# Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking

William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A. Tucker[1]

**Abstract.**
Reducing the spread of false news remains a challenge for social media platforms, as the current strategy of using third-party fact-checkers lacks the capacity to address both the scale and speed of misinformation diffusion. Research on the "wisdom of the crowds" suggests one possible solution: aggregating the evaluations of ordinary users to assess the veracity of information. In this study, we investigate the effectiveness of a scalable model for real-time crowdsourced fact-checking. We select 135 popular news stories and have them evaluated by both ordinary individuals and professional fact-checkers within 72 hours of publication, producing 12,883 individual evaluations. Although we find that machine learning-based models using the crowd perform better at identifying false news than simple aggregation rules, our results suggest that neither approach is able to perform at the level of professional fact-checkers. Additionally, both methods perform best when using evaluations only from survey respondents with high political knowledge, suggesting reason for caution for crowdsourced models that rely on a representative sample of the population. Overall, our analyses reveal that while crowd-based systems provide some information on news quality, they are nonetheless limited—and have significant variation—in their ability to identify false news.

1. To whom correspondences should be addressed: joshua.tucker@nyu.edu. WG performed the statistical analyses for the paper and created the tables and figures. ZS and WG wrote the first draft of the manuscript. KA oversaw the distribution of articles to the crowd and to the professional fact-checkers. ZS oversaw recruitment of professional fact-checkers. ZS, WG, KA, JN, RB, and JAT designed the research and revised the draft of the manuscript. All of the authors contributed to the overall research design of the article pipeline.



doi:10.54501/jots.v1i1.15

*"The issue here is there aren't enough [fact-checkers]... If you get enough data points from within the community of people reasonably looking at something and assessing it over time, then the question is: can you compound that together into something that is a strong enough signal that we can then use that?"* — Mark Zuckerberg, CEO of Facebook (Mark Zuckerberg 2019)

## 1 Introduction

One key obstacle to curbing the spread of fake news online is identifying fake news articles accurately and quickly. The volume of news—both true and false, unbiased and misleading—is so great that simply classifying false or misleading articles in a timely manner poses an immense challenge. Indeed, Facebook has 1.91 billion active users globally (Facebook 2021), relative to just a few hundred global fact-checking organizations (Bell 2019; Duke Reporters' Lab 2016). Most fake news stories have completed their circulation on social media within days after publication (Vosoughi, Roy, and Aral 2018), leaving little time for professional fact-checkers (PFCs) and traditional media outlets to effectively address fake news at the scale and speed of the online information ecosystem. These challenges remain acute in countries with robust fact-checking networks, to say nothing of contexts where professional fact-checking systems are relatively underdeveloped and underresourced (Haque et al. 2018).

To overcome these challenges, Facebook and Twitter have suggested the possibility of crowdsourcing fact-checking—using groups of ordinary users to assess the veracity of news articles (Mark Zuckerberg 2019; Collins 2020). Notably, Twitter has begun experimentation with Birdwatch, a community-based system for users to add additional information to either corroborate or correct a tweet (Coleman 2021). While this approach is promising due to the scale of these platform's user bases and has the normative appeal of including the user community in moderation decisions, the effectiveness of this method in real time is unknown.[2]

In this manuscript we assess whether fact-checking in real time can be effectively accomplished by crowdsourcing evaluations of news articles from groups of ordinary people. Using a novel, preregistered, replicable, and transparent mechanism for selecting popular news stories within 24 hours of their publication, we investigate whether responses from survey participants are able to match the article-level evaluations from a panel of PFCs. To do so, we evaluate two distinct choices that need to be made when setting up crowdsourced fact-checking systems: aggregating the crowd using simple rules or machine learning (ML); and relying on the general population ("random crowds") or limiting crowds to people with attributes that are expected to be related to increased accuracy in identifying the veracity of news ("select crowds"). To the best of our knowledge, this study is the first public-facing research to assess whether crowdsourced fact-checking in real time might be a viable option for identifying the veracity of news.

Our investigation yields three primary findings. First, no approach based on the evaluations of the crowd, regardless of its sophistication, yields particularly high accuracy relative to a PFC. Indeed, no method can reliably distinguish between true and false news at rates most observers would consider reliable, and all methods demonstrate high levels of false positives (i.e., evaluating an article to be false that is not false). Given that false articles are often algorithmically demoted on social media platforms, this rate of false positives has the potential to restrict acceptable content. Second, combining crowds with machine learning algorithms typically improves performance across the

2. Allen et al. 2021 has explored crowdsourcing methods, but not using real-time data. See the Literature Review section for more discussion of the differences in study designs.

board relative to using crowds with simple rules, and this approach shows particular promise when identifying news that is identified as *not* false. Third, when using both simple rules and machine learning methods, select crowds, particularly those with high political knowledge, are superior to random crowds.

## 2 Literature Review

Current approaches from both platforms and scholars to identify false news remain limited. Although there is evidence that algorithms can assist in selecting articles with a high probability of being false (and thus used to automatically send content to PFCs) (Kim et al. 2018), platforms are still unable to keep pace with the scale and velocity of information diffusion through social networks (Horwitz 2020). Additionally, while source credibility is used as a proxy for article veracity in a number of studies (e.g., Grinberg et al. 2019; Guess, Nagler, and Tucker 2019; Allen et al. 2020), Allcott and Gentzkow (2017) estimate that only roughly half of the articles from low-credibility domains are indeed false, thus raising the spectre of unacceptably high numbers of false positives from a domain-based algorithm. A domain-based identification strategy will also fail at identifying misinformation from novel news sources. At the same time, there is evidence that identifying false or misleading content can lead to a meaningful decrease in its impact, as tagging an article as false or misleading noticeably decreases the likelihood an individual will believe it (Pennycook et al. 2017).

A crowdsourced approach might be appealing for three reasons. First, given the size of these platforms' user bases (Perrin and Anderson 2020), this approach provides tremendous scaling potential and could significantly increase the number of news articles that can be fact-checked, especially in countries without developed professional fact checking organizations. Second, a large literature on collective intelligence suggests that aggregated estimates from groups of ordinary people converge on accurate judgements, even if the average individual estimate in the group is not especially accurate (Golub and Jackson 2010; Surowiecki 2005; Woolley et al. 2010; Woolley, Aggarwal, and Malone 2015). Extending this research to whether or not the "wisdom of the crowds" can be applied to fact-checking news in real time is therefore also a compelling scientific question. Third, there are potential normative benefits from moving from a world where platforms employ non-representative specialists to classify news as true or false to one where users of the platform make decisions that are aggregated using transparent rules (Coleman 2021).

A vast and growing literature on the power of the "wisdom of the crowds," or collective intelligence, provides evidence for the efficacy of aggregating non-expert responses in domains ranging from forecasting political and economic events (Budescu and Chen 2015; Griffiths and Tenenbaum 2006) to predicting sporting outcomes and weather trends (Herzog and Hertwig 2011; Hueffer et al. 2013). Recently, the utility of crowdsourcing has also gained momentum in a number of scientific fields, including medicine (Tucker et al. 2019), geology (Comber et al. 2016), and astronomy (Raddick et al. 2010). However, crowdsourced or metapredictors have their limits, and performance varies significantly across different categories of tasks (Simoiu et al. 2019). It remains unclear whether fact-checking is a domain where the crowd can be effectively employed to identify misinformation and curb its spread.

Allen et al. (2021) explore the potential for a crowd-based method for fact-checking, and their findings offer encouraging evidence as to the crowd's ability to fact-check news, especially when asked to evaluate headlines rated true by PFCs. While their study provides useful evidence for the crowd's ability to evaluate news headlines, their design

is structured around three important features that are worth noting. First, a significant portion of their sample of articles, which were provided by Facebook, were months or years old by the time they were included in the study. It is unclear to what extent their findings generalize to the evaluation of articles immediately after publication—the period during which news articles are most likely to be seen (Vosoughi, Roy, and Aral 2018). Second, their sample was composed of articles flagged by Facebook's nontransparent internal systems to be potentially problematic. Previous work has shown that article sampling methods have the potential to undermine external validity (Clemm von Hohenberg 2020). While the Facebook sample is "representative of what social media platforms would have directed to professional fact-checkers" (Allen et al. 2021), it is important to replicate these findings with a different theoretically motivated sampling frame, such as popularity. Finally, we ask respondents to evaluate the full article, rather than just the headline; we do so because it more closely mimics the current fact-checking standards employed by social media platforms. We therefore complement the findings of Allen et al. (2021) by testing the efficacy of crowdsourced fact-checking in real time using a transparent, preregistered method for sampling popular articles.

There are a number of factors that may inhibit the ability of the crowd to classify the veracity of news articles in real time. The "wisdom of the crowds" literature specifically warns that systematic error or bias among individuals will undermine the group's accuracy (Simmons et al. 2011). There is robust evidence of partisan-motivated reasoning as individuals seek out information that confirms their partisan identity (Van Bavel and Pereira 2018; Bolsen, Druckman, and Cook 2014; Druckman and McGrath 2019). In particular, partisans exhibit large biases when evaluating ideologically concordant fake news (Aslett et al. 2021). It could be the case that partisan-motivated reasoning compromises the ability for crowds to accurately assess news articles, especially given that the supply of fake news is ideologically imbalanced on social media (Guess, Nyhan, and Reifler 2018). In addition, prior exposure to false stories increases one's belief in its veracity, and so individuals may be especially inaccurate when evaluating the most viral stories that have garnered significant exposure online (Fazio et al. 2015; Pennycook, Cannon, and Rand 2018; Wittenberg et al. 2020).

Assessment of the veracity of fake news operates in an antagonistic information environment in which news items are deliberately designed to trick the user into believing their credibility. In canonical examples of collective intelligence, the crowd's evaluations are used to recover factual information, such as the weight of an ox. In such cases, the ox does not benefit from obscuring its weight. While the usefulness of crowdsourced judgments has also been found in information environments that resemble fake news—e.g., identifying phishing websites (Moore and Clayton 2008) and other cybersecurity threats (Sharifi, Fink, and Carbonell 2011)—it remains unclear whether fact-checking is a domain in which the crowd can correctly categorize news stories that are intentionally deceptive.

Fact-checkers are skilled professionals, and there is evidence that their training equips them with the ability to arrive at warranted conclusions as to the credibility of information more quickly and accurately than other groups (Wineburg and McGrew 2017). Previous studies suggest that although ordinary people can distinguish between lower- and higher-quality news sources (Pennycook and Rand 2019a), they have high error rates when assessing the veracity of false news headlines and full articles (Aslett et al. 2021; Pennycook and Rand 2019b). While processes that aggregate evaluations have been shown to mitigate errors, a crowd composed of individuals who have a low probability of accurate judgements might fail to produce significant improvement over the average individual (Grofman, Owen, and Feld 1983). Therefore, a crowdsourced approach might fall victim to the simple fact that ordinary people are unable to distinguish between true

and false news articles in real time, especially before fact-checking sites like Snopes or PolitiFact have published their evaluations.

Finally, while a crowd composed of the general public may prove to be ineffective, recent literature on collective intelligence highlights the importance of crowd composition. In particular, while there are certain tasks for which general crowds (i.e., composed of randomly selected individuals) are effective, other tasks benefit from select crowds (i.e., composed of individuals with certain characteristics) (Mannes, Soll, and Larrick 2014; Goldstein, McAfee, and Suri 2014). Following this logic, crowds balanced by ideology (and thus offsetting partisan-motivated reasoning in either direction) or composed of people with high political knowledge (Clayton et al. 2020) might improve performance relative to randomly selected crowds. Recent work also shows that, in contexts where simple aggregation techniques may fail to produce accurate results, the use of machine learning to extract meaningful signals from the crowd's evaluations can be particularly effective (Laan, Madirolas, and Polavieja 2017). Taken together, this literature suggests looking beyond simple aggregation methods to fully test the viability of crowdsourced fact-checking, specifically examining if either select crowds or the use of machine learning can improve the performance of crowd-based predictions.

Here, we use both simple aggregation rules and machine learning to test whether crowds of ordinary people are able to match the article-level evaluations from a panel of PFCs. Combining these two approaches – simple aggregation rules vs. machine learning – with random versus selective crowds gives us four versions of crowdsourced fact-checking to compare: (1) random crowds with simple rules; (2) select crowds with simple rules; (3) random crowds with machine learning; and (4) select crowds with machine learning.[3] In the following section, we explain our real-time article selection mechanism, the collection of evaluations from ordinary people and PFCs, the construction of crowds, and the different simple aggregation rules and machine learning approaches we use to assess the four different approaches to crowdsourced fact-checking.

## 3 Real-Time Data Collection

Any method attempting real-time false news identification must be tested on articles as they are published. A key innovation of our research design is collecting a theoretically motivated sample of articles within 24 hours of their publication, prior to any publicly available third-party fact-checking, and sending them to be evaluated by a representative sample of Americans within 72 hours of publication. As a result, the veracity of an article was unknown when it was included in our study, which requires dealing with several challenges. Below we discuss how we sourced news articles in real time using a preregistered method that was replicable, ideologically balanced, and transparent. We then lay out the survey methodology that was employed to collect the evaluations of these articles, just after publication, from both ordinary people and PFCs.

3. From the perspective of the platforms, simple rules with random crowds would have obvious benefits. One of the most often heard claims from Facebook founder Mark Zuckerberg is that he does not want Facebook to be the "arbiter of truth." If simple rules with random crowds could generate accurate assessment of the accuracy of news, then Facebook could get out of the fact-checking business and offload news evaluations to randomly selected users. This would be attractive to the platforms insofar as the "user community" could be said to be in charge of determining what content is demoted on the platform; simple aggregation rules would make this process easier to explain to the public. As we demonstrate in the remainder of the manuscript, however, such approaches are the least successful of the four types we analyze.

### 3.1 Article Collection

A key challenge for testing the efficacy of crowdsourced fact-checking is sourcing articles and collecting evaluations in real time. To this end, we developed a pipeline for collecting evaluations of popular news articles by both PFCs and ordinary respondents within 72 hours of an article's publication. Importantly, we did not select specific articles to include based on our own assessment of their appropriateness for the research, which has introduced questions of external validity for previous studies in which respondents evaluate the veracity of news articles (Clemm von Hohenberg 2020). Instead, each day of the study we used a preregistered algorithm to bind ourselves to choosing the most popular article that had appeared in the previous 24 hours from streams of three distinct low-quality news sources (one left-leaning, one right-leaning, and one without a clear partisan lean) that we constructed before the beginning of the study.[4] The streams were designed to ensure a balance of liberal- and conservative-leaning news, as well as a balance of false and true content.[5] For each day of our study, we sent our three articles—one from each of the streams—within 24 hours of their publication for evaluation (see Appendix A.1 and A.5 for a detailed study overview).[6]

To construct these "low quality" news streams, we used lists of websites known to produce fake news. Our data collection proceeded in two periods. In the first period, which ran from November 2019 through February 2020, we used all low-quality news sources from Allcott, Gentzkow, and Yu (2019), which itself combines well-known lists of false news websites. A total of 99 websites were active at the time of our study. In the second period,[7] which ran in May and June 2020, we supplemented the Allcott, Gentzkow, and Yu (2019) list with domains identified by NewsGuard, a web extension that provides live informational feedback on online news source reliability (NewsGuard 2021), as frequently producing false news related to the pandemic. A total of 45 websites were added. We then classified these low-quality sources by partisan lean (liberal, conservative, or unclear) to create an ideologically balanced sample of articles.[8] Finally, on each of the 39 days of our study, we selected the most popular article over the previous 24 hours from each of the streams as measured by social interactions using Feedly, an RSS news aggregator.[9] Taken together, this procedure enabled us to collect a sample of popular articles on a given day, balanced by partisan lean. Our sample for this paper

4. We also included two streams of mainstream news (one left-learning and one right-leaning). We focus here only on articles from low-quality news sources for two reasons. First, previous literature suggests that while members of the public are able to correctly identify true news from mainstream sources, they struggle to differentiate articles from low-credibility websites that are false from those that are not (Aslett et al. 2021; Pennycook and Rand 2019a). Second, virtually all articles from mainstream news sources were labelled "true" by PFCs, and so would be a relatively trivial task for which a crowdsourced approach is not necessary.

5. As will be made apparent below, we find that even a substantial number of articles from low-quality news sources are evaluated as true by our PFCs. Thus a sample of articles from low-quality sources produces both false and true stories, while a sample of articles from high-quality sources primarily produces only articles that are ranked as true by PFCs.

6. The data utilized in this manuscript were collected as part of a larger project that also included two additional article streams from mainstream (i.e., not low quality) new sources, one left-leaning and one right-leaning. Given (as detailed below) that each respondent evaluated three articles, this means that while some may have only received articles from the low-quality streams to evaluate, others will have evaluated one or two articles from the mainstream threads in addition to one article from the low-quality threads. For full details on the data collection and survey design, see the Appendix. For the purposes of the current study, we evaluate the performance of crowds only on articles from low-quality news sources.

7. In the second period, the sample was split evenly between the most popular articles concerning COVID-19 and the most popular articles not concerning COVID-19. No such distinction or split was made in the first sample.

8. Two research assistants coded each website for partisan lean; if they were split, a third coder was used to break the tie.

9. For more information on Feedly's popularity measure, see https://feedly.com/i/entry/NBHkLLj8YGLLEyGA+0mSpEvCPJ4mKcxBYbHNP0YqkfY=:1570c7dc2d6:bd400f4:e3157ec0.

consists of 135 articles from low-quality sources, 93 of which were sourced in the first period of data collection and 42 of which were sourced in the second period. Of these articles, 57 were classified as "false" by the fact-checkers, 58 were classified as "true," 4 were "could not determine" (CND), and 16 had no fact-checker consensus. Given the goal in this article is to identify if articles are false or not, 57 articles were therefore classified as "false" and the remaining 78 were classified as "not false."

### 3.2 Collecting Respondent Evaluations

After articles were collected each day, they were sent out as part of an online survey to be evaluated by approximately 90 respondents aged 18 and over who were recruited by the survey research firm Qualtrics.[10] No respondent could take a survey more than once. We also sent each day's articles to a panel of six PFCs from leading national news publications.[11] In the first period of data collection, PFCs and respondents were required to complete the survey within 24 hours. In the second period of data collection, we delayed sending the surveys to respondents by 24 hours, which enabled us to collect PFC evaluations before sending out the articles to respondents.[12] Given that our pipeline only produced articles for evaluation that had been published in the past 24 hours, all of the evaluations in the study were collected within 72 hours after publication.

An important feature of this study is that we collected crowd evaluations in a manner that would approximate a scalable crowdsourced fact-checking system. While there exist current models for much more involved crowdsourced systems, such as Public Editor,[13] these models require substantial training and time to produce evaluations. Instead, our research design examines the efficacy of a simple, straightforward crowdsourced fact-checking system that requires only a minor investment of time on the part of those who make up the crowd, and therefore has the potential to match the scale of social media platforms.[14] To this end, the participants in our surveys, who were compensated by Qualtrics, evaluated the articles without prior training and within a moderate timeframe. It should also be noted that Aslett et al. (2021), who used a similar design, found that offering respondents additional compensation for correctly evaluating the veracity of news articles did not increase their accuracy; therefore, we would not expect larger incentives to improve the crowds' accuracy relative to what we report in this study.

The respondents and the PFCs rated the veracity of the central claim of each article on both a categorical scale ("true," "false/misleading," or "could not determine") and a

10. As noted previously in footnote 6, these data were collected as part of a larger project that included two additional streams of mainstream news. We asked each respondent to evaluate only three articles because results from pretests showed attrition increased when we asked subjects to evaluate more than three articles over the course of the survey. Thus the research design involved recruiting approximately 140–160 respondents—balanced on age, gender, partisanship, and education—each day, meaning each article was evaluated by approximately 90 respondents. We used census proportions that approximated to: partisanship - 1/3 who self-identify as a Democrat, 1/3 who self-identify as moderate, 1/3 who identify as Republican; gender - 1/2 who self-identify as male and 1/2 who self-identify as female; education - 2/3 have no high school, a high school degree, or partial college; 1/3 have a college degree or more; age - 30% ages 18–34, 34% ages 35–54, and 35% ages 55+.

11. The PFCs were recruited through emails sent to national media employees and organizations; each PFC was either currently working or had previously worked as a fact-checker for a reputable national media outlet, such as the Atlantic, the New Republic, or NPR. None of the PFCs were employed by the outlets included in our article selection streams so as to avoid any conflicts of interest. For each article evaluation, the PFC was compensated $10. Not all of the PFCs responded every day, but we never had fewer than four respond; the modal number of daily PFC evaluations was five.

12. The delay enabled us to immediately communicate the PFC evaluations to respondents once they had completed their own evaluations, thus reducing the potential for an article to misinform a respondent about pandemic-related topics.

13. For more information, see https://www.publiceditor.io/

14. To put another way, all that is required to implement the system we are testing is existing online survey capacity, which companies such as Qualtrics now offer in over 90 countries around the world.

seven-point veracity scale (from "1 - definitely false" to "7 - definitely true"). Importantly, "false" and "misleading" were combined, as pilot surveys indicated respondents had difficulty distinguishing between the two categories. Therefore, for the purpose of this paper, "false" refers to content that is evaluated as false and/or misleading. In addition, respondents completed a political knowledge test, a digital literacy test, and a number of standard questions related to sociodemographic status.[15] In total, our data consists of 12,883 "respondent evaluations" across the 135 articles. Taken together, our study captures real-time evaluations from both survey respondents and PFCs on a sample of popular news articles from low-quality news sources balanced on partisan lean.

## 4 Methods

The central research question of this manuscript is whether crowds of reasonable size—using various theoretically informed aggregation rules—are able to produce an evaluation of an article that would match the evaluation produced by PFCs *in real time*. To be clear, we are interested in whether we could use a crowd of ordinary people to replace a PFC in delivering a quick evaluation within the first 72 hours after an article appeared online. To this end, we first illustrate how we built the crowds utilized in our analyses from the individual survey responses we collected. We then turn to an explanation of how we split these crowds into train and test sets for supervised machine learning (ML) models. Finally, we describe how we utilize the evaluations from the panel of PFCs to introduce benchmarks that can be used to assess the performance of the crowd.

### 4.1 Crowd Creation

To measure the performance of multiple crowdsourced approaches to fact-checking, we constructed multiple crowds of plausible sizes (i.e., crowds that could be reasonably used in a fact-checking system) by aggregating individual survey responses of the same article. Here, we analyzed crowds that ranged from a size of one to 25 individuals. Crowds were limited to a maximum of 25 for both theoretical and practical reasons. First, our analysis of simple rules (discussed below) illustrates that the performance of crowds essentially plateaus once crowds have roughly 20 members, and our preliminary analysis revealed that increasing crowds to a size greater than 25 did not lead to an increase in performance. Practically speaking, our machine learning models used information from each member of the crowd as inputs, so each additional member increases the number of features by a factor of three, which translates into increased model complexity and training time. Finally, aggregate statistics, such as the mean and mode, based on crowds will change very little with marginal increases in crowd size and thus have very little ability to meaningfully change machine learning predictions. These factors led us to limit crowds in our analyses to no larger than 25.

Using the approximately 90 distinct evaluations of each of our 135 articles, we constructed these simulated crowds. We created a "crowd" associated with an individual article by randomly sampling with replacement from the 90 responses associated with that article. For example, for a given article, to construct a crowd we randomly selected 10 individual evaluations of that article from the roughly 90 evaluations of that article present in our survey data. How the crowd evaluated that article—i.e., how we would combine the 10 evaluations in the crowd—was then determined by a variety of methods and compared to the PFC evaluations of that article (discussed below).

15. For full details of the survey, see Appendix A.2.

It was also important that crowds be comparable. For example, it is important to examine how crowds of size 10 perform, on average, and then contrast that to the performance of crowds of size 11. The most appropriate way to do that is to simply add one additional person to each crowd of 10, so that in the comparison the crowds are largely the same, save for the addition of one extra person. In order to accommodate this, each crowd we simulated was deliberately oversized—containing 60 respondents. We sampled 60 rather than 25 so that if we wished to sample crowds with specific characteristics, such as having high political knowledge, there would be a sufficient number of responses with that characteristic within the 60. Construction of sample crowds of smaller size, typically 10 or 25, were built by subsampling these larger crowds. In instances where we wished to select crowds with select characteristics—for example, higher political knowledge—we scanned through each member of the larger crowd of 60 and selected the first ten individuals that met this criteria.[16] As our goal was to estimate how crowds would do on average, we bootstrapped 500 crowds of 60 individuals, which allowed for a large and robust dataset that was nonetheless still manageable in terms of the scale of the data and computation.[17]

This method also ensured as much similarity as possible when comparing crowds of different sizes or respondent composition. Importantly, smaller crowds are always included in larger crowds, so that comparison of performance of crowds of 10 versus 25 in size means that 15 new evaluations have been added to the crowds of 25—not that 25 new evaluations were sampled. This allowed a consistent comparison between different crowd types and methods.

### 4.2 The Train–Test Split

In our analyses, we use two distinct approaches to aggregating a crowd of individual respondents' evaluations of an article into a single crowd evaluation: simple aggregation rules and machine learning. The former—for example, taking the mode of the crowd—are straightforward and can be assessed on their performance on the entire dataset. Conversely, our machine learning method, in which we treat the evaluations of an article by a crowd as inputs into supervised ML models, requires subdividing data into a training and test set. Consequently, ML models will only be assessed on their performance on the test set of data. Therefore, when we present our performance assessments of our simple aggregation rules, we do so not only on the entire dataset, but also on the test set for purposes of direct comparison with ML results.[18]

### 4.3 PFC ratings

To assess the efficacy of the crowd, a crowd's evaluation needs to be compared against the veracity of the article in question. In more traditional crowdsourcing contexts, the ground truth is conspicuous and so the comparison proves relatively straightforward

16. We were not able to create select crowds larger than 10 individuals due to the limited number of respondents who had a particular characteristic.

17. We sample approximately rather than exactly 500 times in order to insert some variation into the sample, which will better reflect a natural information environment.

18. Given the structure of our data—bootstrapped crowds based on 135 articles—it was key that our train–test split occur at the article level, rather than at the individual crowd level. If we were to split at the crowd level, our algorithms would have crowds evaluating articles in the test set, where those same articles were evaluated by different crowds in the training set. Besides being a form of leakage whereby information about the test set is made available in the training set, it would also be unlikely that any crowd-based fake news algorithm would be applied to articles that have already been evaluated by other crowds. Given this split, articles were assigned to either the train, validation, or test set, with observations then being sorted into the train, validation, or test set based on the article that was being evaluated by that crowd. All ML algorithms were then trained and tuned using the training and validation sets, with results presenting performance only on the test set.

(e.g., the market goes up or down; a sports team wins or loses; the cow actually does weigh 1,198 pounds). If our PFCs unanimously agreed on every article, there would be little reason to not simply take the point of unanimous agreement as the "ground truth": an article would be true or false, or it would be impossible to determine the article's veracity given the available information based on the assessment of the PFCs. However, in many cases, our PFCs did not unanimously agree with each other, introducing challenges for determining the veracity ("ground truth") of articles in which there is not uniform agreement.[19] Encouragingly, though, the disagreement we find is not out of line with previous work observing variation in the evaluations of Snopes and PolitiFact (Lim 2018),[20] despite the fact that the fact-checking process in our study does not perfectly reflect the more extensive process undertaken in newsrooms or fact-checking organizations (Graves 2016). Accordingly, we take the mode of the fact-checkers as our measure of "ground truth."[21]

As our primary research question is whether the crowd can be used to distinguish the set of articles that are problematic from those that are not, we focus on examining whether the crowd can predict if a panel of PFCs evaluated an article as false or not. Of the 135 articles in our study, 57 articles are coded as "false" according to this rule (i.e., the mode of the PFCs is false or misleading), and 78 articles are coded as not meeting this standard (i.e., mode is true, could not determine, or there is no mode[22]). Note that for purposes of classification below, a "positive" refers to an article that is classified as "false."

### 4.4 Baseline Performance and PFC Benchmark

Crowd performance cannot be assessed without context, specifically a benchmark or baseline by which to assess how well crowds are doing relative to an alternative. We have two such metrics. Firstly, there is *baseline performance*, which provides a minimum or floor for performance. As our outcome variable ("false" or "not false") is binary, baseline performance is simply predicting the majority category in the dataset. In our test set, 53.9% of observations are "not false" and the remainder are "false." By predicting "not false," a naive classification method could achieve 53.9% accuracy; therefore, for a classification method to have value, it must achieve accuracy greater than 53.9%. In the entire dataset, baseline performance is 58%.

Secondly, beyond *baseline performance*, classification approaches are typically compared to a human standard, in this case the performance of a professional fact-checker. In order to construct a professional human standard, we created a *PFC benchmark*, which measures how well we might expect an individual PFC to distinguish between false news and news that is not false.

19. To be clear, this is not intended as an indictment of the professional fact-checking industry. We asked our PFCs to simply answer our questions and to return our surveys within 24 hours (which, to reiterate, were themselves sent out within 24 hours of the article first appearing online). These are not the typical working conditions for PFCs, nor necessarily the type of outputs they normally produce. We did so for two reasons. First, we wanted the fact-checkers to evaluate the articles in the same period of time as the respondents, thus having access to the same information. Second, there were resource constraints for completing this study, and employing six fact-checkers to complete a standard fact-check for each article would have been prohibitively expensive.

20. A table including the agreement and Fleiss Kappa of the professional fact-checkers is in Appendix A.3.

21. Another solution to establishing "ground truth" could have been to simply employ a single PFC. While using one PFC would have made classification more straightforward, we think that, given the variation between PFCs, it is important for any study that attempts to classify news in real time to use a panel of PFCs to reduce the arbitrariness of the "ground truth" classification based on which the individual PFC was selected for inclusion in the study.

22. While an article not having a modal evaluation from fact-checkers might suggest the article is not entirely true, it does not necessarily indicate that it is false, either. Knowing that platforms suppress the visibility of articles that are evaluated to be false or misleading, we use a strict definition of false and thus count articles without a mode as "not false."

To do so, recall that each article is designated as either "false" or "not false" based on the mode of a group of PFCs (typically five). In order to evaluate how well an individual PFC would do, we calculated how often an individual PFC could predict the mode of the other PFCs in the group. If they did match, we marked that individual fact-check as correct, and if they did not, we marked it as incorrect. This benchmark is similar to the one constructed in Allen et al. (2021), in which they used the average correlation among three PFCs.

We did this for every PFC of every article in our study and found that PFCs correctly predicted the modal answer of the other PFCs 72.8% of the time.[23] Generally speaking, this means that a PFC can predict what other PFCs would say about an article from a low-quality source at about 72.8% accuracy. It is important to remember here that our PFCs were also just asked to provide their assessment of the article's accuracy in the same manner as our survey respondents, and therefore did not carry out a full-fledged fact-checking exercise that might involve days of researching the topic, sourcing original documentation, and conducting interviews with experts (e.g., Politfact 2020). Consequently, we would expect PFCs to have an even higher benchmark score were they able to complete a full fact-check. Therefore, this benchmark serves as a floor insofar as we would hope for crowd performance to meet or exceed this standard. In short, the *PFC benchmark* provides a rough yardstick for expert-level performance, but likely underestimates it. We also performed the same exercise for the test set, which is a subset of the data used for testing Bayes' rule and the machine learning algorithms. For the test set, the PFC benchmark was, reassuringly, very similar at 69.4%. In this case, it means that for any method used on the test set, comparable performance to PFCs would be an accuracy of 69.4%.

## 5 Results

With these data and benchmarks in hand, we are able to test the efficacy of our four different approaches to crowdsourced fact-checking. Recall that in addition to varying membership in our crowds—*random crowds* versus *select crowds*—we also consider two different classes of methods for extracting crowd evaluations (simple rules and ML-based methods). We begin with *simple rules*—in this case, taking the modal response from crowds of various sizes and of various compositions. We then employ three *machine learning* models to extract signals from crowds using more information than simply the crowd's categorical evaluation; specifically, these models use additional evaluation metrics provided by the crowd, as well as sociodemographic information about the members of their crowd (e.g., partisanship).

### 5.1 Simple rules

In the first set of analyses, we take the modal response of general crowds (crowds composed of randomly selected individuals) and select crowds (crowds composed of individuals with characteristics informed by previous studies, such as ideological balance and high political knowledge).

In the first aggregation, we begin by randomly selecting one respondent from each crowd, then increase each crowd by incrementally adding one additional respondent without replacement from the larger crowd of 25. To measure performance at a particular crowd size, we use the same binary false/not false variable (1 = crowd modal response is "false" or "misleading," and 0 = crowd modal response is "true," "could not determine," or there

23. The corresponding measure for performance on mainstream sources is 92%.

is no mode). If the crowd's evaluation matches the PFC evaluation, it is considered correct; otherwise, it is considered incorrect.

As shown in Panel A of Figure 1 on the facing page, the proportion of crowds that match the PFC mode—which we refer to as crowd performance—ranges from 57.4% at n = 1 to 61.5% at n = 25. This result is striking for two reasons. First, even at crowds of size 25, performance remains significantly below the PFC benchmark (72.8%). Second, our results do not show that performance significantly increases as the size of the crowd increases, as is often the case in more traditional contexts where collective intelligence is demonstrated (Boland 1989).

As people's assessment of news is often influenced by partisan-motivated reasoning, ideologically balanced crowds might attenuate the biases that could inhibit crowdsourced article evaluations. Therefore, in Panel B, we use respondents' self-identified ideology and create ideologically balanced crowds between 3 and 15 respondents—increasing by a factor of 3 and including an equal proportion of liberals, moderates, and conservatives. Performance ranges from 59.1% to 61%, marginally worse than the crowd without ideological balance. Given that our initial sample of respondents was roughly balanced by partisanship, it is perhaps unsurprising that this sampling method does not generate improvement.

If we believe that people "know the truth when they see it," then treating "could not determine" as "false" and recomputing the mode could lead to an improvement in accuracy. However, Panel C reveals that doing so instead generates a substantial decrease in performance, ranging from 57.4% to 62.9%, suggesting that respondents who respond "could not determine" are in fact uncertain as to the veracity of the articles. Taken together, the results from Panels A-C of Figure 1, which remain below the 72.8% PFC benchmark, suggest that simple rules-based methods for aggregating evaluations from general crowds are unlikely to approximate the performance of a PFC.

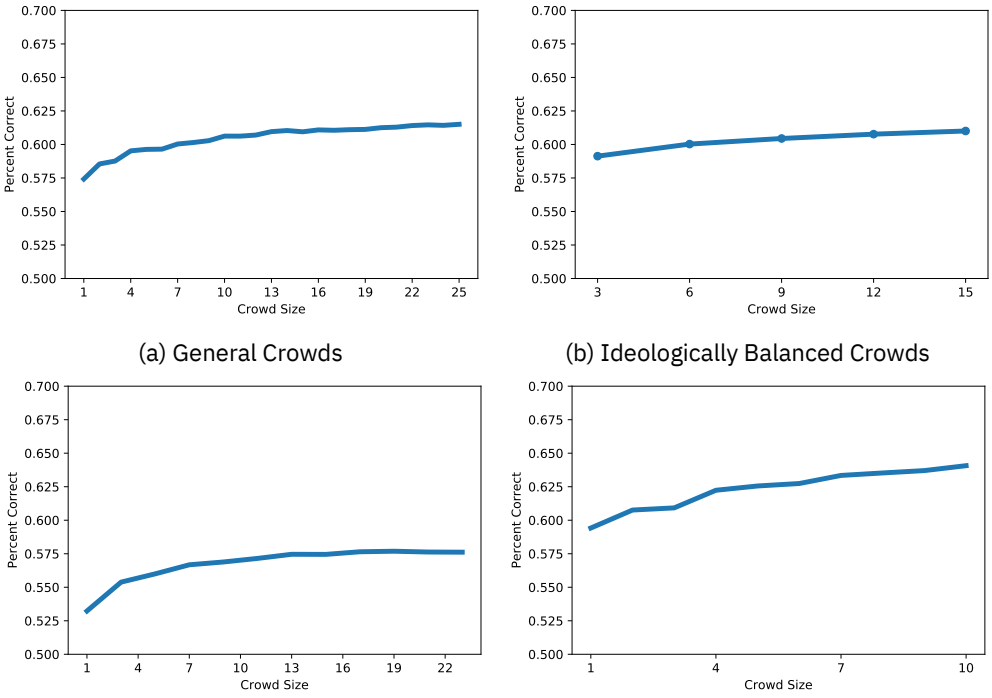
In addition to general crowds drawn from the overall sample of respondents, we measure the performance of select crowds—i.e., crowds composed of individuals who have a characteristic that has been shown to be associated with high performance at this task. Having high political knowledge (Clayton et al. 2020), which we define as individuals who successfully answered all four political knowledge questions on the survey, is one such characteristic. Overall, just over half of respondents (50.7%) were classified as having high political knowledge. These questions were designed to identify those with a level of political knowledge consistent with a well-informed lay person, not a specialist. The questions used to measure political knowledge are listed in Appendix A.2.

In Panel D, we create crowds between 1 and 10 respondents with high political knowledge.[24] These select crowds generate improvement over general crowds, with fewer overall evaluations. Notably, crowds of 10 individuals with high political knowledge achieve performance of just over 64%—nearly halving the gap between the performance of general crowds (Panel A) and the PFC benchmark.

Another plausible transparent approach would be to use simple Bayesian inference to estimate if an article is false or not. Applying Bayes' rule in this context is fairly straightforward, where the prior and all conditional probabilities are estimated in the training set and then applied to observations in a test set (see Appendix A.4 for a more detailed explanation).[25] As shown in Table 1, Bayes' rule outperforms all other simple aggrega-

24. We only increased our crowd size to 10 because in some of our bootstrapped crowds, there were not more than 10 respondents who met our definition of high political knowledge.

25. For full details of the train–test split, please see the Machine Learning-Based Prediction section. Bayes' rule is not evaluated on the training set, just like the ML model, because Bayes' rule utilizes data from the training set to make its predictions.




(a) General Crowds

(b) Ideologically Balanced Crowds

(c) Could Not Determine as False

(d) High Political Knowledge

Figure 1: Percentage of crowd evaluations that match PFC evaluations (y axis) at various crowd sizes (x axis). (A) The proportion of general crowds, increasing in size by 1, that match the modal response from PFCs. (B) The proportion of ideologically balanced crowds, increasing in size by 3, that match the modal response from PFCs. (C) Reclassifying all "could not determine" responses as "false," the proportion of crowds, increasing in size by 1, that match the modal response from PFCs. (D) The proportion of crowds made up of people only with high political knowledge, increasing in size by 1, that match the modal response from PFCs.

tion rules in accuracy on the test set, with the exception of the high political knowledge crowd. Importantly though, Bayes' rule "learns" the distribution of evaluations with different article types ("false" and "not false"), and thus is somewhere in between a simple aggregation method and the more complex ML methods.

For each rule in Figure 1, we report in Table 1 aggregate statistics at the largest crowd size we tested.[26] We find that no simple rule performs particularly well, but that performance was highest when we took the modal response of select crowds composed of respondents with high political knowledge. In addition to overall performance, we specify the true and false positive and negative rates. Any system that evaluates the veracity of articles balances a tradeoff between these measures. Notably, for every rule included in our study, the false positive rate is greater than 40%. This means that no matter which rule we used, more than 40% of the time a crowd evaluated an article as "false," it was "not false" according to the mode of the panel of PFCs. Finally, for purposes of comparing performance across simple aggregation and ML-based rules, we have included accuracy results for the test set.[27]

| Crowd Composition | Crowd Size | Rules | Test Accuracy | All Data Accuracy | True Positive Rate | False Positive Rate | True Negative Rate | False Negative Rate |
|---|---|---|---|---|---|---|---|---|
| General | 25 | Mode | .55 | .62 | .57 | .42 | .63 | .37 |
| Ideologically Balanced | 15 | Mode | .56 | .61 | .56 | .45 | .62 | .38 |
| General | 25 | Mode, Treat CND as F/M | .63 | .58 | .50 | .50 | .78 | .22 |
| High Political Knowledge | 10 | Mode | .60 | .64 | .60 | .40 | .66 | .34 |
| General | 10 | Bayes | .60 | NA | .57 | .43 | .62 | .38 |
| General | 25 | Bayes | .58 | NA | .57 | .43 | .59 | .41 |

Table 1: Performance of crowds in Figures 1 A-D, with additional statistics of true and false positive and negative rates. Additionally, we report the crowd accuracy on the articles included in the test set. CND = Could Not Determine; F/M = False or Misleading; Treat CND as F/M refers to the test where we treated "could not determine" responses as "false or misleading."[28]

### 5.2 Machine Learning-Based Prediction

We also tested whether a more sophisticated machine learning-based approach, based on the same crowds, could better predict PFC article ratings. In the previous section, we examined simple rules where evaluations of the crowd were used to make predictions about the veracity of an article directly. By contrast, in a ML-based approach, the information from the crowds was used as inputs to supervised ML or statistical learning algorithms, where the data for each observation were the ratings of a crowd of lay people about a given article, and the label for each observation was the evaluation of that article

26. The True Positive Rate is defined as TP/(TP + FP) and the False Positive Rate is FP/(TP + FP). The corresponding Negative Rates are defined as the same, but using negatives rather than positives.

27. The test set is randomly determined and is by design a separate set of articles from the training set, and thus this sample of articles has the potential to be easier or harder for the crowd to evaluate. In order to test the performance of each simple aggregation rule on the test set, we use the defined method but limit our results to the articles in the test set. We found that performance on the test set was noticeably lower and showed little improvement with larger crowds.

28. Bayes' rule results rely on a training set to make predictions; as such, their performance can comparatively be assessed only on the test set, and not on the entire set. Consequently, there are no results shown for Bayes' rule for accuracy on the entire set.

by the PFCs. Crucially for the sake of comparability, the *same* crowds that were used in the prior analysis were also used here; specifically, the same randomly generated crowds were used when training and evaluating the ML algorithms as were used in the analyses in the previous section.

For all ML algorithms, we used a small number of original features (three per respondent in the crowd) from the crowd. Specifically, the categorical evaluation ("true," "false/misleading," "could not determine") of each article by each respondent in the crowd was used, as well as a Likert-based evaluation (1-7 from "completely false" to "completely true") and the partisanship of each respondent.[29] Additional features were then generated based on these inputs. Specifically, additional features include the mode of the crowds' categorical responses, the count of each type of categorical response, as well as the mean, range, and variance of the Likert-based evaluations. For a full list of features, see Appendix A.7.

We evaluate three algorithms — elastic net, random forest, and neural networks (NNs)[30] — using these features. The algorithms chosen were employed based on the task at hand and the best currently available tools. Of the three, elastic net is the most straightforward. The elastic net model used here is just a simple linear sigmoid model (essentially a logistic regression), but with L1 and L2 regularization. Regularization is a method to penalize models for having large coefficients that bias models towards simpler predictions, which in turns helps models to ignore noise in the training set.

Random forest models are more sophisticated and do a better job accommodating nonlinear data. Random forests are constructed by aggregating many decision trees together, known as bagging. An individual decision tree is constructed by selecting a random feature in the data, finding a value of that feature that splits a random subset of observations efficiently (i.e., into more homogenous groups), and then selecting another feature and repeating the process on those subgroups until the tree reaches a certain size. After generating many trees, predictions can be made by identifying the predictions each decision tree would make for an individual observation, and using the majority prediction as the final prediction for that observation. Random forests have many advantages, primarily that they are nonlinear and are thus able to accommodate a variety of behaviors in the data, and they are unlikely to overfit (treat noise in the training set as useful information) because so many trees are typically used (usually over 1,000). Their ease of training and implementation make them one of the most common off-the-shelf ML models.

Finally, neural networks are among the most sophisticated nonlinear ML models. Their inner workings are difficult to summarize, but essentially NNs take the input data and make a series of linear transformations to it, then apply a nonlinear function to those transformations, and then repeat the same process of linear transformation followed by nonlinear function, until all the predictions are combined to make a final prediction. Neural networks are the basis for many of the most important innovations in machine learning, and their sophistication ranges from straightforward to complex. More than any other ML models available, they can accommodate nonlinear behaviors in data and have consequently become the premier model of choice for working with nonstandard data such as text or images. While the NN models used here are fairly straightforward, they should be able to accomodate complexities in the data that other models may miss and must be part of any analysis that attempts to evaluate the potential of ML-based methods. Beyond ML-based methods, we also examine a specialized subset of crowds, specifically high political knowledge, based on our simple rules findings.

29. For full wordings of the question used to elicit this data, see Appendix A.2.
30. Neural networks contained six hidden layers utilizing ReLU and softmax for final predictions using pytorch (Paszke et al. 2019).

### 5.3 Machine Learning-Based Results

Table 2 presents the accuracy of the various ML-based methods. In general, these methods were almost universally superior to simple rules-based approaches, but the relative performance varied significantly. In this case, baseline performance on the test set was 53.9%, meaning that 53.9% of the observations were "not false," and we would expect any algorithm to achieve a minimum performance of 53.9% or better, which most algorithms achieved. Yet no algorithm was able to meet the PFC benchmark performance of 69.4% on the test set.

Notably though, NN-based methods with crowds of size 25 were able to achieve 68% percent accuracy on the test set, just below the PFC benchmark of 69.4%. The use of a larger crowd coupled with a more sophisticated approach yields results that are near the same level as a human expert standard, a significant improvement over both heuristic rules and simpler algorithmic approaches. Similarly, relatively small crowds of 10 evaluations by individuals with high political knowledge, when paired with an NN, achieved almost 68% performance. That being said, it is important to remember that our PFC benchmark is a low estimate of a professional human standard, and even with this low benchmark no algorithm was able to meet it.

| Method | Crowd Size | Crowd Composition | Test Accuracy | True Positive Rate† | False Positive Rate† | True Negative Rate† | False Negative Rate† |
|---|---|---|---|---|---|---|---|
| Random Forest | 10 | Random | .58 | .56 | .44 | .58 | .42 |
| Elastic Net | 10 | Random | .56 | .56 | .44 | .57 | .43 |
| Neural Net | 10 | Random | .66 | .71 | .29 | .64 | .36 |
| Random Forest | 25 | Random | .60 | .58 | .42 | .61 | .39 |
| Elastic Net | 25 | Random | .56 | .54 | .46 | .57 | .43 |
| Neural Net | 25 | Random | .68 | .73 | .27 | .66 | .34 |
| Random Forest | 10 | High Political Knowledge | .63 | .64 | .36 | .63 | .37 |
| Neural Net | 10 | High Political Knowledge | .68 | .69 | .31 | .67 | .33 |

Table 2: The accuracy of various ML algorithms used on different crowd sizes and compositions.

† The original values for this column, published on October 28th, 2021, have been updated to reflect corrected calculations.

Interestingly, only neural network models were able to achieve significant improvements in performance relative to the baseline. This performance was driven largely by improvements in true positive rates. Generally speaking though, all the non-NN based algorithms were slightly worse at identifying "not false" articles, with false negatives rates between 37 and 43%. NN-based methods marginally improved on this metric as well, boasting false negative rates between 33% and 36%. Despite the superiority of NN-based models, even the best models have false positive and negative rates on the order of 30%. Consequently, none of the methods can be considered especially adept at distinguishing false and not false news.[31]

Additionally, in order to assess the robustness of these estimates to different news environments, we simulated test sets where articles labelled "False" were 2%, 10%, 25%, and 90% of the set.[32] We then evaluated the performance of the two best-performing algorithms (NNs with crowds of 25 and NNs with crowds of 10 high political knowledge)

31. This paragraph has been updated from a previously published version of this article to reflect the updated calculations made to Table 2.

32. Therefore, the relative baseline performances measures were 98%, 90%, 75%, and 90%, respectively.

on these test sets. We found that rates of performance—e.g., false positive, true positive, etc.—were very stable across these different test sets. At the same time, accuracy was consistently below baseline performance, which can be attributed to the false positive and false negative rates.[33]

This has important implications for judging the performance of these models. If a model trained in one environment were to be shifted to another, where the distribution of "false" news was extremely different, this evidence suggests the model would be highly inaccurate. Essentially, these models are sensitive to data they are trained on: not just the individual articles or crowds, but rather how often articles of certain types appear. If the setting changes, the ability of these models to make accurate predictions deteriorates considerably. Consequently, none of the methods identified in this paper seem appropriate as a standalone approach to fact-checking in any information environment, regardless of the distribution of "false" or "not false" news.

### 5.4 Feature Importance

In order to evaluate what was potentially driving the performance of the NN models, we explored the relative importance of different features in the NN model trained on crowds of 25 via SHAP analysis (Lundberg and Lee 2017).[34] SHAP values (based on the game theoretic concept of Shapley values) estimate the average marginal contribution of a feature—essentially permuting over all the possible orderings of feature inclusion, estimating marginal effects, and then averaging them to attempt to estimate the contribution of a feature to the prediction. This is in contrast to standard permutation testing,[35] which only measures the marginal loss in performance due to withdrawing a feature. Consequently, SHAP is able to provide estimates of feature importance that take into account the importance of the feature with respect to other features—thus potentially compensating for features that may be very similar. Features that may appear to have minimal importance in permutation testing may in fact be more important in making predictions when their relative contribution is taken into account.

SHAP values for the NN model of crowds of 25 are displayed in Figure 2 on the following page, where each dot is an individual prediction. Specifically this means that each dot is based on the prediction the NN made based on one crowd of 25 individuals. Dots are stacked on one another when they would otherwise occupy the same space on the graph. The horizontal axis measures how much that feature, for that observation, contributed to increasing or decreasing the final predicted value of the algorithm for that crowd. In this case, an increase corresponds to a greater likelihood the algorithm predicted the article was fake news. Finally, the color of the dot indicates if the value was low or high relative to the mean for that feature.

One of the primary advantages of NNs is their ability to discover and address nonlinearities that may elude researchers. Our earlier analysis strongly suggested the signal from the crowd was not easily addressed with a linear approach, as the elastic net was essentially unable to make meaningful predictions with crowd data, and simple aggregation techniques were similarly limited. Consequently, some unintuitive behaviors from NNs are possible. In this case, our SHAP analysis produced individual estimates of feature importance and impact that fit expectations.

The most important feature for this model was the ratio of true to false and CND responses.

33. When baseline measures are high, for example 95%, and false positive or false negative rates exceed the rate of the minority label, in this case 5%, then accuracy will almost always be below baseline performance.
34. For SHAP analysis of the NN model trained on crowds of 10 with high political knowledge, see Appendix A.6.
35. See Appendix A.7 for permutation analysis.

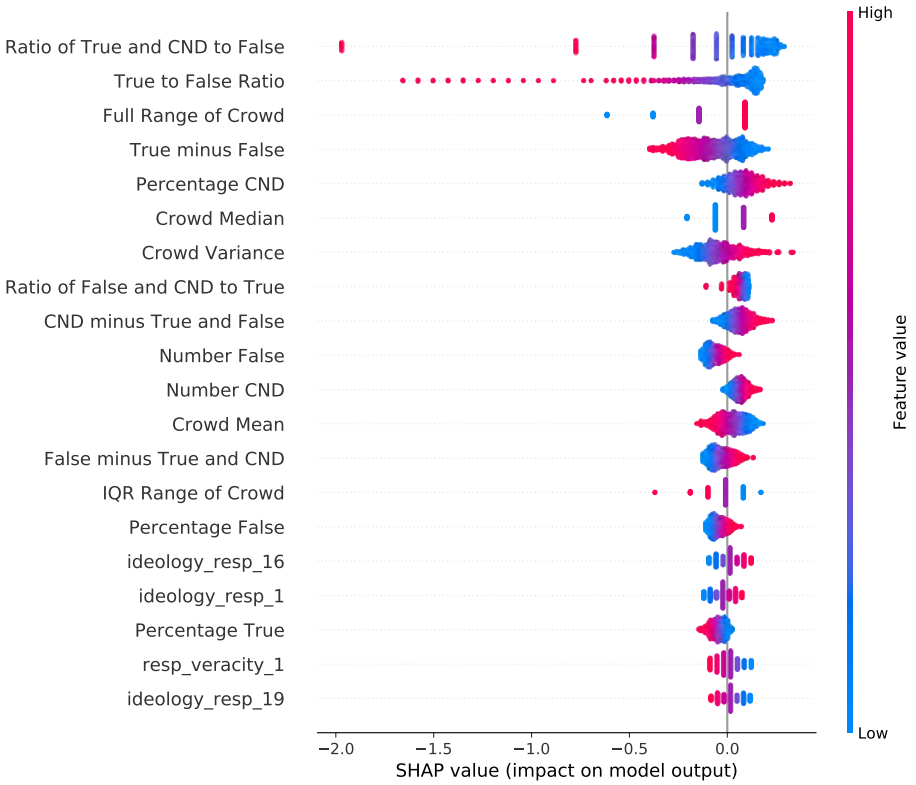



Figure 2: We use SHAP analysis for the NN model of crowds of 25 as a method for identifying feature importance. This graph shows the impact of each feature on every prediction the model made, with each dot representing the impact of that feature (in the y axis), on a specific observation. The x axis indicates the impact that feature had on that observation's prediction, while the color indicates the actual value of that feature for that observation. For example, the dots to the right of "crowd variance" show that high values (those in red) increased the model output value. As the model labeled false articles as "1" and not false as "0", higher values indicate more likely false. Thus higher crowd variance values increased the likelihood the model predicted an article was false.

The lower the value of this ratio, the more likely the model was to predict an article was false, precisely as would be expected. The other top feature— "true to false" —follows the same pattern. Also of note, crowd variance and "Full Range of Crowd" were positively related to a prediction of false: the greater the variance in answers, the more likely the model was to predict false. Somewhat surprisingly, the median and mean of the crowd were not among the most important features.

SHAP analysis for this model suggests that this NN largely relied on features in intuitive ways. At the same time, other ML models, with access to the same features, were unable to achieve the same levels of accuracy as NN-based methods, which are, compared to the other ML methods, the most complex, opaque, and nonlinear. This suggests that translating the signal of the crowd into usable information may not be a simple or linear process.

## 6 Discussion and Conclusion

Using crowds of lay people is both pragmatically and normatively attractive for addressing the diffusion of false or misleading news. *Pragmatically*, it provides a method that can potentially match the scale and velocity of the current online news ecosystem. *Normatively*, it provides a potentially more inclusive and community-based approach to regulating content than simply allowing a small group of highly trained experts to decide what is false or misleading.

The results presented here illustrate both the pitfalls and promise of such an approach. Most notably, no method was able to meet or exceed the relatively low standard of the PFC benchmark for news from low-quality sources, which, as discussed below, likely underestimates performance at this task under normal professional circumstances. Crowds of lay persons with straightforward transparent rules were simply unable to approach professional fact-checkers in their performance. Select crowds, specifically crowds composed of those with high political knowledge, were able to roughly halve the gap between baseline performance and the performance of professional fact-checkers.

Despite the limits of simple rules applied to crowds, some ML-based methods did approach the PFC benchmark using a crowd of 10 untrained lay people, although there are significant caveats to this claim, which we address below. Importantly, further analysis would be necessary to learn if even the performance we observed would be robust to a change in setting. We have found that these ML models can approach PFC performance only when their training environment strongly matches the environment in which they are applied, but real-world applications would almost by definition involve applying the models in environments that have evolved in new directions from the ones in which they were trained. In addition, the performance measured by professional fact-checkers in this study is likely an underestimation of their true performance. The PFCs, as discussed, did not perform a complete professional fact-check of the articles (although they were all professional fact-checkers!) for this study. They were, in essence, simply the most expert population available to perform an evaluation of these articles. If these individuals were given the same task, but with the time and resources to do a complete and thorough professional fact-check, their agreement rate would likely be higher. Therefore the finding that ML-based methods approached the PFC standard come with strong qualifications.

On a more practical level, there are three significant tradeoffs to be made when using crowds to address false or misleading content. First, there is a tradeoff between performance and transparency. The most transparent and easily explained systems that directly translate crowd evaluations to a final prediction—what we call simple rules—per-

form relatively poorly at assessing the veracity of articles. Even relatively large crowds of different types and sizes are inferior to the performance of one PFC. Parler, the alternative social network, employed a simple crowd-based system for content moderation in which a jury of five users evaluated content; the system required four jurors to be in agreement in order to flag a piece of content (Fried 2020). Our findings suggest this approach likely prioritized transparency over accuracy.

Our NN method for aggregating crowds achieved the best performance, almost meeting the PFC benchmark. Though ML methods offer superior performance to simple rules, this improved performance comes with a more opaque method that makes content decisions potentially more difficult to scrutinize and evaluate. Any system that employs lay people to identify the veracity of news content will necessarily need to find a balance between the normatively more satisfying transparency of simple rules and the pragmatic value of ML-based methods.

Our findings also show that there is a tradeoff between representativeness and performance. We find that methods that utilize the most representative crowds, where all respondents in our data are used and chosen at random to participate, are inferior to systems where we limit crowds to those with greater political knowledge. We find that using only crowds with high levels of political knowledge increases accuracy over using both simple rules and ML-based methods. While this paper only examines one selection criteria—political knowledge—it is possible other selection criteria could further improve performance.

This tradeoff again illustrates that normative preferences for representative crowds are in tension with the goals of the task. It is unambiguous that expertise matters—one PFC can, on average, better predict the veracity of an article than a crowd of 25 lay persons. Our research shows that even among lay people, differences in political knowledge are sufficient to translate into differences in crowd performance. Once again, any system that employs crowds for this task will have to address and balance these potentially competing concerns.

Finally, there is a tradeoff between false and true positive and negative rates. Any fact-checking system—whether using trained PFCs or crowds of laypeople—will have to balance which errors are more acceptable for the platform and the community. In this study, the highest performing simple aggregation rules and ML models misclassified articles rated "not false" by the PFCs as "false" almost 30% of the time[36]. Consequently, if platforms acted on these recommendations (e.g., Rosen 2019), they would downweight or ban large amounts of truthful articles. Taken together, our results suggest that a crowdsourced approach is unlikely to be able to replace a journalistic fact-checking system; nonetheless, it remains possible that crowdsourcing could potentially be part of a larger pipeline, although this potential is not explored here.

Even accounting for these tradeoffs, there is the potential that a crowdsourced system that uses machine learning—and thus is trained on a specific set of articles—could fall prey to a rapidly changing information environment. The introduction of new informational dynamics, especially around salient and dynamic events such as a pandemic, election, or terror attack, would likely produce challenges for an ML-based system. The model would make predictions on a new sample of articles with different characteristics (e.g., different distribution of false versus not false; different ideological fault lines) than the sample of articles used as training data, and it is unclear how well those models would perform. And while dynamic informational environments may introduce limitations for ML-based systems, it is often those contexts in which accurate fact-checking is most

36. This sentence has been updated from the orginally published version of this article to reflect the updated calculations made to Table 2

exigent.

If a crowd-based system were implemented, there is also the risk that strategic motivations would lead users to intentionally misevaluate articles. These motivations may not have been present in an online survey in which respondent evaluations did not have the potential to impact the diffusion of information. However, if a social media platform were to employ ordinary users to evaluate articles, there might be incentives to miscategorize information for political (or other) ends, thus possibly further lowering the accuracy of a crowd-based system.

Despite these tradeoffs and shortcomings, there is clearly some information signal in crowds. With crowds of sufficient size, some demographic or metadata on crowd composition, and appropriate aggregation methods—in this case, a neural network based on simple features—performance is significantly improved relative to baseline and approaches the PFC benchmark. Moreover, we were able to achieve this level of performance despite significant limitations faced by an academic research team that would likely not be faced by a large social media platform, most notably a somewhat limited number of articles that were used to evaluate performance when compared to the volume of potential news such a system could eventually be used to evaluate. Given that ML algorithms would almost certainly improve significantly with more data, we expect that the results we reported here provide a baseline level of performance and could be improved upon.

Overall, there are many approaches being proposed to control the dissemination of false or misleading online news. The first step in all such methods requires the identification of problematic articles. Our research suggests that real-time crowdsourced fact-checking does possess genuine information for which social media platforms or others may find potential uses. So long as the tradeoffs are accounted for and the limitations are recognized, such a system may offer a viable tool—as part of what certainly needs to be a larger toolkit—to combat the spread of online misinformation. Nonetheless, we find little evidence that a crowd-based approach, on its own, is sufficient to identify false news.

## References

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31, no. 2 (May): 211–236. https://doi.org/10.1257/jep.31.2.211.

Allcott, Hunt, Matthew Gentzkow, and Chuan Yu. 2019. "Trends in the diffusion of misinformation on social media." *Research & Politics* 6, no. 2 (April): 205316801984855. https://doi.org/10.1177/2053168019848554.

Allen, Jennifer, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. "Scaling up fact-checking using the wisdom of crowds." *Science Advances* 7, no. 36 (September 3, 2021): eabf4393. https://doi.org/10.1126/sciadv.abf4393.

Allen, Jennifer, Baird Howland, Markus Mobius, David Rothschild, and Duncan J. Watts. 2020. "Evaluating the fake news problem at the scale of the information ecosystem." *Science Advances* 6, no. 14 (April): eaay3539. https://doi.org/10.1126/sciadv.aay3539.

Aslett, Kevin, William Godel, Zeve Sanderson, Nathaniel Persily, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker. 2021. "An Externally Valid Method for Assessing Belief in Popular Fake News." *Unpublished Manuscript.*

Bell, Emily. 2019. "The Fact-Check Industry." Columbia Journalism Review. https://www.cjr.org/special:report/fact-check-industry-twitter.php/.

Boland, Philip J. 1989. "Majority Systems and the Condorcet Jury Theorem." *The Statistician* 38 (3): 181. https://doi.org/10.2307/2348873.

Bolsen, Toby, James N. Druckman, and Fay Lomax Cook. 2014. "The Influence of Partisan Motivated Reasoning on Public Opinion." *Political Behavior* 36, no. 2 (June): 235–262. https://doi.org/10.1007/s11109-013-9238-0.

Budescu, David V., and Eva Chen. 2015. "Identifying Expertise to Extract the Wisdom of Crowds." *Management Science* 61, no. 2 (February): 267–280. https://doi.org/10.1287/mnsc.2014.1909.

Clayton, Katherine, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, et al. 2020. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media." *Political Behavior* 42, no. 4 (December): 1073–1095. https://doi.org/10.1007/s11109-019-09533-0.

Clemm von Hohenberg, Bernhard. 2020. *Truth and Bias: Robust findings?* Preprint. Open Science Framework, July 30, 2020. https://doi.org/10.31219/osf.io/yj2rn.

Coleman, Keith. 2021. "Introducing Birdwatch, a community-based approach to misinformation." Twitter, January 25, 2021. https://blog.twitter.com/en:us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.

Collins, Ben. 2020. "Twitter is Testing New Ways to Fight Misinformation — Including a Community-Based Points System." NBCNews.com, February 20, 2020. https://www.nbcnews.com/tech/tech-news/twitter-testing-new-ways-fight-misinformation-including-community-based-points-n1139931.

Comber, Alexis, Peter Mooney, Ross S. Purves, Duccio Rocchini, and Ariane Walz. 2016. "Crowdsourcing: It Matters Who the Crowd Are. The Impacts of between Group Variations in Recording Land Cover." Edited by Timothy C. Matisziw. *PLOS ONE* 11, no. 7 (July 26, 2016): e0158329. https://doi.org/10.1371/journal.pone.0158329.

CrowdTangle Team. 2021. "CrowdTangle."

Druckman, James N., and Mary C. McGrath. 2019. "The evidence for motivated reasoning in climate change preference formation." *Nature Climate Change* 9, no. 2 (February): 111–119. https://doi.org/10.1038/s41558-018-0360-1.

Duke Reporters' Lab. 2016. "Fact-Checking," October 17, 2016. https://reporterslab.org/fact-checking/.

Facebook. 2021. *Form 10-Q.* https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/dcf20641-cba6-4b5c-b60e-4b40b52811a4.pdf.

Fazio, Lisa K., Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. 2015. "Knowledge does not protect against illusory truth." *Journal of Experimental Psychology: General* 144, no. 5 (October): 993–1002. https://doi.org/10.1037/xge0000098.

Fried, Ina. 2020. "Parler Exec Defends Allowing Conspiracy Theories, Misinformation." Axios, November 13, 2020. https://www.axios.com/newsletters/axios-login-3eb66608-1976-4097-97a0-f4d532f2a1b5.html?utm:source=newsletter.

Goldstein, Daniel G., Randolph Preston McAfee, and Siddharth Suri. 2014. "The wisdom of smaller, smarter crowds." In *Proceedings of the fifteenth ACM conference on Economics and computation,* 471–488. EC '14: ACM Conference on Economics and Computation. Palo Alto California USA: ACM, June. https://doi.org/10.1145/2600057.2602886.

Golub, Benjamin, and Matthew O Jackson. 2010. "Naïve Learning in Social Networks and the Wisdom of Crowds." *American Economic Journal: Microeconomics* 2, no. 1 (February 1, 2010): 112–149. https://doi.org/10.1257/mic.2.1.112.

Graves, Lucas. 2016. *Deciding what's true: the rise of political fact-checking in American journalism.* New York: Columbia University Press.

Griffiths, Thomas L., and Joshua B. Tenenbaum. 2006. "Optimal Predictions in Everyday Cognition." *Psychological Science* 17, no. 9 (September): 767–773. https://doi.org/10.1111/j.1467-9280.2006.01780.x.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake news on Twitter during the 2016 U.S. presidential election." *Science* 363, no. 6425 (January 25, 2019): 374–378. https://doi.org/10.1126/science.aau2706.

Grofman, Bernard, Guillermo Owen, and Scott L. Feld. 1983. "Thirteen theorems in search of the truth." *Theory and Decision* 15, no. 3 (September): 261–278. https://doi.org/10.1007/BF00125672.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." *Science Advances* 5, no. 1 (January 18, 2019): eaau4586. https://doi.org/10.1126/sciadv.aau4586.

Guess, Andrew, Brendan Nyhan, and Jason Reifler. 2018. "Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign." *European Research Council* 9 (3): 4.

Haque, Md Mahfuzul, Mohammad Yousuf, Zahedur Arman, Md Main Uddin Rony, Ahmed Shatil Alam, Kazi Mehedi Hasan, Md Khadimul Islam, and Naeemul Hassan. 2018. "Fact-checking Initiatives in Bangladesh, India, and Nepal: A Study of User Engagement and Challenges." *arXiv:1811.01806 [cs]* (November 5, 2018). arXiv: 1811.01806. http://arxiv.org/abs/1811.01806.

Herzog, Stefan M, and Ralph Hertwig. 2011. "The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition." *Judgment and Decision making* 6 (1): 58–72. http://journal.sjdm.org/11/rh18/rh18.html.

Horwitz, Jeff. 2020. "Facebook's Fact Checkers Fight Surge in Fake Coronavirus Claims." *Wall Street Journal* (March 30, 2020). Accessed April 13, 2020. https://www.wsj.com/articles/facebooks-fact-checkers-fight-surge-in-fake-coronavirus-claims-11585580400.

Hueffer, Karsten, Miguel A Fonseca, Anthony Leiserowitz, and Karen M Taylor. 2013. "The wisdom of crowds: Predicting a weather and climate-related event." Publisher: Society for Judgment and Decision Making, *Judgment and Decision Making* 8 (2): 91–105.

Kim, Jooyeon, Behzad Tabibian, Alice Oh, Bernhard Schölkopf, and Manuel Gomez-Rodriguez. 2018. "Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation." In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining,* 324–332. WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining. Marina Del Rey CA USA: ACM, February 2, 2018. https://doi.org/10.1145/3159652.3159734.

Laan, Andres, Gabriel Madirolas, and Gonzalo G. de Polavieja. 2017. "Rescuing Collective Wisdom when the Average Group Opinion Is Wrong." *Frontiers in Robotics and AI* 4 (November 6, 2017): 56. https://doi.org/10.3389/frobt.2017.00056.

Lim, Chloe. 2018. "Checking how fact-checkers check." *Research & Politics* 5, no. 3 (July): 205316801878684. https://doi.org/10.1177/2053168018786848.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems,* edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Mannes, Albert E., Jack B. Soll, and Richard P. Larrick. 2014. "The wisdom of select crowds." *Journal of Personality and Social Psychology* 107 (2): 276–299. https://doi.org/10.1037/a0036677.

Mark Zuckerberg. 2019. Facebook. https://www.facebook.com/zuck/videos/10106612617413491/.

Moore, Tyler, and Richard Clayton. 2008. "Evaluating the Wisdom of Crowds in Assessing Phishing Websites." In *Financial Cryptography and Data Security,* edited by Gene Tsudik, 5143:16–30. ISSN: 0302-9743, 1611-3349 Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-85230-8:2.

NewsGuard. 2021. "Coronavirus Misinformation Tracking Center," May 13, 2021. https://www.newsguardtech.com/special-reports/coronavirus-misinformation-tracking-center/.

Paszke, Adam, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. 2019. "Pytorch: An imperative style, high-performance deep learning library." *Advances in neural information processing systems* 32:8026–8037.

Pennycook, Gordon, Adam Bear, Evan T Collins, and David G Rand. 2017. "The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings." Publisher: INFORMS, *Management Science.*

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand. 2018. "Prior exposure increases perceived accuracy of fake news." *Journal of Experimental Psychology: General* 147, no. 12 (December): 1865–1880. https://doi.org/10.1037/xge0000465.

Pennycook, Gordon, and David G. Rand. 2019a. "Fighting misinformation on social media using crowdsourced judgments of news source quality." *Proceedings of the National Academy of Sciences* 116, no. 7 (February 12, 2019): 2521–2526. https://doi.org/10.1073/pnas.1806781116.

———. 2019b. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning." *Cognition* 188 (July): 39–50. https://doi.org/10.1016/j.cognition.2018.06.011.

Perrin, Andrew, and Monica Anderson. 2020. "Share of U.S. adults using social media, including Facebook, is mostly unchanged since 2018." *Pew Research Center* (July 31, 2020). https://www.pewresearch.org/fact-tank/2019/04/10/share-of-u-s-adults-using-social-media-including-facebook-is-mostly-unchanged-since-2018/.

Politfact. 2020. "The Principles of the Truth-O-Meter: PolitiFact's methodology for independent fact-checking." https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/.

Raddick, M. Jordan, Georgia Bracey, Pamela L. Gay, Chris J. Lintott, Phil Murray, Kevin Schawinski, Alexander S. Szalay, and Jan Vandenberg. 2010. "Galaxy Zoo: Exploring the Motivations of Citizen Science Volunteers." *Astronomy Education Review* 9, no. 1 (December). https://doi.org/10.3847/AER2009036.

Rosen, Guy. 2019. "Remove, Reduce, Inform: New Steps to Manage Problematic Content." About Facebook, April 10, 2019. https://about.fb.com/news/2019/04/remove-reduce-inform-new-steps/.

Sharifi, Mehrbod, Eugene Fink, and Jaime G. Carbonell. 2011. "SmartNotes: Application of crowdsourcing to the detection of web threats." In *2011 IEEE International Conference on Systems, Man, and Cybernetics,* 1346–1350. 2011 IEEE International Conference on Systems, Man and Cybernetics - SMC. Anchorage, AK, USA: IEEE, October. https://doi.org/10.1109/ICSMC.2011.6083845.

Simmons, Joseph P., Leif D. Nelson, Jeff Galak, and Shane Frederick. 2011. "Intuitive Biases in Choice versus Estimation: Implications for the Wisdom of Crowds." *Journal of Consumer Research* 38, no. 1 (June 1, 2011): 1–15. https://doi.org/10.1086/658070.

Simoiu, Camelia, Chiraag Sumanth, Alok Mysore, and Sharad Goel. 2019. "Studying the "Wisdom of Crowds" at Scale." Section: Technical Papers, *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7, no. 1 (October 28, 2019): 171–179. https://ojs.aaai.org/index.php/HCOMP/article/view/5271.

Surowiecki, James. 2005. *The wisdom of crowds.* Anchor Books.

Tucker, Joseph D., Suzanne Day, Weiming Tang, and Barry Bayus. 2019. "Crowdsourcing in medical research: concepts and applications." *PeerJ* 7 (April 12, 2019): e6762. https://doi.org/10.7717/peerj.6762.

Van Bavel, Jay J., and Andrea Pereira. 2018. "The Partisan Brain: An Identity-Based Model of Political Belief." *Trends in Cognitive Sciences* 22, no. 3 (March): 213–224. https://doi.org/10.1016/j.tics.2018.01.004.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The spread of true and false news online." *Science* 359, no. 6380 (March 9, 2018): 1146–1151. https://doi.org/10.1126/science.aap9559.

Wineburg, Sam, and Sarah McGrew. 2017. *Lateral Reading: Reading Less and Learning More When Evaluating Digital Information.* SSRN Scholarly Paper ID 3048994. Rochester, NY: Social Science Research Network, October 6, 2017. https://doi.org/10.2139/ssrn.3048994.

Wittenberg, Chloe, Adam J Berinsky, Nathaniel Persily, and Joshua A Tucker. 2020. "Misinformation and its correction." Publisher: Cambridge University Press New York, *Social Media and Democracy: The State of the Field, Prospects for Reform* 163.

Woolley, Anita Williams, Ishani Aggarwal, and Thomas W. Malone. 2015. "Collective Intelligence and Group Performance." *Current Directions in Psychological Science* 24, no. 6 (December): 420–424. https://doi.org/10.1177/0963721415599543.

Woolley, Anita Williams, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. "Evidence for a Collective Intelligence Factor in the Performance of Human Groups." *Science* 330, no. 6004 (October 29, 2010): 686–688. https://doi.org/10.1126/science.1193147.

## Authors

**William Godel** is a PhD candidate at the NYU Wilf Family Department of Politics and currently an intern at Facebook Core Data Science.

**Zeve Sanderson** is the Executive Director of the NYU Center for Social Media and Politics.

**Kevin Aslett** is a postdoctoral scholar at the NYU Center for Social Media and Politics.

**Jonathan Nagler** is a Professor of Politics at the NYU Wilf Family Department of Politics and a Director of the NYU Center for Social Media and Politics.

**Richard Bonneau** is a Professor of Biology and Computer Science at NYU and a Director of the NYU Center for Social Media and Politics.

**Nathaniel Persily** is the James B. McClatchy Professor of Law at Stanford Law School, with appointments in the departments of Political Science, Communication, and FSI.

**Joshua Tucker** is a Professor of Politics at the NYU Wilf Family Department of Politics, Director of the Jordan Center for the Advanced Study of Russia, and a Director of the NYU Center for Social Media and Politics.

## Acknowledgements

This research was approved after an exempt review by the NYU IRB Board (NYU-IRB-FY2019-3511).

We thank David Rand, Paul Resnick, and their research teams for their comments and feedback. We thank the Hewlett Foundation for their generous support of this project. The NYU Center for Social Media and Politics is also generously supported by the Knight Foundation, the Charles Koch Foundation, Craig Newmark Philanthropies, the Russell Sage Foundation, the Bill and Melinda Gates Foundation, and the Siegel Family Foundation.

## Data Availability Statement

Replication files are available at https://github.com/SMAPPNYU/crowdsourcing_factchecking. Our preregistration of the article collection mechanism can be found at https://osf.io/dp8ze/.

## Ethical Standards

This research was approved after an exempt review by the NYU IRB Board (NYU-IRB-FY2019-3511).

As discussed in Footnote 12, during our data collection covering the pandemic, we delayed sending articles to respondents for evaluation by 24 hours. During this period, PFCs assessed the articles. This delay enabled us to immediately communicate the PFC evaluations to respondents once they had completed their own evaluations, thus reducing the potential for an article to misinform a respondent about pandemic-related topics.

## Funding Statement

We thank the Hewlett Foundation for their generous support of this project. The NYU Center for Social Media and Politics is also generously supported by the Knight Foundation, the Charles Koch Foundation, Craig Newmark Philanthropies, and the Siegel Family Foundation.

## Keywords

Crowdsourcing, fake news, fact-checking, machine learning, misinformation

# A Appendices

## A.1 Article selection

We set up the following five streams: liberal mainstream news; conservative mainstream news, liberal low-quality news, conservative low-quality news, and low-quality news sites with no clear political orientation. Our mainstream news list is built from the top 100 news sites by US consumption between 2016 and 2019 from Microsoft Research's Project Ratio, which we then classified by ideological lean (liberal or conservative) and selected the top ten websites from each ideological category. To construct the low-quality streams, we use all low-quality news sources from Allcott, Gentzkow, and Yu (2019), which itself combines well-known lists of false news websites; a total of 99 websites were active at the time of our study. We then classified these low-quality sources by partisan lean (liberal, conservative, or unclear).[37] Finally, we selected the most popular article over the previous 24 hours from each of the streams, using CrowdTangle for the mainstream sources and Feedly RSS feeds for the low-quality sources.[38] On CrowdTangle, a data analytics platform owned by Facebook, we selected the article from the mainstream news streams that had the highest overperforming score (CrowdTangle Team 2021).[39] On Feedly, a news aggregator, we selected the article from each of the three low-quality news streams that was most popular as measured by interactions on Feedly and on social media platforms.[40] Taken together, the five articles provide a diverse sample of true and potentially problematic popular articles published on a given day, balanced between ideological lean.

For each of the 45 days of our study, which ran from November 2019 to June 2020, we therefore sent out five articles within 24 hours of their publication for evaluation (see Appendix A.5 for a detailed study overview). Here, we were concerned with only those articles coming from low-quality or questionable news outlets and so did not incorporate the articles from the mainstream news streams. Altogether, our sample for this paper consists of 135 articles from low-quality sources.

## A.2 Survey Questions

From the survey, information from the following questions was used to provide respondent evaluations of questions.

For the categorical evaluation of the article, the question was:

What is your assessment of the central claim in the article?

- **True.** The central claim you are evaluating is factually accurate.
- **Misleading and/or False.** Misleading: the central claim takes out of context, misrepresents or omits evidence. False: The central claim is factually inaccurate.
- **Could Not Determine.** You do not feel you can judge whether the central claim is true, false, or mislead.

For the veracity score of an article, respondents were asked:

37. For both mainstream and low-quality websites, two research assistants coded each website for partisan lean; if they were split, a third coder was used to break the tie.
38. Because Facebook has banned many of the pages associated with low-quality domains known to produce false news, CrowdTangle could not be used to measure popularity for the three low-quality streams.
39. For more information on CrowdTangle's overperforming measure, see https://help.crowdtangle.com/en/articles/1141056-how-is-overperforming-calculated.
40. For more information on Feedly's popularity measure, see https://feedly.com/i/entry/NBHkLLj8YGLLEyGA+0mSpEvCPJ4mKcxBYbHNP0YqkfY=:1570c7dc2d6:bd400f4:e3157ec0

Now that you have evaluated the article, we are interested in the strength of your opinion. Please rank the article on following scale:

1. **- Definitely NOT TRUE**
2. 
3. 
4. 
5. 
6. 
7. **- Definitely TRUE**

For the political ideology response, the question was:

Where would you place yourself on this scale?

- Extremely Conservative
- Conservative
- Slightly Conservative
- Moderate: Middle of the road
- Slightly Liberal
- Liberal
- Extremely Liberal
- Haven't thought much about it

For political knowledge, respondents were asked to answer four questions. Those questions were:

1. Which party currently has the most members in the US House of Representatives in Washington, DC?
    - Republican Party
    - Democratic Party
2. Who is the current Speaker of the US House of Representatives?
    - Nancy Pelosi
    - Mitch McConnell
    - Chuck Schumer
    - Steve Scalise
3. What job or political office is now held by Boris Johnson?
    - Nancy Prime Minister of Australia
    - Prime Minister of Canada
    - Prime Minister of the United Kingdom
    - Secretary-General of the United Nations
4. Who is the current US Secretary of State?

- Rex Tillerson
- John Sullivan
- Jim Mattis
- Michael Pompeo

### A.3 PFC Agreement Table

| Time Period | Agreement | Fleiss' Kappa |
|---|---|---|
| Period One Data | 45.45 | .41 |
| Period Two Data Covid | 42.5 | .46 |
| Period Two Data Noncovid | 33.33 | .35 |

### A.4 Bayes' Rule

As explained in the article, we use Bayes' rule to predict if an article is false or not. Below is the equation used for Bayes' rule and a written explanation of the terms:

$$P(\text{False}|\text{Observations}) = \frac{P(\text{False}) * P(\text{Observations}|\text{False})}{P(\text{Observations})}$$

Where $P(\text{False})$ is the prior and:

1. $P(\text{Observations}|\text{Not False})$
2. $P(\text{Observations}|\text{False})$
3. $P(\text{Observations}) = P(\text{False}) * P(\text{Observations}|\text{False}) + (1 - P(\text{False})) * P(\text{Observations}|\text{Not False})$

(1) and (2) are multinomial distributions where the proportions of each distribution are calculated by counting the number of each answer type from respondents ("true," "false/misleading," "could not determine"), corresponding to the article type ("false" or "not false"). For example, calculating (2) for a given set of observations is done by calculating what proportion of respondents evaluated all "false" articles in the training set as "false/misleading," what proportion evaluated "false" articles as "true," and what proportion evaluated "false" articles as "could not determine." These three proportions are used to generate a multinomial distribution. The actual probability of observing a given distribution of evaluations (labeled observations in the formula) from one crowd is then calculated by inputting the proportions of the crowd that chose each answer into the probability mass function of the multinomial distribution. (3) is then calculated by using the law of total probability, where (2) is multiplied by the prior and added to one minus the prior times (1). Once the prior and conditional distributions are estimated from the training set, for any observation in the test set all we must do to calculate the Bayes' probability the article under consideration is false is to calculate the distribution of answers from the crowd and apply Bayes' rule. For the purpose of predicting, any value over .5 was a "false" prediction and anything under was a "not false" prediction.

### A.5 Overview of Study Design

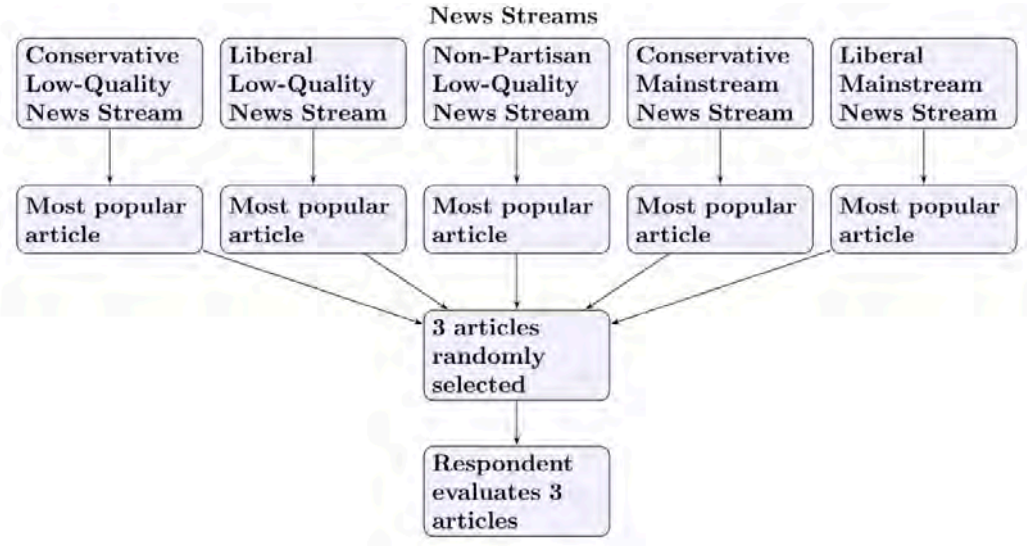



Figure 3: This diagram depicts the article selection method we used to source the most popular article from five news streams and collect respondent evaluations.

### A.6 Full List of Features

For convenience of explanation, let T be the number of respondents in a given crowd that answered "True" in their article evaluation, let F be the number of respondents who evaluated "False/Misleading," and let C be the number of respondents who answered "Could Not Determine."

The full list of features, including their names and how they were composed, used in the ML algorithms were:

1. Crowd Mean = Mean of Crowd Veracity Score
2. Crowd Median = Median of Crowd Veracity Score
3. Crowd Variance = Variance of Crowd Veracity Score
4. IQR Range of Crowd = Interquartile Range of Crowd Veracity Score
5. Full Range of Crowd = Range of Crowd Veracity Scores
6. Mode of Crowd T = Binary Measure of if the Mode of the Crowd is "True"
7. Mode of Crowd F = Binary Measure of if the Mode of the Crowd is "False/Misleading"
8. Bayes' Post = Bayes' Posterior Probability of Being "False/Misleading"
9. Number True = $T$
10. Number False = $F$
11. Number CND = $C$
12. True Minus False = $T - F$
13. True to False Ratio = $\dfrac{T}{(F+1)}$

14. CND Minus True and False = $C - T - F$
15. True Minus CND and False = $T - C - F$
16. False Minus CND and True = $F - T - C$
17. Percentage Could Not Determine = $\frac{C}{(C + F + T)}$
18. Percentage True = $\frac{T}{(C + F + T)}$
19. Percentage False = $\frac{F}{(C + F + T)}$
20. False and Could Not Determine to True = $\frac{(F + C)}{(T + 1)}$
21. True and Could Not Determine to False = $\frac{(T + C)}{(F + 1)}$

In addition to the above features, for each individual respondent in the crowd, the algorithms were provided with the categorical response of that respondent, the veracity score of that respondent, and the political orientation of that respondent (coded from -3 to 3).

### A.7 Additional SHAP Analysis and Permutation Testing

In permutation testing, one feature at a time is randomly permuted and then the performance of the model is evaluated, thus providing an estimate of the performance of the model when that feature is useless noise. This process is repeated for each feature 100 times, and the average change in performance of the model is used as the final metric. Here permutation testing was performed on an NN model trained on crowds of 25 and an NN model trained on crowds of 10 with high political knowledge.

No feature in either model explored leads to large drops in performance according to permutation testing. Randomizing an individual feature never results in a loss of performance much larger than one percentage point. Figure 5 shows the percentage of the total performance gain that was lost by randomizing that feature.

These findings strongly suggest that ML-based methods achieve their increases in performance by using many features in conjunction with one another. This is precisely the advantage of NNs and, when combined with our earlier findings, illuminates why simple aggregation rules are limited in their ability to predict the veracity of articles. This is also unsurprising given the nature of features used in these models. The features are strongly related; many are mathematical manipulations or functions of one another. This provides an abundance of perspectives on the same data for the models to utilize, but it also means it is unlikely any one feature is especially important in making predictions.

### A.8

We could also remove members of the crowd who respond “Could Not Determine” and then recompute the mode from those who made a clear determination as either “True” or “False/Misleading.” Dropping “Could Not Determine” responses generates the slimmest of improvements compared to the original mode measure, with performance ranging from 53.2% to 57.6%.

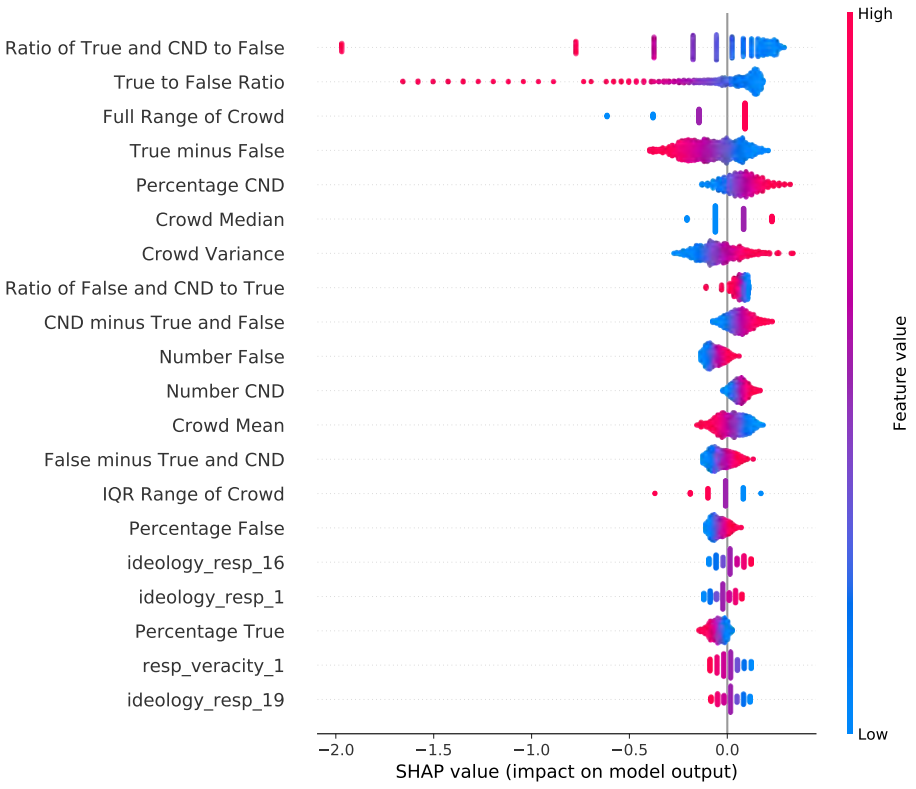



Figure 4: We use SHAP analysis for the NN model of crowds of 10 with high political knowledge as an alternative method for identifying feature importance. This graph shows the impact of each feature on every prediction the model made, with each dot representing the impact of that feature (in the y axis), on a specific observation. The x axis indicates the impact that feature had on that observation's prediction, while the color indicates the actual value of that feature for that observation. For example, the dots to the right of "crowd variance" show that high values (those in red) increased the model output value. As the model labeled false articles as "1" and not false as "0," higher values indicate more likely false. Thus higher crowd variance values increased the likelihood the model predicted an article was false.

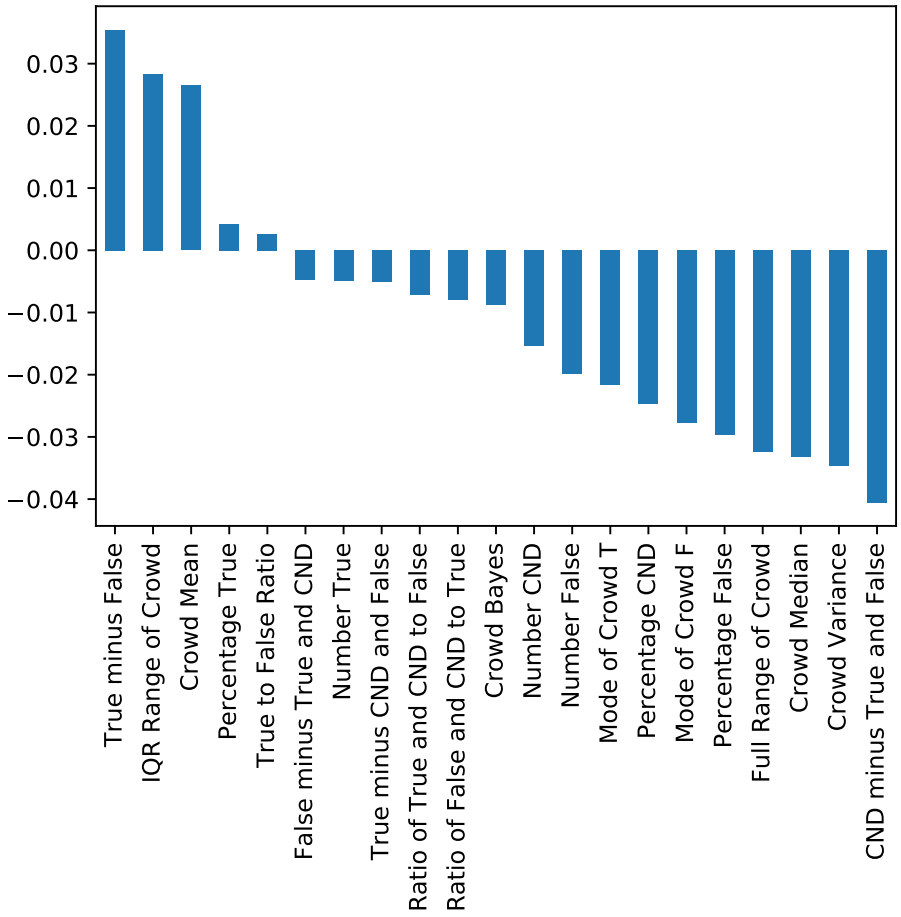



Figure 5: The percentage performance loss by randomizing each feature. Positive values indicate the percent of the performance gain that was lost by randomizing that feature, while negative values indicate the gain, likely due to overfitting on the feature.

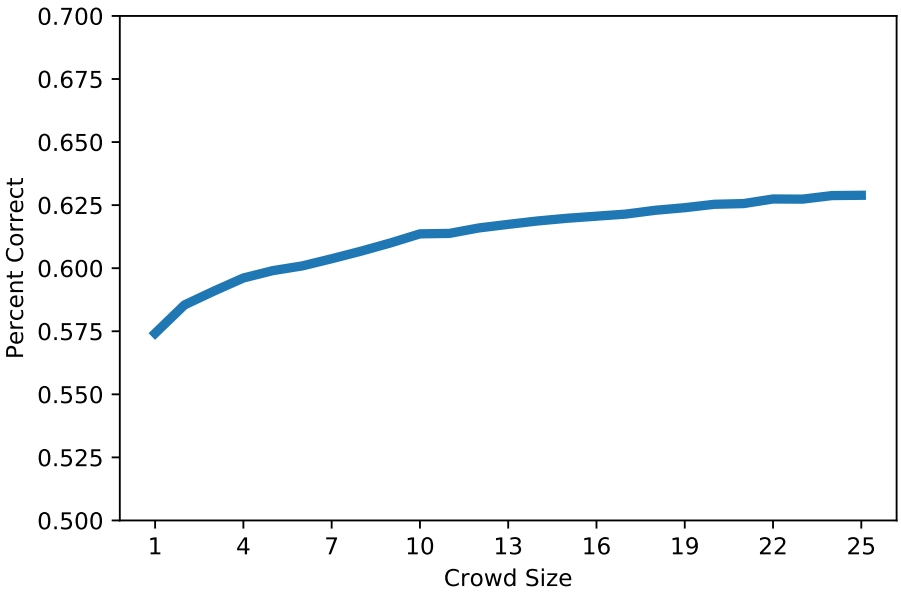



Figure 6: Dropping all respondents in our crowds who evaluated an article as "Could Not Determine," the proportion of crowds, increasing in size by 1, that match the modal response from PFCs.