user's implicit and explicit opinion bias. Experiments on 580 quotes from PolitiFact's fact checking corpus of 2016 U.S. presidential candidates show that Crowd-O-Meter is precise and accurate for two news modalities: text and video. Our analysis also reveals which are the most informative cues of a person's vulnerability.

I. Introduction

Social media have taken on the important roles of news gathering and circulating. Current estimates indicate that 62% of U.S. adults receive news on social media (Gottfried and Shearer 2016). This underscores an important concern about the unchecked influence of fake news. According to popular media reports, fake news are shared more often than factual news (Silverman 2016). This issue begs an important question of how users of social media can know when to trust the information they read.

Social media platforms started initiatives to reduce the circulation of misinformation. For example, some platforms enable users to report suspicious stories and then employ fact-checking organizations to provide "disputed accuracy" labels (Mosseri 2016). Unfortunately, fact checking relies on the costly efforts of domain experts in political science and journalism. Moreover, manual fact checking is so time consuming that its results may come too late to prevent the cir-

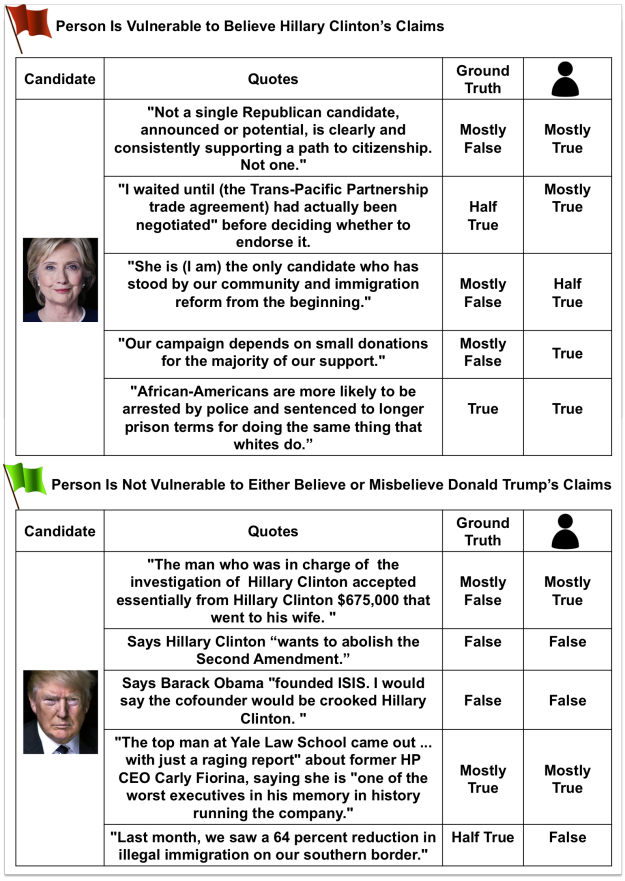

Figure 1: Our goal is to determine whether a person will consistently form a singular belief about the truthfulness of political claims made by a subject of bias, in this example from the 2016 U.S. presidential candidates. In the top example, the person consistently believes quotes from Clinton to be more true than domain experts. In the bottom example, the person does not show a vulnerability to believe quotes from Trump to be consistently more true (or more false) than domain experts. We propose a system, Crowd-O-Meter, which uses measures of a person's explicit bias (e.g., party preference) and implicit bias (i.e., worker behavior when rating the quotes) to predict if (s)he is vulnerable.

culation of fake news, which can spread fast on the internet. Alternatively, state-of-the-art automatic methods for dealing with fake information on social media have proposed fact checking based on linguistic analysis of the stories (e.g., (Mitra, Wright, and Gilbert 2017)) and balancing a user's exposure to factual and fake stories (Farajtabar et al. 2017).

In this work, we instead propose a *user-centered* approach as a first step towards empowering social media platforms to detect users who are vulnerable to misinformation. We propose a method for detecting whether a person is consistent in his/her belief about the truthfulness of political claims. For example, as observed in Figure 1, one person consistently believes claims made by Hillary Clinton to be true even that most of the claims are not actually true. In contrast, Figure 1 also exemplifies a person who is inconsistent in whether (s)he believes quotes from Donald Trump are true. Our proposed method automatically identifies whether a user is consistently biased in his/her beliefs based on a combination of implicit behavioral cues and explicitly shared political views. In other words, our method indicates if a person is vulnerable, whether because (s)he consistently believes the claims are true or because (s)he consistently believes the claims are not true. Our solution can be applied to filter a collection of users to a smaller set in order to more efficiently identify users who are vulnerable to believing false claims.

In this work, we focus on the following three questions:

- Can we train a machine to automatically predict if a crowd worker will harbor a consistent belief about the truthfulness of political claims?
- What features best predict whether a crowd worker will hold a consistent view of the truthfulness of the claims?
- How does the data modality, i.e., text versus video, impact a machine's ability to predict whether a crowd worker will hold a consistent view of the truthfulness of the claims?

In our work, we used statements made during the 2016 U.S. presidential election campaign. We created a new dataset called the "Political Claims Factualness Dataset," for which we collected judgments from U.S. located crowd workers on the factuality of quotes from the front runners Donald Trump and Hillary Clinton. We compared the workers' judgments with those of domain experts. We considered crowd workers' behavioral traces (unconscious) and explicit opinion bias (conscious) separately and jointly to train and test different versions of our Crowd-O-Meter prediction system. Some of our findings are:

- Crowd-O-Meter is up to 26 percent points more accurate than the best we can achieve today (i.e., chance predictions).
- Features that measure implicit bias are typically better predictors than features that measure explicit bias.
- The predictive power of Crowd-O-Meter to detect a person's vulnerability is typically stronger for claims observed in a video than for claims read in text.

We will make the "Political Claims Factualness Dataset" with ground truth and 7 crowd worker annotations per data point available to the research community. The insights we gained about how to measure and predict vulnerability in crowd work, while tested only in the domain of political discourse, have the potential to generalize to other areas.

The rest of the paper is organized as follows. Section II explains the state of the art in three related areas: (1) dealing with fake news on the internet, (2) defining implicit and explicit bias, and (3) handling bias in crowdsourcing. Section III explains our proposed crowdsourcing methodology, including our implicit and explicit bias metrics and the prediction model used by our Crowd-O-Meter system. Section IV discusses our experimental setup and the performance of our prediction system. Sections V and VI discuss potential applications of our system and conclude with future work.

II. Related Work

Fact Checking News on the Internet

We define "fake news" as articles that are verifiably discussing false claims (Allcott and Gentzkow 2017). Due to high adoption of social media in informing U.S. adults about everyday news (Gottfried and Shearer 2016), many concerns are raised about the trustworthiness and credibility of information circulating on social media. Thus, the problem of identifying and dealing with fake news has attracted a lot of attention. The task of fact checking is sensitive to human bias, as is known, for example, from legal court cases. People typically either show very strong support for or against the topic of analysis. Presence of such strong opinions can cloud the critical thinking and decision making (Kang et al. 2012). We summarize the offered solutions to mitigate the presence of fake news as follows:

Manual Intervention: Facebook recently suggested a report-and-flag framework to recognize and remove false stories from the platform (Mosseri 2016). In the "report stage," the community identifies and reports suspicious stories. Reported stories are next sent out to fact checkers to evaluate the credibility and validity of selected stories. In the "flag stage," fake stories are labeled by the "disputed by 3rd party fact-checkers" flag. Users are also notified once they are about to share a disputed story. Such a labor-intensive approach requires domain experts of journalism and political science fields to get involved and, thus, the process is costly and slow. The scheme also can suffer from malicious attacks on real stories, planned by spammers and malicious adversaries (Farajtabar et al. 2017). Our proposed system does not require additional human input. Rather it relies on users' interaction with the platform to learn if they could be advocates of fake news.

Intervention based on Linguistic Content: As the importance of linguistic content has been revealed in many fact-checking studies (Arif et al. 2016; Liao and Shi 2013; Liu, Burton-Jones, and Xu 2014), linguistic features of text have been used to detect controversial text (Mitra, Wright, and Gilbert 2017; Zhao, Resnick, and Mei 2015; Zeng, Starbird, and Spiro 2016). Although such studies could reduce the amount of human effort needed in the fact checking paradigms, they might not generalize to fake stories generated in the future. Spammers could use novel strategies that could potentially outdate current factuality classifiers. With

Crowd-O-Meter, we approach the problem from a different, user-centered perspective. Relying on a users' interaction on a platform to predict how vulnerable (s)he might be in believing false claims could offer a valuable social media strategy to avoid promoting false stories.

Network Activity Intervention using Reinforcement Learning: With the goal to match people's exposure to real news to their exposure to fake news, (Farajtabar et al. 2017) developed a reinforcement learning model that aims to optimize the propagation of the real news through the network. This model requires a network graph of connections (user u follows user v), and makes assumptions of users following each other and immediately reading their shared stories. As noted by the authors themselves, this assumption is not realistic as people may only be offline at certain times and so miss updates from their network. Our method does not make assumptions about a user's network. Each user is studied individually, and analysis of his/her implicit and explicit opinion biases is predictive of his/her reaction to false claims.

Bias

Bias Definition: Psychologists and sociologists have defined bias as a property of people with the following characteristics (Guerra et al. 2011; Walton 1991): (1) A lack of appropriate critical doubt that leads the biased party to lean toward a specific side of an argument instead of assessing the other side in a critically appropriate manner; (2) A lack of proper logic in argumentation; (3) A visible position of the biased party toward a subject (e.g., favoring one subject over others); (4) A personal gain of the party associated with the outcome of an argument.

Bias Types: Two types of bias are widely studied in the literature: implicit and explicit bias (Dovidio, Kawakami, and Gaertner 2002). Implicit bias refers to people's evaluations and assessments that are made without their full awareness or control over the subject and are often automatically activated. A common method to measuring such bias is by using the "Implicit Association Test" (IAT by A. G. Greenwald, D. E. McGhee, & J. L. K. Schwartz,1998), which measures the reaction time that "captures the strength with which social groups [..] are implicitly or automatically associated with good/bad evaluations and other characteristics" (Jost et al. 2009). On the other hand, explicit bias refers to attitudes that are often assessed and collected through a self-report assessment and reflect a person's beliefs and ideology. One study indicated the impact of "implicit bias among physicians, its dissociation from conscious (explicit) bias, and its predictive validity" (Green et al. 2007). Other studies indicated that implicit bias collected through IAT and explicit self-reported bias are systematically related (e.g., (Greenwald, Nosek, and Banaji 2003; Hofmann et al. 2005)). Unlike prior work, our aim is to uncover the potential of implicit and explicit metrics for predicting the bias of people to have a consistent belief about the truthfulness of political claims.

Crowdsourcing a Person's Bias

Measuring Implicit Bias: Our goal is to measure a user's implicit opinion bias via his/her behavioral traces, also called "task fingerprinting." This concept was introduced by (Rzeszotarski and Kittur 2011) to teach a machine to predict the quality of crowd work in the absence of ground truth; e.g., using worker clicks, key presses, mouse scrolling. Since then, multiple studies have employed task fingerprinting to detect poor-quality results based on the user's behavior (Sameki, Gurari, and Betke 2015; Sameki, Gurari, and Betke 2016). A recent study analyzed user's event logs to detect poor-quality crowd and expert work (Kazai and Zitouni 2016). As an application of such efforts, (Birnbaum et al. 2013) used behavioral data to identify interviewer fabrication in surveys by asking annotators to fabricate the data intentionally. Finally, (Dang, Hutson, and Lease 2016) developed a freely-available framework to empower researchers to collect behavioral traces on their tasks while working with Amazon Mechanical Turk. Our work builds on prior work by demonstrating the potential of task fingerprinting in a new application of detecting the bias of a crowd worker to hold a consistent belief about the truthfulness of political claims.

Measuring Explicit Bias: Previous surveying efforts have investigated the general demographic setting of crowdsourcing platforms such as Amazon Mechanical Turk (Paolacci, Chandler, and Ipeirotis 2010; Ross et al. 2010) and the viability of crowdsourcing platforms to collect high quality data via surveys (Behrend et al. 2011). In general, workers appear to be truthful when providing self-report information (Rand 2012; Shapiro, Chandler, and Mueller 2013). Our proposed Crowd-O-Meter system identifies a crowd worker's explicit bias a priori to uncover any conscious opinion biases of a crowd worker to the subject of interest.

III. Methods

We here describe the "Political Claims Factualness Dataset" that we created and our crowdsourcing platform, Crowd-O-Meter, that assesses if a crowd worker is vulnerable to hold a consistent belief about the truthfulness of political claims.

New Factualness Dataset

Fact checking is the task of assessing a public figure's accuracy of claims. Fact checkers are widely employed by news agencies in the process of creation and refinement of news articles (Vlachos and Riedel 2014). Examples are PolitiFact, a Pulitzer-award fact checking platform that analyzes the validity and accuracy of claims by elected officials and others who speak up in American politics, and Full Fact, a British platform for fact checking of U.K. political claims.

We created a database of presidential candidates' public claims using the PolitiFact fact-checking platform. This website compiles public figures' statements and categorizes each statement's accuracy into one of the following categories: *true*, *mostly true*, *half true*, *mostly false*, *false*, and *pants on fire*. The fact checking process is as follows. Staffers first collect provocative and questionable statements they hear or read. A group of experts next investigate the factualness of each claim by collecting additional data about the claim. The experts then suggest a rating and add the quote to their Truth-O-Meter tool along with a list of supporting resources to help readers judge whether they agree with the

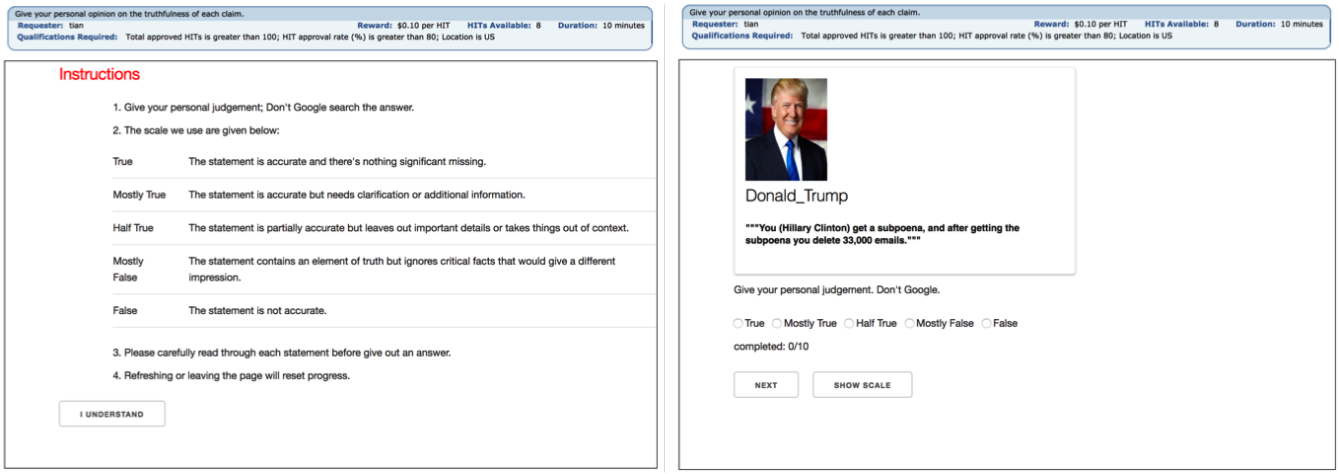

Figure 2: Crowd-O-Meter-Text interface with the annotation instructions (left) and task interface (right).

ruling or not. Our database includes quotes found in PolitiFact in two data modalities: text and video.

TextModal: We crawled PolitiFact for the claims made by US presidential candidates Donald Trump and Hillary Clinton and selected 290 claims by each candidate. Out of the 290 extracted claims from Donald Trump, 180 claims were not factual, i.e., either false or mostly false. Moreover, 82 of the 290 claims by Clinton were not factual as well. This sampling accurately reflects the natural balance of truthfulness of quotes on the PolitiFact website.

VideoModal: We next curated televised video clips showing the candidates Donald Trump and Hillary Clinton when they stated a subset of the quotes included in *TextModal*. Collection of the video dataset was motivated in part because prior work has suggested televised images have an impact on people's perception of presidential debates (Druckman 2003; Kraus 1996). We extracted 20 video clips in which the PolitiFact's extracted quotes are spoken by the candidate. 10 of the 20 claims in the videos were not factual, i.e., either false or mostly false. The videos evenly represent the appearance of Donald Trump and Hillary Clinton.

Predicting Crowd Worker Vulnerability

Crowdsourcing Task Design: Our crowdsourcing platform guides a crowd worker through a two-step process, first showing him/her instructions, and then the interface with the task (Figure 2). A crowd worker's task is defined for him/her to provide a judgment about the factualness of 10 quotes from presidential candidates Donald Trump or Hillary Clinton, 5 quotes per candidate. Due to the fact that crowd workers could leave our platform after performing only one task, we grouped 10 quotes into one task in order to make sure we have enough signals per user to evaluate him/her across quotes. This design choice provided us with at least 10 results per crowd worker.

We chose a multi-page interface where each question is presented to the crowd worker on a different page for the following reasons: first, by presenting one quote per page, we ensure that the collected behavioral traces only reflect that quote. Second, Crowd-O-Meter randomizes the questions to eliminate the potential bias that might happen among workers, due to specific ordering of presented quotes.

Crowd workers' judgments on the factualness of each quote are collected from a five-point Likert scale. Likert scales are widely adopted in many research areas, including social science (Garland 1991). Our system matches the five-point Likert scale shown to the crowd worker to five definitions used by the *PolitiFact* fact checking platform:

- TRUE: The statement is accurate and there is nothing significant missing.
- MOSTLY TRUE: The statement is accurate but needs clarification or additional information.
- HALF TRUE: The statement is partially accurate but leaves out important details or takes things out of context.
- MOSTLY FALSE: The statement contains an element of truth but ignores critical facts that would give a different impression.
- FALSE: The statement is not accurate or makes a ridiculous claim.

Establishing the True Vulnerability of a Crowd Worker: Our measure for indicating if a user is consistently biased to a belief about the truthfulness of quotes is based on two factors: subject of bias (e.g., Donald Trump) and the user's position toward that subject (e.g., a user might have strong negative bias toward a subject and be completely neutral toward another subject). We propose a metric, *Valence Diff*, that characterizes a user's vulnerability to consistently believe (or not believe) claims from each subject separately.

Our metric *Valence Diff* aims to accumulate over the direction of worker w's annotation bias for all of the quotes (s)he has annotated about the subject of bias *SB*. It measures the absolute difference between two numbers: (1) Positive-Valence: the number of quotes from a subject of bias *SB* that are rated more positively by worker w than the actual rating they received from domain experts (ground truth) and

(2) NegativeValence: the number of quotes from a subject of bias *SB* that worker w rated more negatively than the ground truth. We hypothesize that the absolute difference between PositiveValence and NegativeValence for worker w can uncover if there is a solid position that worker w holds toward that subject of bias. A small absolute value can be an indicator of lack of bias and thus, lack of vulnerability in (positively or negatively) over-reacting to the quote. A large absolute difference, in contrast, can identify a strong position of worker w toward that subject of bias. For instance, a worker w may be consistently more trusting than what reality (expert-curated ground truth) may support. The *Valence Diff* definition follows our intuition that users with a strong opinion for or against a subject might be more vulnerable in believing false positive or negative claims from that subject.

In order to assign a vulnerable versus non-vulnerable label to a crowd worker, we set a threshold of 40% of all annotated quotes by the worker to focus on cases with a strong bias. If the absolute difference exceeded the threshold, we assigned a "vulnerable" label to the annotator. Otherwise, a worker was classified as "not vulnerable". We computed this value per individual worker per subject of bias. While we will show our threshold value (40%) is successful in our experiments, a valuable area for future work is investigating what is the optimal threshold.[1]

Features Characterizing a Person's Implicit Bias: We recorded each worker's unique ID along with his/her measured behavioral traces. Each recorded entry consists of an event type, a millisecond-precision time stamp, an instance (quote) identifier, and a worker identifier. From these logs, our system extracts the following task-level features that reflect how the crowd worker interacted with the platform:

- *Time per Question.* As done in prior work (Vijayanarasimhan and Grauman 2009; Carlier et al. 2014; Rzeszotarski and Kittur 2011), we measure the time per task. Our system incorporates the time a crowd worker reads and thinks about a quote and answers a question about it as a feature in the prediction model.
- *Time to First Response.* We also captured the time between revealing the quote and the selection of the first response on the Likert scale.
- *Total Time per Subject of Bias*. We captured two separate time points per subject of bias, which reflected how long the worker spent on all quotes of a specific subject of bias (e.g., time spent for answering fact checking questions for Donald Trump versus time spent for answering questions for Hillary Clinton). Our motivation comes from the hypothesis that implicit bias might lead the workers to be less critical of a subject of interest's quotes and thus, they might quickly identify his/her quotes as true.
- *Normalized Time per Question.* Text characteristics such as text length (e.g., word count) could affect a crowd worker's time spent per task. A worker may require more time to read and analyze a long and complicated text. We use a metric that normalizes against the effect of word count: Time-per-Task/Word-Count. For tasks with videos as input, we captured total time per question from the moment that the video stopped playing, and thus, we removed this normalized measure.
- *Answer Switch.* Inspired by prior work (Rzeszotarski and Kittur 2011), we designed our system to register workers' mouse clicks on the 5-point Likert scale options of the interface. It thus captures the number of times a user switches answers among the five options.
- Hover Time. Our system also captures how much time a worker spent around each of the five answer options of the Likert scale (e.g., hover time on *true* option).

Features Characterizing a Person's Explicit Bias: As described earlier, explicit bias refers to attitudes that are often assessed and collected through a self-report questionnaire and reflect a person's beliefs and ideology. We hypothesize that a questionnaire could measure crowd workers' explicit preferences by creating an experience in which they have to make conscious choices. We designed a post-test questionnaire to capture individual MTurk worker's explicit preferences on contradicting subjects. Our survey questions covered the following criteria:

- *Personal:* Crowd workers were asked about their age group, gender, and education level.
- *Political Party Affiliation and Interest:* We included questions about whether they supported (voted for) a presidential candidate or not. We asked the question (Gallup 2016): "In politics, as of today, what do you consider yourself: a Republican, a Democrat, or an Independent?" *Political Knowledge:* Our system gauges MTurk workers' political knowledge based on responses to a 5-point Likert scale ("strongly disagree" to "strongly agree") to the following statements: "I followed the U.S. presidential election," "I paid close attention to the 2016 U.S. presidential campaigns," "I know a lot about the Democratic party," and "I know a lot about the Republican party."
- *Media Use:* To measure media use, our system asked MTurk workers to indicate their frequency of reading, watching, or listening to the news on a 5-point scale ranging from "less than once per week" to "more than once per day," with "once per week," "3–5 times per week," and "once per day" as other options (adapted from (Eveland et al. 2005)). Our system also collected their judgments on the use of social media based on responses to a 5-point Likert scale ("strongly disagree" to "strongly agree") to the following statements: "I follow most of my news from social media (i.e. Twitter, Facebook, etc.). "

Prediction System: Crowd-O-Meter uses as input the features that model a specific crowd worker's implicit and explicit bias and outputs a prediction about whether the crowd

[1]Future work could also examine how to distinguish a person who incorrectly assigns an equal number of negative and positive judgements with random variation from a person who has domain expertise (e.g., journalism domain expert) and so correctly rates the truthfulness of every claim every time.

worker holds a consistent belief. A random forest classification model is employed for the prediction model. Crowd-O-Meter classifies a crowd worker as "vulnerable" or "not vulnerable" to the subject of bias. In our application, the subject may be the presidential candidate Donald Trump or presidential candidate Hillary Clinton. The class "vulnerable" represents either the possibility that a crowd worker is positively or negatively influenced by his/her bias. This means, for example, if a crowd worker is vulnerable to believe Trump then the worker is more likely to believe a false quote from Trump.

We implemented our random forest classifier with Python's Scikit-Learn library using 10 trees. Once trained, this prediction model learns the unique weighted combination of the aforementioned implicit and explicit bias features that is predictive of whether a worker is vulnerable. The resulting trained random forest model implicitly reveals the importance of each feature in making a prediction. In particular, for each of the decision trees, a feature is rated as more informative if it is closer to the top of the tree. This is because, during training, it was selected as yielding the greatest measured information gain across the training examples from the remaining features. Consequently, each feature's importance is its average importance across the 10 decision trees in our random forest classifier.

IV. Experiments

We now describe our studies to evaluate the predictive power of the Crowd-O-Meter to uncover a crowd worker's vulnerability from worker's implicit and explicit biases. We addressed the following Research Questions (RQ):

- RQ1: Can we train a machine to automatically predict a crowd worker's vulnerability to consistently believe (or not believe) political claims?
- RQ2: What features best predict a crowd worker's vulnerability?
- RQ3: How does the data modality, i.e., text versus video, impact a crowd worker's vulnerability?

Annotation Tool Settings: We chose the Amazon Mechanical Turk (AMT) marketplace to recruit crowd workers, knowing that prior research showed the presence of a diverse pool of contributors with different political affiliations (Huff and Tingley 2015; Levay, Freese, and Druckman 2016). We only accepted AMT workers who live in the United States, had previously completed at least 100 tasks, and maintained an approval rating of at least 80%. We accepted and compensated all crowd workers who participated in our tasks. To collect annotations from crowd workers, we embedded a task hosted on our private Amazon Web Services (AWS) server into the AMT framework as an external task.

Experimental Details: We collected annotations from crowd workers for a task that contains 10 factual statements; 5 of the 10 statements are quotes from Donald Trump and the other 5 are quotes from Hillary Clinton. To capture the variability of crowd behaviors that may arise due to workers with differing implicit and explicit biases, we collected five annotations on each quote. We posted all tasks simultaneously while randomizing the order of quotes in each task. We alloted a maximum of ten minutes to complete each HIT and paid $0.10 per HIT. Each quote mentioned the name of the candidate who made the claim (e.g., Donald Trump: "The Obama Administration agreed to take thousands of illegal immigrants from Australia."). We hypothesized that such association would trigger a conscious or unconcious opinion bias in workers who are pro/anti a candidate. For tasks with video input, the video is first shown to crowd workers and upon completion of the video, the question on factualness on the mentioned claim is enabled. In total, our dataset included 2,900 and 1,000 crowdsourced labels for text and video respectively.

Crowd-O-Meter Classification Performance: We first evaluated Crowd-O-Meter on our TextModal dataset, which includes 580 text quotes from presidential candidates (290 per candidate). We processed the 2,900 crowdsourced results to create a label of "vulnerable" or "not" for the 59 unique workers who created the text labels. We also collected the implicit and explicit bias features for the 59 crowd workers via our crowdsourcing platform. Finally, we used 5-fold cross validation to train and test our prediction system. We enriched our analysis of our Crowd-O-Meter system by analyzing implicit and explicit features jointly as well as separately. We conducted the latter analysis to augment our analysis of what are the key factors that are most predictive of whether a worker is vulnerable to be consistently biased.

To the best of our knowledge, no prior work has addressed predicting a user's vulnerability to consistently believe (or not believe) political claims via crowdsourcing. Consequently, the best a system can achieve today is to randomly decide if a worker would be vulnerable. For this reason, we compare our systems to a *Chance* baseline which returns a random class label per worker toward each specific *subject of bias* (i.e., Hillary Clinton and Donald Trump).

We evaluated and compared our Crowd-O-Meter systems and the *Chance* baseline by generating precision-recall curves using each prediction method's confidence (Figure 3; Donald Trump, left; Hillary Clinton, right). We also calculated the average precision (AP) for each prediction method. As observed, our proposed system that employs *All Features* yields a large improvement compared to the *Chance* baseline; e.g., the AP score improves by 26 percentage points (0.58 to 0.84) for quotes by Trump and 14 percentage points (0.51 to 0.65) for quotes by Clinton. Despite the significant variety of quote topics and differences in candidates, our Crowd-O-Meter systems produce quite accurate results.

We next investigated what makes our prediction system successful by evaluating the predictive power of our Crowd-O-Meter system when it is trained and tested on implicit features and explicit features separately (Figure 3; purple and red curves respectively). As observed, both implicit and explicit features outperform the *Chance* baseline. Such improvements suggest that both unconscious (implicit features) and conscious (explicit features) cues are effective in-

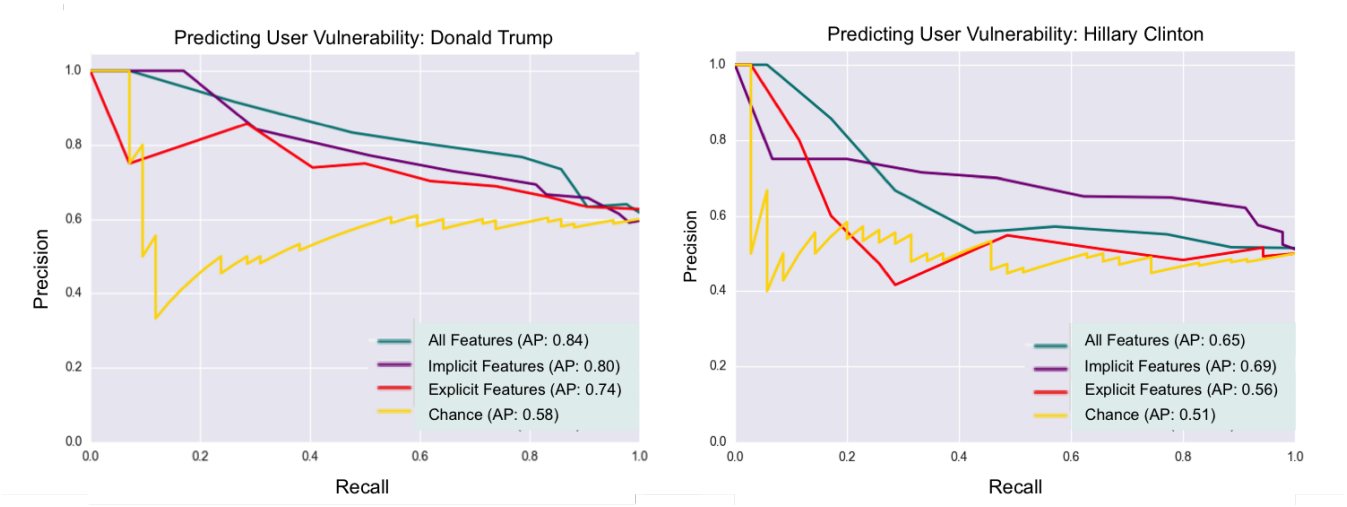

Figure 3: Shown are precision-recall curves and average precision (AP) scores for all prediction systems; Trump (left), Clinton (right). Our Crowd-O-Meter system (i.e., *All Features*) outperforms today's status quo approach (i.e., *Chance* predictions) by up to 26 percentage points (i.e., Trump), showing the promise of our proposed novel task to model human vulnerability. Our results also highlight that a user's unconscious behavioral data (purple curve) is a better predictor than relying on a user's explicitly shared information (red curve) alone.

dicators of user vulnerability. Interestingly, we observe that relying on implicit features alone yields a greater predictive performance for our proposed task over relying on explicit cues alone. The predictive power of implicit features is exciting because unconscious traces (behavioral cues) are much easier to obtain than requiring a user to opt in and share explicit information about himself/herself.

We also analyzed the resulting random forest prediction models to uncover which specific features are most predictive. We found for the prediction model learned from all cues that a crowd worker's "hover time over the (mostly) false Likert scale options" was most predictive; i.e., 20% and 12% of the predictive power came from this feature for the Trump and Clinton quotes respectively. When examining the most predictive features for the model trained on the implicit features alone, we found the most informative features for both Trump and Clinton quotes were user's "hover time" over the five Likert scale options. These findings enrich our previous finding that implicit features have stronger predictive power than explicit features. We suspect the hover time is most predictive because the crowd workers tended to harbor a strong belief about the truthfulness of each quote before even reflecting on the content of the quote.

Input Modality: Text versus Video: We next investigated how changing the data modality (i.e., text versus video) affected the predictive power of Crowd-O-Meter to predict crowd worker's vulnerability. We conducted two experiments using the same quotes for both experiments. One experiment used as input to the crowdsourcing platform the textual quotes (Crowd-O-Meter-Text). The second experiment used as input to the crowdsourcing system the original video clips from the presidential debate where the quote was stated (Crowd-O-Meter-Video). Two groups of independent crowd workers were recruited to take part in each experiment. For each experiment, we recruited 100 workers to complete each task resulting in a total of 1,000 annotations (10 videos/quotes per candidate x 100 workers) per candidate for each data modality (text and video).

As in the previous study, we compared the predictive power of Crowd-O-Meter-Text and Crowd-O-Meter-Video using different combinations of features (Table 1; all, implicit, explicit). We evaluated with respect to four different evaluation metrics. We found that across all evaluation metrics and all features, Crowd-O-Meter typically yielded better predictions for the video modality than the text modality. This finding aligns with previous findings about the impact of televised images on viewers (Kraus 1996) and suggests that watching a subject of bias might trigger more conscious or unconscious personal biases. Among the two candidates, Crowd-O-Meter led to the best prediction using the combination of implicit and explicit features for Donald Trump videos and using explicit features for Hillary Clinton videos. This suggests that collecting both conscious and unconscious features can help to uncover vulnerability.

We again analyzed the resulting random forest prediction models to uncover which specific features are most predictive. We found that 23% of the predictive power comes from the explicitly stated support for the candidate for the Trump quotes and 16% of predictive power comes from the average time to respond for Clinton quotes. When examining the most predictive features for the model trained on the implicit features alone, we found the most informative features for both Trump and Clinton quotes was the "average time to the first response". We hypothesize that predictions are on average better for video than text because the crowd workers experience stronger implicit reactions when observing a person's appearance and gestures.

	PolitiFact Text Dataset (# Annotations):	**Donald Trump** (1,000)		**Hillary Clinton** (1,000)	
Features	Input Modality:	Text	Video	Text	Video
a. All Features	Average Precision (AP)	0.78	**0.79**	0.52	**0.64**
	True Positive Rate (TPR)	**0.71**	0.67	0.46	**0.57**
	True Negative Rate (TNR)	0.71	**0.93**	0.6	**0.85**
	Accuracy (ACC)	0.71	**0.81**	0.53	**0.71**
b. Implicit Features Only	Average Precision (AP)	**0.84**	0.57	**0.71**	0.56
	True Positive Rate (TPR)	0.73	**0.86**	0.62	**0.66**
	True Negative Rate (TNR)	0.61	**0.84**	**0.79**	0.70
	Accuracy (ACC)	0.68	**0.85**	**0.70**	0.68
c. Explicit Features Only	Average Precision (AP)	0.75	**0.77**	0.48	**0.71**
	True Positive Rate (TPR)	0.69	**0.81**	0.47	**0.67**
	True Negative Rate (TNR)	0.60	**0.78**	0.66	**0.85**
	Accuracy (ACC)	0.66	**0.80**	0.57	**0.77**

Table 1: Evaluation and comparison of Crowd-O-Meter-Text and Crowd-O-Meter-Video. The combination of implicit and explicit features typically led to better predictions for video than text. This demonstrates that implicit and explicit features were better indicators of vulnerability when users were exposed to televised videos.

V. Discussion

Quality Control in Crowdsourcing. Our work highlights how to detect a crowd worker's vulnerability to harbor a consistent bias. Such information can be valuable in future crowdsourcing experiments when (1) recruiting crowd workers as well (2) evaluating the quality of a crowd worker's judgements. Moreover, we found that implicit features were typically the most informative features for predicting worker vulnerability. Consequently, our work highlights the promise of discovering whether a crowd worker holds a consistent bias without the need to change the crowdsourcing task itself.

Connection to a Social Science Methodology. The Implicit Association Test (IAT) has been widely used by social scientists to measure the strength of associations between concepts (e.g., human race, sexual orientation) and evaluations (e.g., good, bad, pleasant, unpleasant) or stereotypes (e.g., athletic, violent, peaceful). IAT systems measure a user's reaction time when responding to a series of pre-defined questions about an explicit bias topic. While Crowd-O-Meter can similarly uncover a crowd worker's unconscious beliefs via implicit cues, our proposed methodology instead makes predictions based on a large variety of implicit behavioral cues without directly asking questions about a person's bias to the subject. We offer our proposed proposed approach as an alternative for uncovering a person's implicit beliefs.

Potential Impact of our Results on Detecting and Mitigating for User Bias on the Open Web. While the process of promoting fake information in social media might depend on many factors, our findings provide promising evidence that incorporating user-centered implicit and explicit characteristics into a learning system could potentially empower social media platforms to characterize users based on their vulnerabilities to believe (or not believe) information. The features we suggest in this study can potentially be generalized to a social media settings. Tracking individual users on social media has become easy with assorted advertising systems. Often users log into different websites to receive a service by using their social media credentials (e.g., Facebook or Google accounts). People often willingly give up information about themselves in exchange for online services. Social media platforms can use explicit cues (e.g., self-reported profiles often indicate a user's political views, age, gender, etc.) and/or implicit behavior (user's browsing behaviors on web pages) to monitor or approve what users can re-post or assign different weight to their interaction within the platform. By analyzing information about a user, a social network platform could estimate this user's opinion bias, and then group users into cohorts with similar behavior. These cohorts could then be used to train algorithms to predict users' patterns of interacting with news articles. Based on these predictions, users' abilities to spread news on the platform could be adjusted. Interactions of users whose actions are predicted to be influenced by a strong opinion bias would receive a different weight than users that are predicted to not harbor a consistent bias toward a subject. A valuable area for future work is to examine the robustness of the proposed idea with a larger number of users to more closely emulate the situation on the open web.

VI. Conclusion

We propose the novel problem of predicting a crowd worker's vulnerability to consistently hold a belief about the truthfulness of political claims. Our proposed Crowd-O-Meter system for this task makes predictions using both implicit and explicit opinion bias cues. Our experiments show our top-performing system can outperform today's status quo approach by 26 percentage points in prediction accuracy. We offer our system as a promising starting point towards the problem of mitigating the impact of false news.

Acknowledgments

The authors would like to thank the Rafik B. Hariri Institute for Computing and Computational Science & Engineering at Boston University for supporting this research and the crowdsourced workers for participating in our experiments.

References

[Allcott and Gentzkow 2017] Allcott, H., and Gentzkow, M. 2017. Social media and fake news in the 2016 election. Technical report, Technical report, National Bureau of Economic Research.

[Arif et al. 2016] Arif, A.; Shanahan, K.; Chou, F.; Dosouto, Y.; Starbird, K.; and Spiro, E. S. 2016. How information snowballs: Exploring the role of exposure in online rumor propagation. In *Proceedings of the 19th Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 466–477. ACM.

[Behrend et al. 2011] Behrend, T.; Sharek, D.; Meade, A.; and Wiebe, E. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods* 43(3):1–14.

[Birnbaum et al. 2013] Birnbaum, B.; Borriello, G.; Flaxman, A. D.; DeRenzi, B.; and Karlin, A. R. 2013. Using behavioral data to identify interviewer fabrication in surveys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 2911–2920. ACM. 10 pp.

[Carlier et al. 2014] Carlier, A.; Charvillat, V.; Salvador, A.; Giro-i Nieto, X.; and Marques, O. 2014. Click'n'Cut: Crowdsourced interactive segmentation with object candidates. In *Proceedings of the International Workshop on Crowdsourcing for Multimedia*, 53–56. ACM.

[Dang, Hutson, and Lease 2016] Dang, B.; Hutson, M.; and Lease, M. 2016. Mmmturkey: A crowdsourcing framework for deploying tasks and recording worker behavior on amazon mechanical turk. In *arXiv preprint arXiv:1609.00945*.

[Dovidio, Kawakami, and Gaertner 2002] Dovidio, J. F.; Kawakami, K.; and Gaertner, S. L. 2002. Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology* 62–68.

[Druckman 2003] Druckman, J. N. 2003. The power of television images: The first Kennedy-Nixon debate revisited. *The Journal of Politics* 65(2):559–571.

[Eveland et al. 2005] Eveland, W. P.; Hayes, A. F.; Shah, D. V.; and Kwak, N. 2005. Understanding the relationship between communication and political knowledge: A model comparison approach using panel data. *Political Communication* 22(4):423–446.

[Farajtabar et al. 2017] Farajtabar, M.; Yang, J.; Ye, X.; Xu, H.; Trivedi, R.; Khalil, E.; Li, S.; Song, L.; and Zha, H. 2017. Fake news mitigation via point process based intervention. In *arXiv preprint arXiv:1703.07823*.

[Gallup 2016] Gallup. 2016. Gallup, Party affiliation. http://www.gallup.com/poll/15370/party-affiliation.aspx.

[Garland 1991] Garland, R. 1991. The mid-point on a rating scale: Is it desirable? *Marketing Bulletin* (2):66–70.

[Gottfried and Shearer 2016] Gottfried, J., and Shearer, E. 2016. News use across social media platforms 2016. Pew Research Center.

[Green et al. 2007] Green, A. R.; Carney, D. R.; Pallin, D. J.; Ngo, L. H.; Raymond, K. L.; Iezzoni, L. I.; and Banaji, M. R. 2007. Implicit bias among physicians and its prediction of Thrombolysis decisions for black and white patients. *Journal of General Internal Medicine* 22(9):1231–1238.

[Greenwald, Nosek, and Banaji 2003] Greenwald, A. G.; Nosek, B. A.; and Banaji, M. R. 2003. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology* 85(2):197–216.

[Guerra et al. 2011] Guerra, P. H. C.; Veloso, A.; W. Meira, J.; and Almeida, V. 2011. From bias to opinion: A transfer-learning approach to real-time sentiment analysis. In *Proceedings of the Seventeenth ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 150–158. ACM.

[Hofmann et al. 2005] Hofmann, W.; Gawronski, B.; Gschwendner, T.; Le, H. T.; and Schmitt, M. 2005. A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality and Social Psychology Bulletin* 31(10):1369–1385.

[Huff and Tingley 2015] Huff, C., and Tingley, D. 2015. Who are these people? evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research and Politics* 2(3):1–12.

[Jost et al. 2009] Jost, J. T.; Rudman, L. A.; Blair, I. V.; Carney, D. R.; Dasgupta, N.; Glaser, J.; and Hardin, C. D. 2009. The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore. *Research in Organizational Behavior* (29):39–69.

[Kang et al. 2012] Kang, J.; Bennett, M.; Carbado, D.; Casey, P.; Dasgupta, N.; Faigman, D.; Godsil, R.; Greenwald, A.; and Mnookin, J. 2012. Implicit bias in the courtroom. *UCLA Law Review*.

[Kazai and Zitouni 2016] Kazai, G., and Zitouni, I. 2016. Quality management in crowdsourcing using gold judges behavior. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 267–276. ACM.

[Kraus 1996] Kraus, S. 1996. Winners of the first 1960 televised presidential debate between Kennedy and Nixon. *The Journal of Communication* 46(4):78–96.

[Levay, Freese, and Druckman 2016] Levay, K. E.; Freese, J.; and Druckman, J. N. 2016. The demographic and political composition of Mechanical Turk samples. *SAGE Open 6* (1):1–17.

[Liao and Shi 2013] Liao, Q., and Shi, L. 2013. She gets a sports car from our donation: rumor transmission in a Chinese microblogging community. In *Proceedings of the 16th Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 587–598. ACM.

[Liu, Burton-Jones, and Xu 2014] Liu, F.; Burton-Jones, A.; and Xu, D. 2014. Rumor on social media in disasters: Extending transmission to retransmission. In *Proceedings of the Pacific Asia Conference on Information Systems*.

[Mitra, Wright, and Gilbert 2017] Mitra, T.; Wright, G.; and Gilbert, E. 2017. A parsimonious language model of social media credibility across disparate events. In *Proceedings*

of the 20th Conference on Computer-Supported Cooperative Work and Social Computing (CSCW). 8 pp.

[Mosseri 2016] Mosseri, A. 2016. News feed fyi: Addressing hoaxes and fake news. http://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

[Paolacci, Chandler, and Ipeirotis 2010] Paolacci, G.; Chandler, J.; and Ipeirotis, P. G. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making* 5:411–419.

[Rand 2012] Rand, D. G. 2012. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology* (299):172–179.

[Ross et al. 2010] Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; and Tomlinson, B. 2010. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI10 Extended Abstracts on Human Factors in Computing Systems*.

[Rzeszotarski and Kittur 2011] Rzeszotarski, J. M., and Kittur, A. 2011. Instrumenting the crowd: Using implicit behavioral measures to predict task performance. In *Proceedings of the Twenty-Fourth Annual ACM symposium on User interface software and technology*, 13–22. ACM.

[Sameki, Gurari, and Betke 2015] Sameki, M.; Gurari, D.; and Betke, M. 2015. Predicting quality of crowdsourced image segmentations from crowd behavior. In *Proceedings of the Third AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*. 2 pp.

[Sameki, Gurari, and Betke 2016] Sameki, M.; Gurari, D.; and Betke, M. 2016. ICORD: Intelligent collection of redundant data - a dynamic system for crowdsourcing cell segmentations accurately and efficiently. In *Proceeding of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 10 pp.

[Shapiro, Chandler, and Mueller 2013] Shapiro, D. N.; Chandler, J.; and Mueller, P. A. 2013. The promise of mechanical turk: How online labor markets can help theorists run behavioral experiments. *Clinical Psychological Science* (1):213–220.

[Silverman 2016] Silverman, C. 2016. This analysis shows how viral fake election news stories outperformed real news on facebook. https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook.

[Vijayanarasimhan and Grauman 2009] Vijayanarasimhan, S., and Grauman, K. 2009. What's it going to cost you? Predicting effort vs. informativeness for multi-label image annotations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2262–2269.

[Vlachos and Riedel 2014] Vlachos, A., and Riedel, S. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science (ACL)*, 18–22.

[Walton 1991] Walton, D. 1991. Bias, critical doubt, and fallacies. *Argumentation and Advocacy* 28:1–22.

[Zeng, Starbird, and Spiro 2016] Zeng, L.; Starbird, K.; and Spiro, E. S. 2016. Unconfirmed: Classifying rumor stance in crisis-related social media messages. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM)*. 4 pp.

[Zhao, Resnick, and Mei 2015] Zhao, Z.; Resnick, P.; and Mei, Q. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the International World Wide Web Conference (WWW)*, 1395–1405. ACM.