



# Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program

Jennifer Allen  
jnallen@mit.edu

Sloan School of Management, MIT  
Cambridge, MA, USA

Cameron Martel  
cmartel@mit.edu

Sloan School of Management, MIT  
Cambridge, MA, USA

David Rand  
drand@mit.edu

Sloan School of Management and  
Department of Brain and Cognitive  
Sciences, MIT  
Cambridge, MA, USA

## ABSTRACT

There is a great deal of interest in the role that partisanship, and cross-party animosity in particular, plays in interactions on social media. Most prior research, however, must infer users' judgments of others' posts from engagement data. Here, we leverage data from Birdwatch, Twitter's crowdsourced fact-checking pilot program, to directly measure judgments of whether other users' tweets are misleading, and whether other users' free-text evaluations of third-party tweets are helpful. For both sets of judgments, we find that contextual features – in particular, the partisanship of the users – are far more predictive of judgments than the content of the tweets and evaluations themselves. Specifically, users are more likely to write negative evaluations of tweets from counter-partisans; and are more likely to rate evaluations from counter-partisans as unhelpful. Our findings provide clear evidence that Birdwatch users preferentially challenge content from those with whom they disagree politically. While not necessarily indicating that Birdwatch is ineffective for identifying misleading content, these results demonstrate the important role that partisanship can play in content evaluation. Platform designers must consider the ramifications of partisanship when implementing crowdsourcing programs.

## CCS CONCEPTS

- Human-centered computing → Empirical studies in collaborative and social computing; Social media;
- Information systems → Crowdsourcing.

## KEYWORDS

crowdsourcing, fact-checking, misinformation

### ACM Reference Format:

Jennifer Allen, Cameron Martel, and David Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5,*



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA  
© 2022 Copyright held by the owner/author(s).  
ACM ISBN 978-1-4503-9157-3/22/04.  
<https://doi.org/10.1145/3491102.3502040>

2022, New Orleans, LA, USA. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3491102.3502040>

## 1 INTRODUCTION

Understanding the role of partisanship in social media interactions is integral to improving online platforms. For example, partisanship underscores potentially harmful online behavior such as toxic political discourse and harassment of counter-partisan politicians and members of the public [30, 31, 39]. Exposure to counter-partisan elites via social media can cause increased polarization [4], and more generally, social media use has been causally linked to polarization: being randomly assigned to deactivate Facebook in the leadup to the 2018 U.S. midterm elections significantly decreased polarization [1].

One common explanation for this seemingly toxic political social media ecosystem is the existence of online "echo chambers," in which users are mostly exposed to content from like-minded others [65]. This idea is largely premised on the observation that people are more likely to be connected to co-partisans online [8, 15], and that shared partisanship causally increases the probability of forming new online connections [45]. Despite being intuitively compelling, however, there is surprisingly little evidence to support the echo chamber hypothesis regarding information exposure. Research finds that connections online are actually less homophilous than offline networks, and the media diets of people on social media are more balanced and moderate than often assumed [7, 24, 28]. Thus, rather than shielding people from interacting with counter-partisans, there is reason to believe that social media actually increases exposure to counter-partisan content.

As a result, it is of substantial importance for researchers to explore how people react to counter-partisan content when they encounter it online. Studies have shown that people are more likely to share news that aligns with their partisanship, regardless of its accuracy [26, 48, 50], and that politicians' tweets about members of the other party - which often evoke anger - receive more shares than tweets about members of their own party [57, 77].

These findings are typically interpreted as implying that users judge cross-partisan content negatively. However, it is often extremely difficult to directly assess how social media users actually perceive and evaluate the content they see online. Instead of direct assessments, researchers typically examine on-platform behaviors (e.g. sharing), which are then treated as proxies for agreement. Yet, recent research has shown that there is often a surprisingly large

disconnect between sharing and belief [21, 50, 51, 63]. As a result, the extent to which social media users actually evaluate counter-partisan content more negatively than co-partisan content remains unclear.

In addition to implications for basic research on social interactions and political psychology, understanding whether users judge counter-partisan content more negatively is also important for social media platforms' efforts to harness the wisdom of user crowds to identify misinformation. Prior work has found that when users are randomly assigned publishers or news headlines to rate, layperson crowds show a high level of agreement with professional fact-checkers [2, 18, 25, 52, 58] - even when they believe their ratings may influence what content is shown by social media companies [20]. However, if users are free to choose what content to rate, partisanship may lead to systematic biases in what posts are chosen, and what ratings are given. Here, we shed new light on the relationship between shared partisanship and the evaluation of other users' content. We do so by leveraging data from Birdwatch, Twitter's recently developed crowdsourced fact-checking platform, which provides clearly quantified data about whether users judge (i) others' tweets as misleading, and (ii) others' comments as helpful [14].

## 1.1 The Birdwatch Platform

Birdwatch operates by allowing participants to identify tweets as misleading or not, write free-response fact-checks of tweets, and evaluate the quality of other participants' fact-checks. When the data for the current research were collected, Birdwatch was in a pilot stage and participation in Birdwatch was available only to a small subset of interested users who applied and were then accepted by Twitter. Twitter aimed to include users from a wide and balanced set of perspectives as pilot participants.

The two main components of Birdwatch are notes and ratings. Notes are the free-response fact-checks participants can write in response to any tweets participants come across and think may or may not be misleading. Notes include various multiple choice questions – most important for this paper is a classification of the tweet as 'Not misleading' or 'Potentially misinformed or misleading' – as well as an open ended text field where participants can explain their classification and include relevant sources which helped them reach their decision. Participants in the Birdwatch pilot can view notes directly on tweets on their Twitter timeline.

The second main component is ratings, which are evaluations of other Birdwatch participants' notes. Participants rate the helpfulness of others' notes, and these ratings are then aggregated by Birdwatch to increase visibility of helpful notes.

For an example tweet, note, and rating aggregation, see Figure 1. Birdwatch also includes a Birdwatch site, where participants and all other Twitter users can view all notes and ratings.

Birdwatch participants can write a note about any tweet they encounter, as well as submit a rating on any note. Additionally, the Birdwatch site has a separate feed of notes which require more ratings for adequate helpfulness aggregation.

## 1.2 The Current Research

In this paper, we examine the relationship between partisanship and behavior on Birdwatch. Importantly, Birdwatch is not focused on political misinformation in particular; Birdwatch users may elect to fact-check any tweet. Furthermore, Birdwatch attempted to reduce partisan motivations by including messaging that emphasized values of building understanding, acting in good faith, and being helpful even to those with whom you disagree. Thus, partisanship may not play a major role in how participants use Birdwatch.

Even if partisanship is associated with fact-checking and helpfulness rating behavior on Birdwatch, it is also unclear *a priori* what relationships may exist. For instance, users may primarily encounter co-partisan content in their newsfeed, and thus may be more likely to evaluate co-partisan rather than counter-partisan tweets. Alternatively, users may focus on, or even actively seek out, counter-partisan tweets to fact-check. Similar dynamics may also play out for helpfulness ratings: Users may preferentially rate notes by co-partisans as helpful; users could rate counter-partisan notes as unhelpful; or some other combination of evaluations. Thus, examining the partisan dynamics at play on Birdwatch helps inform discussions of partisan behavior on social media, and is critical for understanding how to better implement features such as crowd-sourced fact-checking.

To this end, we ask the following research questions:

- (1) Is shared partisanship an important predictor of whether tweets are rated as misleading, and if so, how?
- (2) Is shared partisanship an important predictor of whether notes (fact-checks) are rated as helpful, and if so, how?

## 2 RELATED WORK

### 2.1 Partisanship and Online Behavior

A great deal of work has explored the role of partisanship in online behavior. While the concept of online "echo chambers," which are information environments where consumers are overwhelming exposed to confirmatory views [65], has received a great deal of popular attention, academic consensus on the extent to which echo chambers actually exist online is lacking. On the one hand, research has shown that there is substantial ideological clustering on social media sites; that people are more likely to form connections with people who have similar political preferences; and that consumption of political content tends to be more homogeneous than non-political content [7, 15, 45]. Lab studies have also demonstrated that when given the choice between media outlets, people tend to engage in "selective exposure" and choose to consume content from outlets that align with their political views [34, 40, 64].

However, observational studies of social media show little evidence for the "echo chamber" hypothesis [27]. Most consumers of news online have relatively moderate political news diets, or otherwise, do not pay attention to news at all [23, 24, 28, 55]. Use of social media has also been found to be associated with increased exposure to counter-attitudinal information, and social recommendations of content online have been shown to blunt the influence of partisan selective exposure [5, 23, 43].

**A**  **Amazon News**  @amazonnews · Mar 24  
Replying to @repmarkpocan  
1/2 You don't really believe the peeing in bottles thing, do you? If that were true, nobody would work for us. The truth is that we have over a million incredible employees around the world who are proud of what they do, and have great wages and health care from day one.

14.3K 19.4K 4.7K 

**B**  **Currently rated helpful** ...  
Informative · Cites high-quality sources

**Potentially misleading** Mar 25  
Amazon has a documented history of labor violations, including pushing employees to work so much they do not have time to use the restroom.  
<https://www.theguardian.com/technology/2020/feb/05/amazon-workers-protest-unsafe-grueling-conditions-warehouse>

<https://www.newsweek.com/amazon-drivers-warehouse-conditions-workers-complains-jeff-bezos-bernie-1118849>

<https://www.npr.org/2020/07/31/897836765/amazon-workers-respond-to-jeff-bezos-testimony-before-congress>

**C**  **Currently not rated helpful** ...  
Sources not included or unreliable · Misses key points

**Potentially misleading** Mar 25  
Amazon workers are treated awfully. Thank you.

**Figure 1:** An example of two Birdwatch notes, along with the focal tweet. (A) is the tweet that has been flagged by Birdwatch users. (B) is a Birdwatch note which labels the tweet in (A) as “Potentially Misleading”. The note shown in (B) has been labeled as “Currently rated helpful” by Twitter, based on the high aggregate helpfulness rating given to it by other Birdwatch users. (C) is a Birdwatch note that also labels the tweet in (A) as “Potentially Misleading”. However, the note in (C) has been labeled as “Currently not rated helpful” by Twitter, likely based on the low aggregate helpfulness rating given to it by other Birdwatch users.

While partisan selective exposure is rarer than expected online, partisan political *behavior* on social media has been robustly documented. Users are more likely to share and retweet content from co-partisans, especially on political topics [7, 12, 15, 16]. Partisans are also much more likely to share fact-checking messages that denigrate their political opponents and boost their political allies [60]. Highly active partisans are also more likely to engage in adversarial interactions with out-party politicians on Twitter [30, 31].

Thus, the majority of empirical work looks at either exposure to, or sharing of, content. Our work contributes to this literature by directly examining *judgments* of content generated by co-partisans versus counter-partisans. People do not necessarily believe much of what they are exposed to [49] or share [50]. Thus, it is an open question as to whether layperson judgments of content will mirror

exposure, where we see less empirical evidence of partisan differences than expected, or of sharing, where we see larger effects of partisanship – or show an entirely different pattern.

## 2.2 Motivated Reasoning

There has been a large amount of work in laboratory settings examining partisan judgment of information. The process by which people use biased cognitive processes in order to arrive at a particular directional outcome is called “motivated reasoning,” and many papers have claimed to observe politically motivated reasoning [36, 37, 41]. For example, an influential study showed that partisans judged confirmatory political claims as higher quality than disconfirmatory claims, and engaged in more counterargument against opposing claims while uncritically accepting supporting claims [67]. This work has also been applied to processing of political misinformation and corrections. Early research showed a “backfire effect,” in which exposure to a correction triggered a counter-argument that actually increased belief in the original misperception [47].

However, recent research has shown that these backfire effects are more likely the exception than the norm, and that corrections typically reduce belief in misinformation on average [66, 75]. Furthermore, studies have shown that even if partisans evaluate co-partisan versus counter-partisan content differently, they might not be exhibiting cognitive bias if partisans have different prior factual beliefs [68, 69, 69]. These different prior beliefs also need not be indicative of less accurate judgments; for example, research has shown that although people are more likely to believe politically concordant news, partisan alignment is not particularly predictive of the extent to which people *differentiate* between false and true news [53, 54].

Importantly, most of these studies use political content that has been hand-picked by experimenters, and thus, may or may not be representative of the content that people actually encounter online. Therefore, it is unclear to what extent partisanship will play a role in how users evaluate news on social media. Research has found that Twitter is less political than has been typically assumed, and that political content only constitutes a small percent of all tweets – just 13% according to a Pew Analysis [46, 74]. Thus, it is possible that fears of partisan motivated reasoning are overblown and that partisanship will not play a major role in the assessment of content online. By examining the role of partisanship in the judgments of Birdwatch participants, we can shed new light on this question.

## 2.3 Crowdsourced Fact-checking

It is not a priori obvious how these individual-level findings of partisanship translate to group-level assessments of content. Lab studies have shown that small groups of laypeople can generate reasonable levels of agreement with the ratings of experts, including on political content, and that aggregating judgments even among politically homogeneous crowds can lead to more accurate and less polarized judgments [2, 9, 10, 20, 25, 52, 58].

However, this research was done in settings where laypeople were assigned which pieces of content to rate. In contrast, a study examining editing of Wikipedia articles found that politically homogeneous groups of editors produced worse-quality and less accurate articles than politically heterogeneous groups [59]. On the subject

of crowdsourced fact-checking specifically, work in computer science has shown that algorithms where users choose which content to flag could be used to efficiently limit the spread of misinformation, but these algorithms have not been applied in practice or with real user flags [38, 71].

Our work sheds important new light on crowdsourced fact-checking in the wild. Characterizing the behavior of Birdwatch participant crowds who are allowed to choose what to rate illuminates whether partisanship plays a large role in how users 1) rate the accuracy of others’ content and 2) judge the helpfulness of fact-checks.

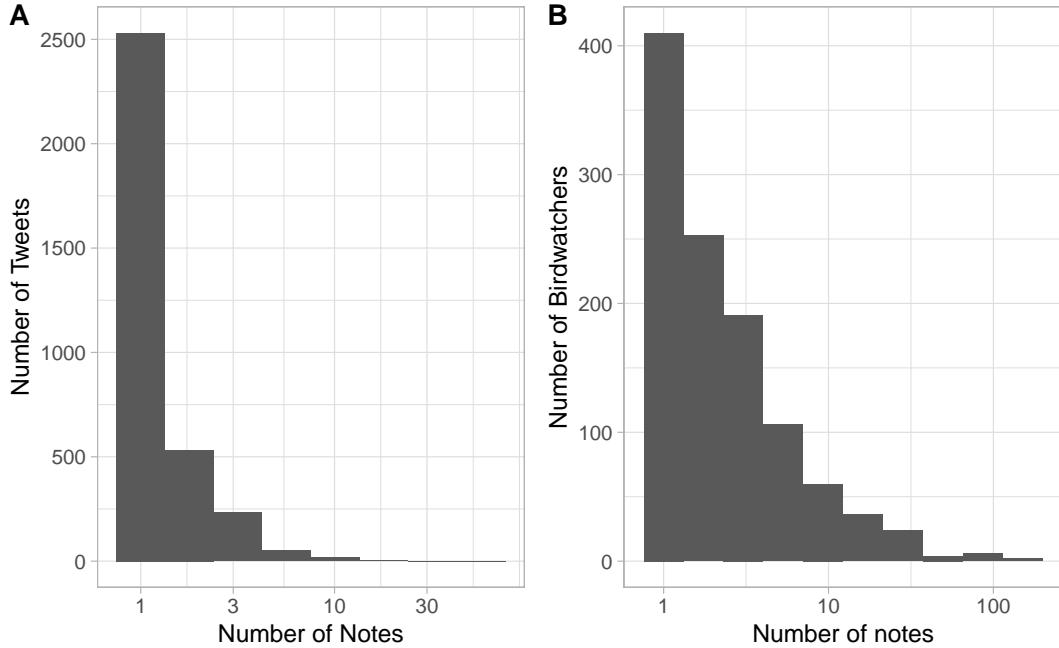
## 3 METHODS

### 3.1 Twitter Datasets

Our analysis of Birdwatch, Twitter’s crowdsourced fact-checking product, used three separate datasets. The first two datasets – the Notes dataset and the Ratings dataset – were provided to us by Twitter, covering all Birdwatch notes and ratings created from the program’s inception on 1/28/21 through 6/29/21. These datasets are very similar to the publicly available datasets found at <https://twitter.com/i/birdwatch/download-data>, except these datasets contained additional information made available for our internal research purposes that allowed us to link the activity of users participating in Birdwatch (the “Birdwatchers”) to their Twitter IDs. The third dataset – the Tweets dataset – was collected by us using the Twitter API.

**3.1.1 Notes Dataset.** The Notes dataset contains the set of 4910 fact-check notes submitted by 1092 unique Birdwatchers. The entry for each note includes the binary classification of the tweet by the Birdwatcher (either “Not Misleading” or “Potentially Misinformed or Misleading”), the Birdwatcher’s Twitter user ID, the tweet ID of the tweet being fact-checked, and a free-text summary written by the Birdwatcher explaining the rationale for their labeling of the tweet. The Notes dataset was highly imbalanced in terms of classifications: 89.6% of the notes in the sample had a classification of “Potentially Misinformed or Misleading.” Thus, notes functioned largely to flag tweets as potentially misleading. Tweets in this dataset received an average of 1.46 notes (median: 1), and Birdwatchers in this dataset rated an average of 4.5 tweets (median: 2). Full histograms of (a) the number of notes received by each tweet and (b) the number of notes submitted by each Birdwatcher who submitted at least one note can be found in Figure 2.

**3.1.2 Ratings Dataset.** The Ratings dataset contained the 28276 ratings of the helpfulness of the Birdwatch notes, submitted by a set of 2359 Birdwatchers. The entry for each rating includes the note ID, the binary helpfulness rating given to the note (either “Helpful” or “Not Helpful”), and the user ID of the Birdwatcher who gave the rating. The distribution of ratings was much more balanced than the distribution of classifications in the Notes dataset: Of the ratings in our dataset, 65.6% were helpful. Each note received 5.9 ratings from Birdwatchers on average (median: 3), and Birdwatchers rated 12.2 notes on average (median: 4). Full histograms of (a) the number of ratings received by each note and (b) the number of ratings submitted by each Birdwatcher who submitted at least one rating can be found in Figure 3.



**Figure 2:** (A) A histogram of the the number of Birdwatch notes received by each tweet. The histogram is long-tailed, and most tweets receive one note, although some receive up to 30. (B) A histogram of the number of notes submitted by each Birdwatcher who wrote at least one note. The histogram is also long-tailed, with most Birdwatchers submitting less than 5 notes, but some submitting more than 100. Note that for both histograms the X-axis is on a logarithmic scale.

**3.1.3 Tweets Dataset.** Finally, we used the Twitter API to pull the full text of tweets about which notes had been written by Birdwatchers, as well as the Twitter user ID of the tweet's author. Most tweets were accessible via the API; however, some were missing due to the tweet author's account being suspended or made private, or because the tweet was deleted. At the time of writing, 89.1% of the 3367 total tweets were available for download. Notes for which the original tweets were missing were kept in the dataset, and the relevant tweet-related features were imputed from the means of the data from the existing tweets.

## 3.2 Features

The review helpfulness literature broadly groups features into two different categories: content features, which are features derived directly from the text of the reviews, and context features, which are features like reviewer characteristics that are not derived from the review itself, but nonetheless can be used to predict helpfulness [19, 61]. Drawing on this literature, in our analysis, we determined quantities related to (1) the content of the note summaries and the tweets and (2) contextual features related to the individual characteristics of the tweeters and the Birdwatchers. We then used these features for our main analyses.

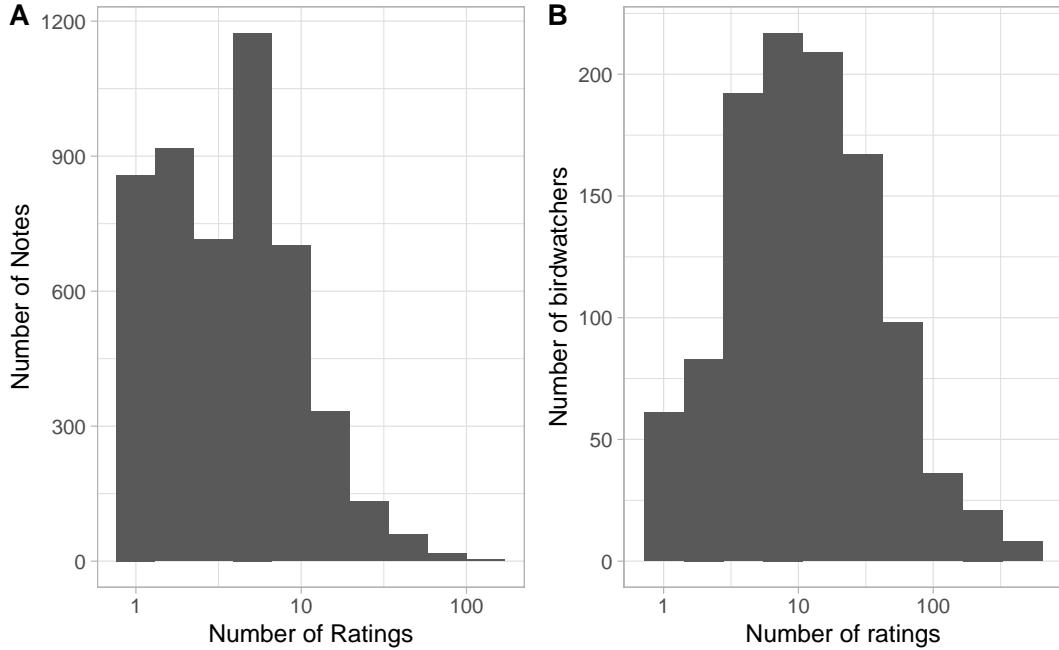
**3.2.1 Content Features.** We extracted several features related to the content of the note summaries and the tweets, which are summarized in Table 1. Length, sentiment, and readability have been shown to improve models of review helpfulness in past studies (for a review, see [19]). Additionally, we included the number of URLs as

an additional feature, since Twitter suggests that Birdwatch users cite sources in their fact-check notes. We generate the same features for both the note summaries and the tweets.

**3.2.2 Context Features.** Additionally, we extract several context-related features focusing on user characteristics, which are summarized in Table 2. We generate all of these features for the (1) tweeters, (2) Birdwatch note writers, and (3) Birdwatch raters, respectively.

We determine users' follower count and statuses count (number of posts the user has made) from the the Twitter API. We use the M3Model package [73], a deep-learning model that uses the user's profile image and textual features of their account, to infer users' gender and age. Most importantly for our key question of interest, we use the approach of Barberá et al. [7] and Barberá [6], which use the accounts a given user follows to predict their partisanship (Democrat versus Republican), where a score of "0" is represents the partisanship of the median Twitter user. We use this score to assign predicted party identities to users, with scores greater than 0.5 classified as "Republican" and scores less than or equal to 0.5 classified as "Democrat".

Due to some accounts being deleted, suspended, or made private, we are able to retrieve the full set of user characteristics from 87.7% of tweeters, 92.9% of Birdwatch note writers, and 92.9% of Birdwatch raters. We use mean imputation to fill in any missing data. Descriptive statistics can be found in Table 3.



**Figure 3:** (A) A histogram of the the number of ratings received by each Birdwatch note. Most notes receive 10 ratings or less, with some notes receiving up to about 100.(B) A histogram of the number of ratings submitted by each Birdwatch user who submitted at least one rating. Note that for both histograms the X-axis is on a logarithmic scale.

**Table 1:** Content related features derived from the tweets and Birdwatch notes, respectively.

Feature name	Description
Length	Length (i.e. character count) of the note summary or tweet
Sentiment	Vader Sentiment score from Hutto and Gilbert [33] for the summary or tweet. [-1,1] scale, where positive values connote positive sentiment.
FK Score	Flesch-Kincaid Reading ease score of summary or tweet. [1,100] scale, with higher values connoting easier reading.
URL Count	Number of URLs in the note summary or tweet.

**Table 2:** Context related features derived for the tweet authors, Birdwatch note writers, and Birdwatch note raters, respectively.

Feature name	Description
Follower Count	Number of followers the user has
Statuses Count	The total number of tweets and retweets the user has posted
Age	Predicted age category using M3Model described in Wang et al. [73]. Categories are <=18, 19-29, 30-39, >=40.
Gender	Predicted gender using M3Model described in Wang et al. [73]. Coded as "female" vs. "not female."
Partisanship Score	Partisanship inferred using the accounts the user follows, using the method from Barberá et al. [7]. [-2.5,2.5] scale, with more positive values indicative of greater affinity for the Republican party.

**Table 3: Descriptive statistics for Birdwatch raters, Birdwatch note writers, and tweeters. Gender, age, and party are predicted values derived from machine learning models. Statuses count and follower count are retrieved using the Twitter API.**

	Rater	Note Writer	Tweeter
<b>Predicted Gender</b>			
% Female	19%	20%	42%
% Not Female	81%	80%	58%
<b>Predicted Age</b>			
% <= 18	22%	17%	4%
% 19-29	30%	30%	15%
% 30-39	24%	26%	24%
% >= 40	24%	28%	57%
<b>Predicted Party</b>			
% Democrat	62%	52%	45%
% Republican	38%	48%	55%
<b>Statuses Count</b>			
Mean	17864	25772	57617
Median	6052	8886	16333
SD	38141	53673	111405
<b>Follower Count</b>			
Mean	3584	11069	3959671
Median	386	517	608718
SD	35192	80472	9368102

### 3.3 Models

Our main analyses consist of comparing the performance of various sets of features on two different classification tasks – (1) predicting whether each note classified its respective tweet as potentially misleading and (2) predicting whether each rating rated its respective note as helpful.

We use random forest (RF) models, which have consistently been shown to give good performance on supervised learning tasks that use social media data [13, 17]. In particular, RF models excel at detecting complex interactions between features, which we expect might be relevant when looking at the potential interactions between the partisanship of the tweeter, note writer, and rater. These analyses allow us to measure the maximum predictive ability of a model both in absolute terms and in comparison to the same type model trained on different sets of features, giving us insight into which types of features are most important for our classification tasks.

We performed hyperparameter tuning of the RF model and repeated 5-fold cross-validation (100 times for a total of 500 scores) separately for each of the feature sets. For our evaluation metric, we use the Area-Under-the-Receiver-Operating-Curve (AUC), due to the unbalanced nature of the data and the fact that we value correct prediction of both classes. We report the average AUC and range of the 500 iterations of the cross-validation procedure. Using alternate evaluation metrics like accuracy and F1-score produced substantively similar findings, see Section A.1.

## 4 RESULTS

### 4.1 Predicting Misleadingness Classification

**4.1.1 Random Forest Models.** Using the Notes dataset, with tweet and tweeter characteristics merged in from the Tweets dataset, we predict the note’s classification, where “0” corresponds to “Not misleading” and “1” corresponds to “Potentially misinformed or misleading”.

We compare the performance of the RF model predicting note classification using 4 different features sets. First, as a baseline, we train a model using a (1) content-level feature set that contains the features related the tweet’s textual content. Then, we compare the results of this content-only feature set to ones that consist of (2) the partisanship scores of the tweeter and note writer, (3) only the partisanship scores as well as all other (demographic and engagement) context features of the tweeter and note writer, and (4) all features. A description of the features included in each model can be found in Table 4.

A comparison on the performance of the RF model predicting whether each note classified its tweet as misleading using the various feature sets can be found in Figure 4A. The estimate of the AUC when predicting classification from our baseline content-level feature set is 0.56 (Range = 0.48 to 0.65). This indicates that on their own, the tweet level textual content features we considered do a relatively poor job of predicting which notes classify their tweets as misleading, since the baseline AUC for a model that guesses randomly is 0.5. In contrast, the estimate of the AUC when predicting classifications from just the partisanship scores of the tweeter and

**Table 4: Feature sets used to train our model classifying the misleadingness of tweets.**

Feature Set	Included Features
Content	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count
Partisanship	Tweeter Partisanship Score, Note writer Partisanship Score
Context	Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship Score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender
All	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count, Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship Score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender

note writer is substantially higher, at 0.84 (Range = 0.77 to 0.89). Next, adding additional context features (demographic and engagement features of tweeters and note writers) slightly increased the AUC to 0.87 (Range = 0.80 to 0.92). This suggests that most of the predictive ability of our context features model comes from tweeter and note writer partisanship scores. Finally, the model using all features has an AUC of 0.85 (Range = 0.79 to 0.91), such that adding content features provided no meaningful benefit beyond context features.

In order to further examine the relative importance of features in our all feature model, we also computed feature importances from one random draw of our cross-validation for the model using all features, which are summarized in Figure 4B. In line with our findings from the partisanship features model, we find the greatest feature importance scores for note writer partisanship (0.16) and tweet writer partisanship (0.13). The next most important features were tweeter follower count (0.10) and note writer follower count (0.10). Overall, the results suggest that the partisanship feature set is highly predictive of truth classifications, more so than our baseline tweet content feature set and comparable to a model containing all content and context features.

**4.1.2 Logistic Regression Model.** One benefit of RF models is that their structure naturally allows them to capture all relevant interactions between features, such as the interaction between the partisanship of the note writer and the tweet writer (i.e. political concordance), allowing them to outperform simple models like logistic regression on classification tasks. However, one drawback of the RF models is that, unlike linear models, it is impossible to identify the direction of the relationship between a feature and the outcome, or to understand which interactions between features are important.

To shed light on these questions, we also conducted a logistic regression model (unregularized) predicting "Potentially misinformed or misleading" classification with standard errors clustered by tweet, tweeter, and note writer, in order to gain insight into the directionality of the important features in our RF models. Our logistic regression model included all features, as well as the interaction between tweeter partisanship score and note writer partisanship score (both z-scored). We include this particular interaction, and not the others, because shared partisanship has been shown to be

a relevant predictor of a variety of other social media behaviors (e.g. sharing, following) [7, 45], and exploring whether the same relationship exists in the Birdwatch dataset is a major focus of our paper. Notably, we find a negative interaction between tweeter partisanship score and note writer partisanship score ( $b=-1.25$ ,  $SE=0.14$ ,  $z=-9.20$ ,  $p<.001$ ), such that *shared* partisanship is associated with not giving 'misleading' classifications. Tweeter and note writer follower count were also both negatively associated with 'misleading' classifications ( $ps < .026$ ); for full regression table, see Section B.

## 4.2 Helpfulness Classification Results

**4.2.1 Random Forest Models.** Next, we performed similar analyses predicting whether each rating rated its note as helpful. Using the Ratings dataset, with tweet and tweeter characteristics merged in from the Tweets dataset and note and note writer characteristics merged in from the Notes dataset, we predict the helpfulness rating of each rating, where "1" corresponds to "Helpful" and "0" corresponds to "Not Helpful".

Similarly to the truth classification task, we compare the performance of the RF model predicting rating-level helpfulness classification using 4 different features sets. However, the features for this model also include note-level content characteristics and rater-level context characteristics. Thus, we have the following feature sets (1) a content-based features based on textual features of the tweet and note, (2) the partisanship scores of the tweeter, note writer, and rater (3) the partisanship scores as well as other demographic and engagement features of the tweeter, note writer, and rater, and (4) all features. A description of the features included in each model can be found in Table 5.

The findings are summarized in Figure 5A. Our baseline content-level model has an AUC estimate of 0.76 (Range = 0.74 to 0.77). This AUC is substantially higher than the corresponding content-only model predicting whether notes classified their tweets as misleading - suggesting that the content features we examine are comparatively more predictive for helpfulness ratings than misleadingness classifications. However, once again our partisanship scores model has a substantially greater AUC estimate of 0.89 (Range = 0.88 to 0.90). And once again, our context features model has an only slightly larger AUC estimate of 0.91 (Range = 0.90 to 0.92). As in our note misleadingness prediction models, these results predicting ratings

**Table 5: Feature sets used to train our model classifying the helpfulness of notes.**

Feature Set	Included Features
Content	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count, Note Length, Note FK Score, Note Sentiment, Note URL Count
Partisanship	Tweeter Partisanship Score, Note writer Partisanship Score, Rater Partisanship Score
Context	Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender, Rater Partisanship Score, Rater Follower Count, Rater Statuses Count, Rater Age, Rater Gender
All	Tweet Length, Tweet FK Score, Tweet Sentiment, Tweet URL Count, Note Length, Note FK Score, Note Sentiment, Note URL Count, Tweeter Partisanship Score, Tweeter Follower Count, Tweeter Statuses Count, Tweeter Age, Tweeter Gender, Note writer Partisanship score, Note writer Follower Count, Note writer Statuses Count, Note writer Age, Note writer Gender, Rater Partisanship Score, Rater Follower Count, Rater Statuses Count, Rater Age, Rater Gender

show that most of the predictive power of the context features model comes from the partisanship score features. Finally, the all features model has an AUC of 0.92 (Range = 0.91 to 0.93). Thus, although the content features were somewhat predictive on their own, adding them to the context features does not meaningfully improve prediction.

Next, we again examined feature importance from one random draw of our all feature model. As can be seen in Figure 5B, the greatest feature importance score is partisanship score of the rater (0.21), followed by partisanship score of the note writer (0.10). Importance scores are also high for number of rater statuses (0.08) and number of rater followers (0.07).

Our helpfulness rating classification model results largely corroborate our main findings from the misleadingness classification models - namely that context features, and specifically partisanship, are highly predictive of both misleadingness and helpfulness ratings.

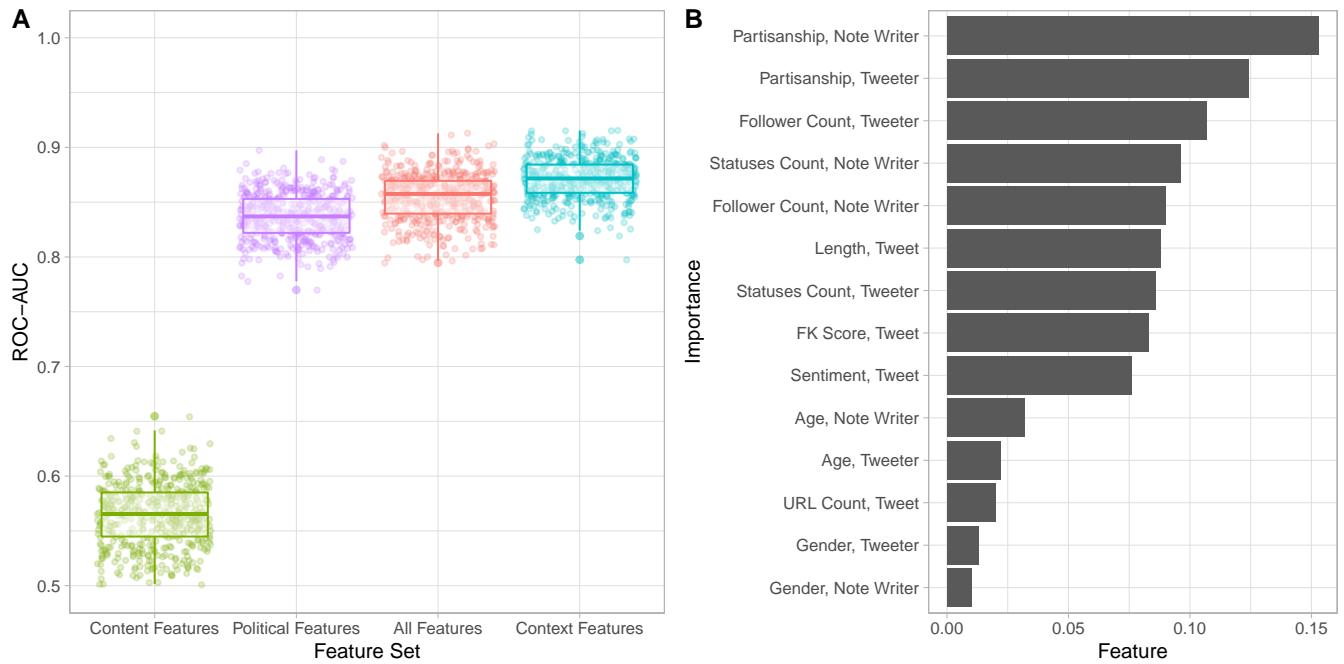
**4.2.2 Logistic Regression Model.** We again conducted a follow-up logistic regression model to examine the directionality of the relationships with key features from our RF models. Our logistic regression model included all features from our helpfulness classification all feature model, as well as all interactions between tweeter, note writer, and rater partisanship scores (all z-scored), and clustered standard errors by note, note writer, and rater, in order to predict helpfulness. We find a positive interaction between note writer and rater partisanship score ( $b=1.27$ ,  $SE=0.07$ ,  $z=17.02$ ,  $p<.001$ ), such that shared partisanship between note writer and rater is associated with notes being rated as helpful. We also observe a (somewhat smaller) negative interaction between tweeter partisanship score and rater partisanship score ( $b=-0.52$ ,  $SE=0.06$ ,  $z=-8.85$ ,  $p<.001$ ), such that shared partisanship between tweeter and rater predicts an unhelpful rating; for full regression table, see Section B. Given that most note classifications are 'misleading', this pattern of results suggests that raters tend to evaluate notes that agree with their partisanship as helpful, and notes that disagree with their partisanship as unhelpful.

### 4.3 Shared partisanship predicts classifications and ratings

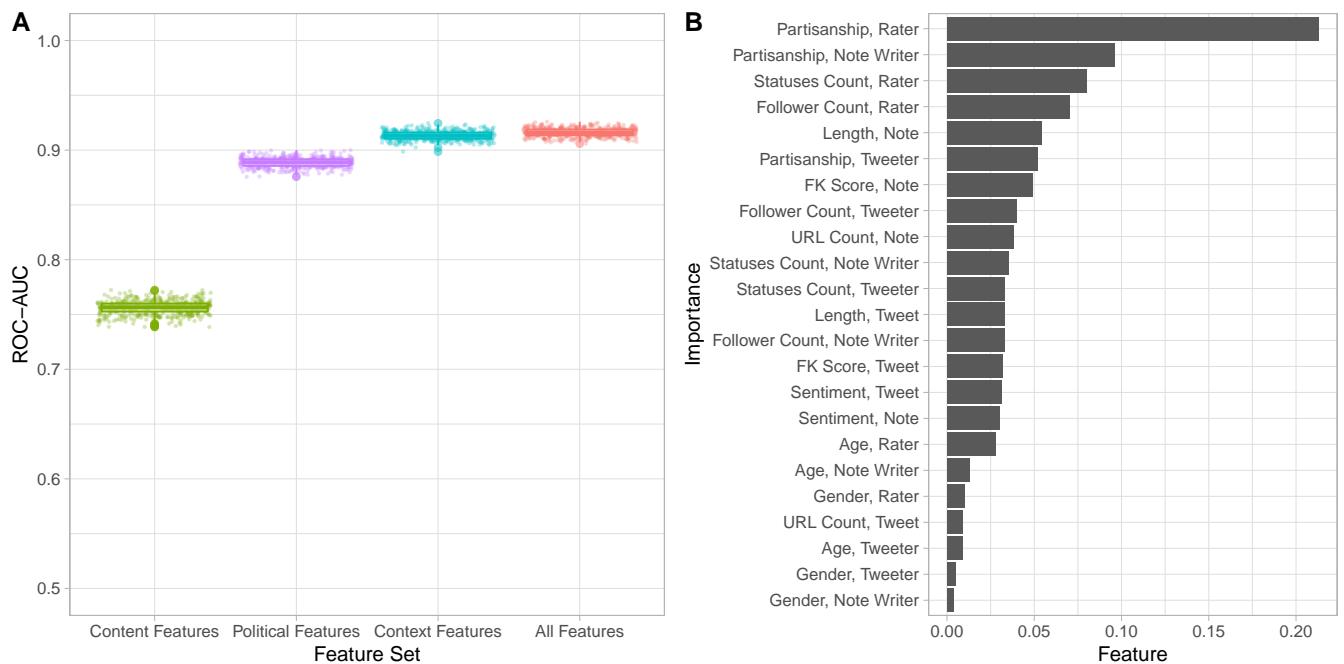
Our results above suggest that shared partisanship is an important feature in both of our models. In particular, the interaction between the partisanship of the tweeter and note writer when predicting the misleadingness classifications, and between the note writer and rater when predicting the helpfulness of ratings are both highly significant and large in magnitude. In this section, we explore those two relationships in further detail.

The relationship between misleadingness classification and the predicted partisanship scores of the note writer and tweeter are shown in Figure 6. For clarity, we also summarize the results using the (binary) predicted party of the tweeter and note writer, where values of the political score greater than 0.5 are coded as "Republican," and less than 0.5 are coded as "Democrat", in Table 6 [7]. Two findings are important to note. First, Birdwatchers are much more likely to write notes about tweets written by counter-partisans than co-partisans. Predicted Democrats are 3X more likely, and predicted Republicans are 1.5X more likely, to submit a note about a tweet by a counter-partisan than by a co-partisan. Second, while the vast majority of note classifications are misleading, Birdwatchers are more likely to classify a counter-partisan's tweet as misleading than a co-partisan. Predicted Republicans rated 97.2% of tweets by predicted Democrats as misleading (compared to 71.3% by predicted Democrats), and predicted Democrats rated 95.5% of tweets by predicted Republicans as misleading (compared to 82.4% by predicted Democrats). Overall, then, Birdwatchers are much more likely to flag counter-partisans' tweets as potentially misleading.

We see similar evidence of strong co-partisan preference when exploring the relationship between helpfulness ratings and the partisanship scores of the note writer and rater; see Figure 7 and Table 7. Unlike with note-writing, Birdwatch users – particularly predicted Democrats – rate more notes from *co-partisans* than counter-partisans. Predicted Democrats are 3X more likely, and predicted Republicans are 1.1X more likely, to rate a note from a co-partisan than from a counter-partisan. Second, Birdwatch users are much more likely to classify a co-partisan's note as helpful than a



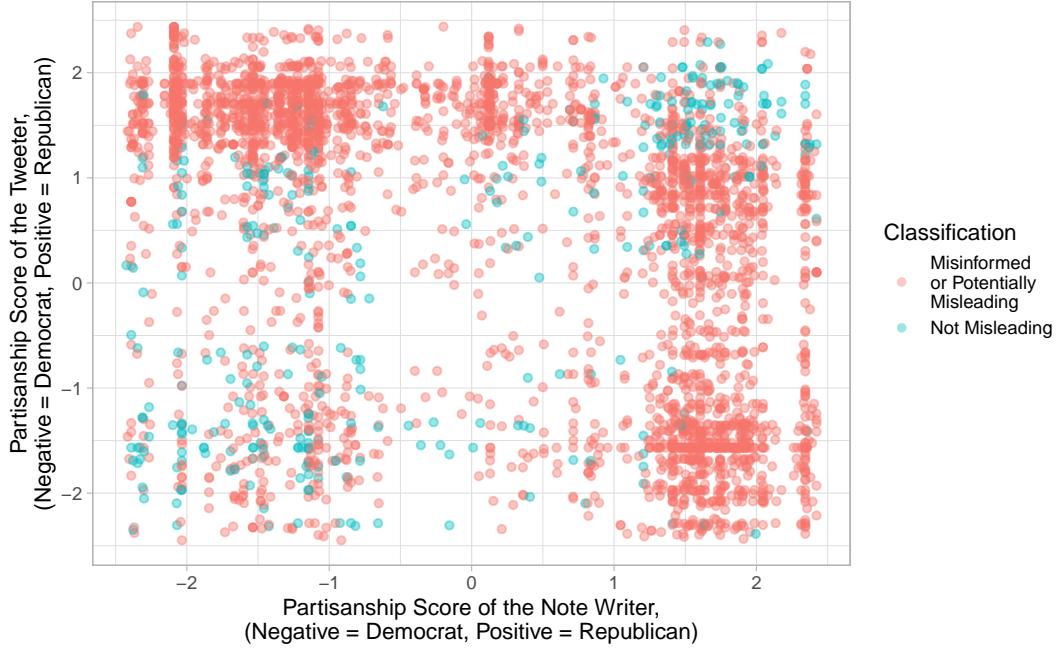
**Figure 4: (A) Comparing the performance of RF models predicting the misleadingness classification of tweets with different feature sets (B) Comparing the relative importance of features for an RF model trained with all features**



**Figure 5: (A) Comparing the performance of RF models predicting the helpfulness ratings of Birdwatch notes with different feature sets (B) Comparing the relative importance of features for an RF model trained with all features**

counter-partisan's. Predicted Republicans rated 83.1% of notes written by other predicted Republicans as helpful (compared to 43.3%

of notes written by predicted Democrats), and predicted Democrats



**Figure 6: Misleadingness classifications of tweets, by the partisanship score of the note submitter and the partisanship score of the tweeter. Each point represents one tweet.**

**Table 6: (1) Note count and (2) Percent of Notes Rated as “Misleading” for different combinations of the Tweeter and Note Writers’ Predicted Parties (Republican or Democrat)**

	Tweeter Democrat		Tweeter Republican	
	Count	Percent Misleading	Count	Percent Misleading
Note Writer Democrat	489	71.3%	1515	95.5%
Note Writer Republican	1003	97.2%	679	82.4%

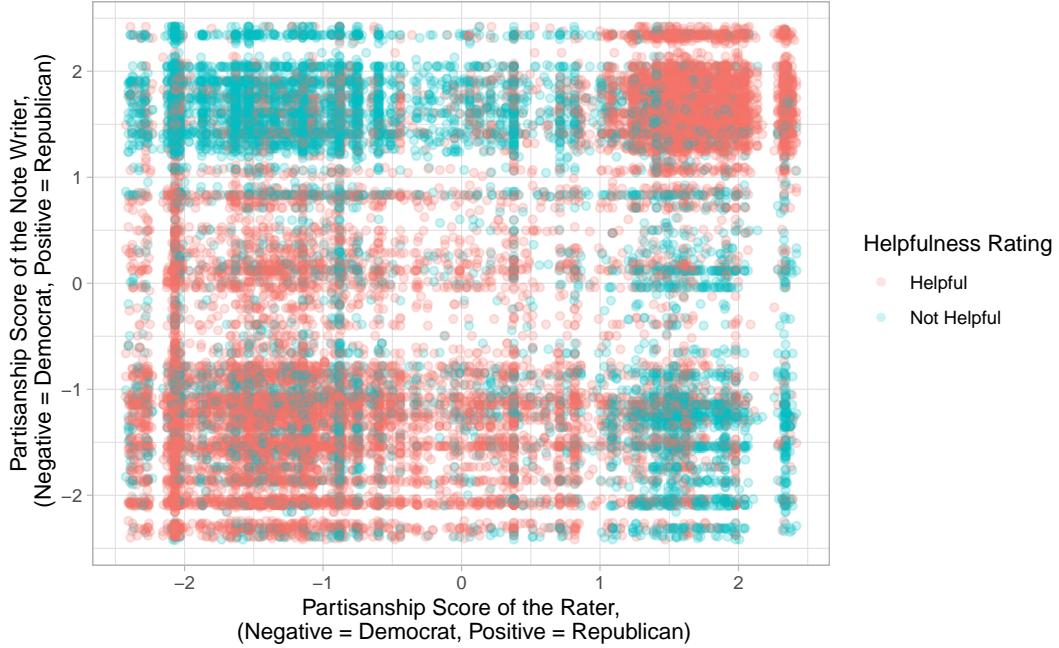
rated 87.1% of notes written by predicted Democrats as helpful (compared to 25.9% of notes written by predicted Republicans).

This preference for concordant notes has implications for the overall average helpfulness ratings of the notes. In Figure 8, we show the relationship between the percent of ratings that are from co-partisans and the overall average helpfulness rating of the note, for notes with at least 5 ratings. There is a strong, positive relationship between the percent of co-partisan ratings and the overall helpfulness rating of the note. For a weighted least squares regression of the average helpfulness rating on the percent of co-partisan ratings, where the weights are the number of ratings for that note, the coefficient on percent of co-partisan ratings is .71 ( $p < .001$ ). This means that for every additional 1% percent increase in ratings by co-partisans, the helpfulness rating rises 0.71%. The model has an  $R^2$  of 0.42, meaning that 42% of the variance in helpfulness ratings is explained by the percent of co-partisan raters.

## 5 DISCUSSION

Here we have shown that shared partisanship is an important predictor of how Twitter users in the Birdwatch program evaluate others’ posts, with tweets from counter-partisans judged as more misleading than tweets from co-partisans, and notes (e.g. fact-checks) from counter-partisans judged as less helpful than notes from co-partisans. We add to the literature on partisan selective exposure and partisan selective sharing by demonstrating a related phenomenon: partisan selective *evaluation*. These findings are notable and perhaps surprising, since much of the content on Twitter is not political in nature, and political content is a fairly small subset of most viral forms of misinformation on social media [46, 62]. It was not a priori obvious, therefore, that partisanship should be such a predictive factor when judging the accuracy of tweets or the helpfulness of fact-check notes – especially when compared to other theoretically relevant features, like the number of sources cited in the note.

Given these findings, it is possible that partisanship is motivating users to volunteer for, and contribute to, Birdwatch in the first place. Other research on crowdsourcing for citizen science has found that



**Figure 7: Helpfulness ratings of notes, by the partisanship score of the rater and the note submitter. Each point represents one note.**

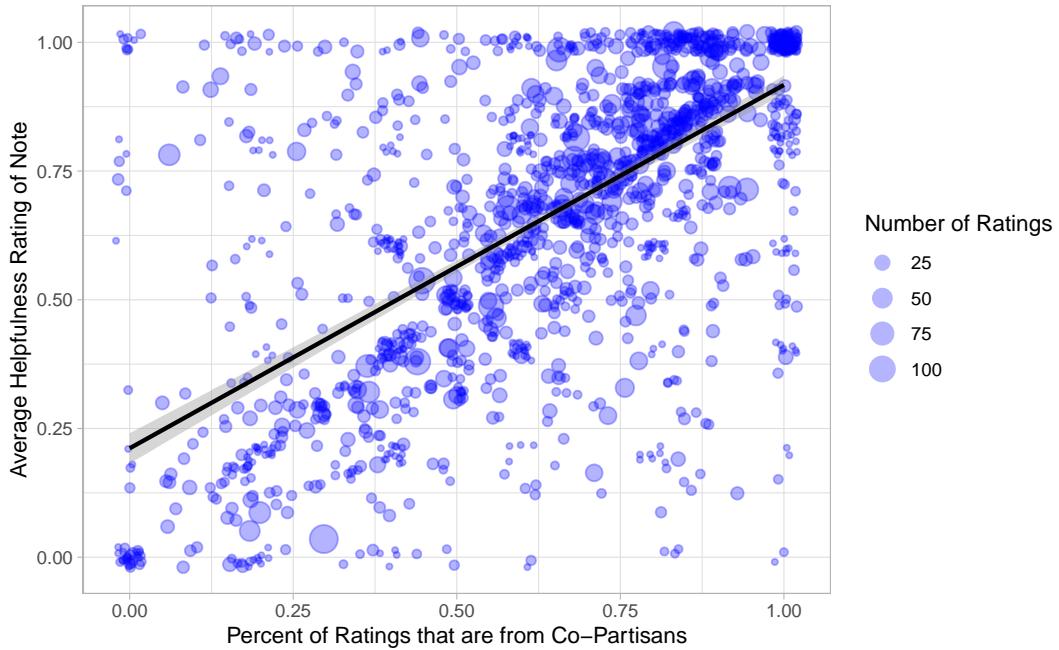
**Table 7: (1) Rating count and (2) Percent of ratings that are “Helpful” for different combinations of the Note Writer and Raters’ Predicted Parties (Republican or Democrat)**

	Note Writer Democrat		Note Writer Republican	
	Count	Percent Helpful	Count	Percent Helpful
Rater Democrat	9459	83.1%	3017	43.3%
Rater Republican	5609	25.9%	6379	87.1%

extrinsic motivations (e.g. status markers within the community) and intrinsic motivations (e.g. belief in the overall goal of the project and individual level interest) are important motivations for participating in these types of project [22, 42, 56]. Partisanship could play into both types of motivations in the case of Birdwatch. In terms of extrinsic motivations, it is possible that the helpfulness feedback system is signalling to Birdwatchers that partisan-aligned political content is most valued by other Birdwatchers, since the pattern of helpfulness votes suggests a partisan cheerleading effect. As for intrinsic motivations, research has shown that partisanship is a highly salient and important part of people’s identities [32]. It is therefore possible that people are participating in Birdwatch to either advance their partisan views, regardless of truth; or because they are genuinely concerned about misinformation generated by users from across the political aisle. Even though viral misinformation is to a large extent non-political in general, it is possible that Birdwatch users are particularly motivated to fact-check partisan information because partisanship is an important part of their identities. Past work has shown that evaluations online are costly to provide and thus scarcer than optimal, so partisan motivations might actually

be beneficial for soliciting notes and ratings in a non-paid platform like Birdwatch [3]. Indeed, past research on Wikipedia suggests that editors who are more politically extreme are more willing to spend time and effort advocating for their viewpoint on Wikipedia articles, and thus, some level of “bias” among editors might spur an optimal level of debate and activity on the platform [59]. A similar dynamic could be happening on Birdwatch.

Importantly, the preferential flagging of counter-partisan tweets we observe does not necessarily impair Birdwatch’s ability to identify misleading content. It is possible that partisans are successfully identifying misinformation from across the aisle (even if they are not scrutinizing content from their own co-partisans as closely), and/or that aggregating ratings from the entire community cancels out bias from both sides (as in [20]). Consistent with this possibility, a preliminary investigation found that among 57 tweets which a majority of Birdwatchers flagged as misleading, 86.0% were also rated as misleading by at least one of two professional fact-checkers (recruited from [2]). Future work should investigate these issues more thoroughly by assessing the veracity of the full set of fact-checked



**Figure 8: Predicting helpfulness ratings as a function of the percent of ratings that are from co-partisans for notes with greater than or equal to 5 ratings. Each point represents one note, where points are sized by the number of ratings received by each note.**

tweets relative to some ground truth (e.g. by having professional fact-checkers evaluate all tweets).

Beyond the specific use case of crowdsourced fact-checking on Twitter, our study contributes to research on partisanship and misinformation more generally. Our observation that Birdwatch participants were much more likely to choose to fact-check counter-partisan tweets provides ecologically-valid support for previous findings from survey experiments suggesting that people subject out-partisan content to more scrutiny than in-partisan content. For example, Taber and Lodge [67] found that partisans scrutinized counter-attitudinal content far more closely than pro-attitudinal content, which they did not critically examine. In their work, exposure to opposing arguments led to ideological polarization rather than moderation, and although we do not measure polarization as an outcome, it is possible that a similar phenomenon could happen in this instance.

Interestingly, the pattern of partisan selection evaluation that we observe on Birdwatch cannot be explained by partisan selective exposure. We inferred user's partisanship based on the accounts they followed, and thus, by construction, users feeds were more likely to contain co-partisan content than counter-partisan content. Nonetheless, users were more likely to post fact-checks of counter-partisan tweets, and, conditional on performing a fact-check, more likely to rate counter-partisan tweets as misleading. Furthermore, such partisan selection is likely driven primarily by disagreement with and motivation to fact-check (potentially misleading) counter-partisan content itself, rather than motivation to fact-check based on partisan account cues. This is because the partisanship of profiles

is likely opaque to users, with some notable exceptions such as accounts of politicians and other political elites. Thus, it is likely that counter-partisan cues in tweeted content itself is motivating partisan fact-checking.

While Birdwatch notes were only viewable by the public on a separate website at the time these data were generated, Birdwatch pilot users could see helpful notes attached to the tweets in their feeds. With this in mind, it is important to consider that the users who followed accounts with a similar political lean to a given tweet's author – and who presumably are thus more likely to come across the tweet organically in their newsfeed – were more likely to be critical of (i.e. rate as unhelpful) notes that marked the tweet as misleading. Thus, the most likely potential consumers of the fact-check were least likely to consider the fact-check helpful. This negative assessment could have important implications for polarization, especially if the fact-checks in question bear more resemblance to partisan “dunking” than to corrections by fact-checkers [57, 72]. While research has shown that fact-checks – even partisan ones – generally decrease belief in misinformation [66, 75], other work has shown that both exposure to counter-partisan content and negative characterizations of counter-partisans can increase polarization [4, 35, 72]. Public corrections could also cause backlash from the original tweeter, as has been shown in a field experiment on Twitter where replying to a misinformation tweet with a link to a fact-check increased the partisan slant and toxicity of the original tweeter's subsequent retweets [44].

Furthermore, the partisan behavior we observe has important implications for the ability of the Birdwatch helpfulness rating system to identify helpful fact-checks. Both our paper and work by others [76] has identified substantial partisan herding in Birdwatch ratings, identifying a potentially substantial flaw in the helpfulness rating system. Perhaps due to these problems, Twitter has been implementing changes to the rating system. While the data analyzed here were being collected, Twitter labeled notes that had at least five ratings and an average helpfulness score of at least 0.84 as "Currently rated as helpful" and highlighted these notes more prominently on their site. Subsequently, in June 2021, Twitter changed their helpfulness labeling algorithm to weight notes by a Birdwatcher's reputation, which is derived in part based on the agreement with consensus rating of the notes they rated in the past [11]. However, if, as we see, Democrats are less likely to submit notes for counter-partisan content than Republicans, then Democrat raters could have a higher reputation due to their greater willingness to engage in partisan cheerleading – rather than higher overall quality. On the other hand, if Twitter just does a simple aggregation of helpfulness scores without reputation rating, they risk a situation where partisan herding could lead to actually helpful notes getting downvoted and unhelpful notes getting boosted due to brigading. It is possible that a different aggregation methodology for helpfulness that balances ratings from parties could prove beneficial, or that Birdwatch should dispense with the helpfulness ratings entirely and instead only focus on classifying tweets as misleading, and/or providing fact-checking notes.

There are important limitations to our study. We cannot identify from these data whether the pattern we observe is the result of politically motivated or otherwise biased reasoning. The observed pattern could also be explained by partisan difference in prior factual beliefs, leading (rational Bayesian) partisans to be more likely to fact-check out-party content simply because it is surprising, rather than because of a political vendetta or bias [70]. Furthermore, it is important to note that the Twitter users who opted in, and were subsequently selected, to participate in the Birdwatch pilot are surely unrepresentative of Twitter users in general, or of Americans more broadly. Men outnumber women 4:1, and the average tweet count Birdwatchers (>25,000) suggests that the users were quite active on Twitter. Additionally, they may be more politically engaged and extreme - and thus more responsive to shared partisanship - than the average Twitter user. Future work should examine how Birdwatchers compare to more representative populations, and evaluate what individual differences predict the relationship between shared partisanship and choosing to rate others' content. Future research should also examine the extent of partisan herding in Twitter replies more generally, rather than just on Birdwatch. Such analyses may shed light on how similar the partisan dynamics observed in a crowdsourced fact-checking setting are to partisan dynamics on Twitter overall. It will be informative to see whether partisan herding is exacerbated by a fact-checking directive, or if similar partisan communication patterns exist (perhaps to an even greater degree) on Twitter outside of Birdwatch.

Furthermore, we recognize that one potential drawback to this research is that, for privacy reasons, we cannot release IDs of the Twitter accounts participating in Birdwatch that were used in our analysis. For transparency, we have posted our code on

OSF: <https://osf.io/acx3j>. Twitter has been releasing anonymized datasets of the notes and ratings from Birdwatch on their site <http://twitter.com/i/birdwatch/download-data> and if de-identified datasets including the relevant co-variates from our analyses become available we will add them to OSF.

In sum, we have shown that shared partisanship is a strong predictor of whether a user rates a tweet as misleading or a fact-check as helpful in the context of Twitter's crowdsourced fact-checking platform Birdwatch. While we do not believe that our findings mean that social media platforms should abandon crowdsourcing as a tool for identifying misinformation, the patterns we observe clearly indicate that it is essential to consider partisan dynamics when designing crowdsourcing systems.

## REFERENCES

- [1] Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. The welfare effects of social media. *American Economic Review* 110, 3 (2020), 629–76.
- [2] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2020. Scaling up fact-checking using the wisdom of crowds. *Preprint at https://doi.org/10.31234/osf.io/9qdza* (2020).
- [3] Christopher Avery, Paul Resnick, and Richard Zeckhauser. 1999. The market for evaluations. *American economic review* 89, 3 (1999), 564–584.
- [4] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, M B Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U. S. A.* 115, 37 (Sept. 2018), 9216–9221.
- [5] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
- [6] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91.
- [7] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science* 26, 10 (2015), 1531–1542.
- [8] Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political analysis* 23, 1 (2015), 76–91. Publisher: Cambridge University Press.
- [9] Joshua Becker, Ethan Porter, and Damon Centola. 2019. The wisdom of partisan crowds. *Proc. Natl. Acad. Sci. U. S. A.* 116, 22 (May 2019), 10717–10722. Publisher: National Acad Sciences.
- [10] Md Momen Bhuiyan, Amy X Zhang, Connie Moon Sehat, and Tanushree Mitra. 2020. Investigating differences in crowdsourced news credibility assessment: Raters, tasks, and expert criteria. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.
- [11] Birdwatch. 2021. Today, we're updating how notes are elevated in Birdwatch! This change will give more weight to contributors whose notes and ratings are consistently found helpful by others. <https://twitter.com/birdwatch/status/1404519791394758657>
- [12] Antoine Boutet, Hyoungshick Kim, and Eiko Yoneki. 2013. What's in Twitter, I know what parties are popular and who you are supporting now! *Social network analysis and mining* 3, 4 (2013), 1379–1391.
- [13] Rich Caruana and Alexandru Niculescu-Mizil. 2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*. 161–168.
- [14] Keith Coleman. 2021. Introducing Birdwatch, a community-based approach to misinformation. [https://blog.twitter.com/en\\_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html](https://blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html)
- [15] Elanor Colleoni, Alessandro Rozza, and Adam Arvidsson. 2014. Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of communication* 64, 2 (2014), 317–332.
- [16] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.
- [17] Mitali Desai and Mayuri A. Mehta. 2016. Techniques for sentiment analysis of Twitter data: A comprehensive survey. In *2016 International Conference on Computing, Communication and Automation (ICCCA)*. 149–154. <https://doi.org/10.1109/ICCA.2016.7813707>
- [18] Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).

- [19] Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 698–708.
- [20] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.
- [21] Ziv Epstein, Nathaniel Sirlin, Antonio Alonso Arechar, Gordon Pennycook, and David Rand. 2021. Social Media Sharing Reduces Truth Discernment. (2021).
- [22] Alexandra Eveleigh, Charlene Jennett, Ann Blandford, Philip Brohan, and Anna L Cox. 2014. Designing for dabblers and deterring drop-outs in citizen science. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2985–2994.
- [23] Seth Flaxman, Sharad Goel, and Justin M Rao. 2013. Ideological segregation and the effects of social media on news consumption. *Available at SSRN* 2363701 (2013).
- [24] Matthew Gentzkow and Jesse M Shapiro. 2011. Ideological segregation online and offline. *The Quarterly Journal of Economics* 126, 4 (2011), 1799–1839.
- [25] William Godel, Zeve Sanderson, Kevin Aslett, Jonathon Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking. *Journal of Online Trust and Safety* 1, 1 (2021).
- [26] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (Jan. 2019), 374–378.
- [27] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. 2018. Avoiding the echo chamber about echo chambers. *Knight Foundation* 2 (2018).
- [28] Andrew M Guess. 2021. (Almost) Everything in Moderation: New Evidence on Americans' Online Media Diets. *American Journal of Political Science* (2021).
- [29] Marek Hlavac. 2018. stargazer: Well-Formatted Regression and Summary Statistics Tables. (2018). <https://CRAN.R-project.org/package=stargazer> R package version 5.2.2.
- [30] Yiqing Hua, Mor Naaman, and Thomas Ristenpart. 2020. Characterizing Twitter Users Who Engage in Adversarial Interactions against Political Candidates. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. <https://doi.org/10.1145/3313831.3376548>
- [31] Yiqing Hua, Thomas Ristenpart, and Mor Naaman. 2020. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 272–282.
- [32] Leonie Huddy, Lilliana Mason, and Lene Aaroe. 2015. Expressive partisanship: Campaign involvement, political emotion, and partisan identity. *American Political Science Review* 109, 1 (2015), 1–17.
- [33] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 8.
- [34] Shanto Iyengar, Kyu S Hahn, Jon A Krosnick, and John Walker. 2008. Selective exposure to campaign communication: The role of anticipated agreement and issue public membership. *The Journal of Politics* 70, 1 (2008), 186–200.
- [35] Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* 22 (2019), 129–146.
- [36] Dan M Kahan. 2012. Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgment and Decision making* 8 (2012), 407–24.
- [37] Dan M Kahan. 2016. The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it. *Emerging trends in the social and behavioral sciences* 29 (2016).
- [38] Jooyeon Kim, 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.
- [39] Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2020. The Distorting Prism of Social Media: How Self-Selection and Exposure to Incivility Fuel Online Comment Toxicity. *Journal of Communication* (2020).
- [40] Silvia Knoblock-Westerveld and Jingbo Meng. 2009. Looking the other way: Selective exposure to attitude-consistent and counterattitudinal political information. *Communication Research* 36, 3 (2009), 426–448.
- [41] Ziva Kunda. 1990. The case for motivated reasoning. *Psychological bulletin* 108, 3 (1990), 480.
- [42] Anne M Land-Zandstra, Jeroen LA Devilee, Frans Snik, Franka Buurmeijer, and Jos M van den Broek. 2016. Citizen science on a smartphone: Participants' motivations and learning. *Public Understanding of Science* 25, 1 (2016), 45–60.
- [43] Solomon Messing and Sean J Westwood. 2014. Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. *Communication research* 41, 8 (2014), 1042–1063.
- [44] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. <https://doi.org/10.1145/3411764.3445642>
- [45] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David G. Rand. 2021. Shared partisanship dramatically increases social tie formation in a Twitter field experiment. *Proceedings of the National Academy of Sciences* 118, 7 (2021). Publisher: National Acad Sciences.
- [46] Subhayan Mukerjee, Kokil Jaidka, and Yphtach Lelkes. 2020. The Political Landscape of the US Twitterverse. (2020).
- [47] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.
- [48] MATHIAS OSMUNDSEN, ALEXANDER BOR, PETER BJERREGAARD VAHLSTRUP, ANJA BECHMANN, and MICHAEL BANG PETERSEN. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review* (2021), 1–17.
- [49] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.
- [50] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
- [51] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.
- [52] Gordon Pennycook and David G Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (2019), 2521–2526.
- [53] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.
- [54] Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* (2021).
- [55] Markus Prior. 2013. Media and political polarization. *Annual Review of Political Science* 16 (2013), 101–127.
- [56] M Jordan Raddick, Georgia Bracey, Pamela L Gay, Chris J Lintott, Phil Murray, Kevin Schwamb, Alexander S Szalay, and Jan Vandenberg. 2009. Galaxy zoo: Exploring the motivations of citizen science volunteers. *arXiv preprint arXiv:0909.2925* (2009).
- [57] Steve Rathje, Jay J Van Bavel, and Sander van der Linden. 2021. Out-group animosity drives engagement on social media. *Proceedings of the National Academy of Sciences* 118, 26 (2021).
- [58] Paul Resnick, Aljohara Alfayez, Jane Im, and Eric Gilbert. 2021. Informed crowds can effectively identify misinformation. *arXiv preprint arXiv:2108.07898* (2021).
- [59] Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. 2019. The wisdom of polarized crowds. *Nat Hum Behav* 3, 4 (April 2019), 329–336. Publisher: nature.com.
- [60] Jieun Shin and Kjerstin Thorson. 2017. Partisan Selective Sharing: The Biased Diffusion of Fact-Checking Messages on Social Media: Sharing Fact-Checking Messages on Social Media. *Journal of Communication* 67, 2 (April 2017), 233–255. <https://doi.org/10.1111/jcom.12284>
- [61] Michael Siering, Jan Muntermann, and Balaji Rajagopalan. 2018. Explaining and predicting online review helpfulness: The role of content and reviewer-related signals. *Decision Support Systems* 108 (2018), 1–12.
- [62] Craig Silverman. 2016. Here are 50 of the biggest fake news hits on Facebook from 2016. *Buzzfeed News* (2016), 1–12.
- [63] Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School Misinformation Review* (2021).
- [64] Natalie Jomini Stroud. 2010. Polarization and partisan selective exposure. *Journal of communication* 60, 3 (2010), 556–576.
- [65] Cass Sunstein and Cass R Sunstein. 2018. # Republic. Princeton university press.
- [66] Briony Swire-Thompson, Joseph DeGutis, and David Lazer. 2020. Searching for the Backfire Effect: Measurement and Design Considerations. *Journal of applied research in memory and cognition* 9, 3 (Sept. 2020), 286–299. <https://doi.org/10.1016/j.jarmac.2020.06.006> Edition: 2020/09/02 Publisher: The Authors. Published by Elsevier Inc. on behalf of Society for Applied Research in Memory and Cognition.
- [67] Charles S Taber and Milton Lodge. 2006. Motivated skepticism in the evaluation of political beliefs. *American journal of political science* 50, 3 (2006), 755–769.
- [68] Ben M Tappin, Gordon Pennycook, and David G Rand. 2020. Bayesian or biased? Analytic thinking and political belief updating. *Cognition* 204 (2020), 104375.
- [69] Ben M Tappin, Gordon Pennycook, and David G Rand. 2020. Rethinking the link between cognitive sophistication and politically motivated reasoning. *Journal of Experimental Psychology: General* (2020).
- [70] Ben M Tappin, Gordon Pennycook, and David G Rand. 2020. Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. *Current Opinion in Behavioral Sciences* (2020).

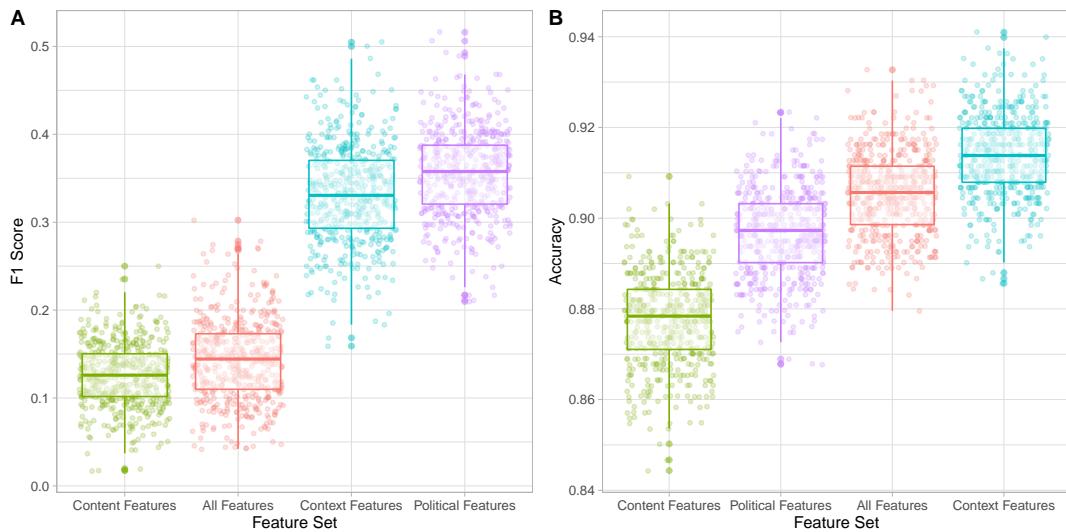
- Sciences* 34 (2020), 81–87.
- [71] Sebastian Tschiatschek, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the The Web Conference 2018*. 517–524.
  - [72] Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)* (2018).
  - [73] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flock, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*. ACM, 2056–2067.
  - [74] Stefan Wojcik and Adam Hughes. 2019. Sizing up Twitter users. *Pew Research Center* 24 (2019).
  - [75] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Polit. Behav.* 41, 1 (March 2019), 135–163. Publisher: Springer Science and Business Media LLC.
  - [76] Taha Yasseri and Filippo Menczer. 2021. Can the Wikipedia moderation model rescue the social marketplace of ideas? *arXiv preprint arXiv:2104.13754* (2021).
  - [77] Xudong Yu, Magdalena Wojcieszak, and Andreu Casas. 2021. Affective polarization on social media: In-party love among American politicians, greater engagement with out-party hate among ordinary users. (2021).

## A MODEL ROBUSTNESS

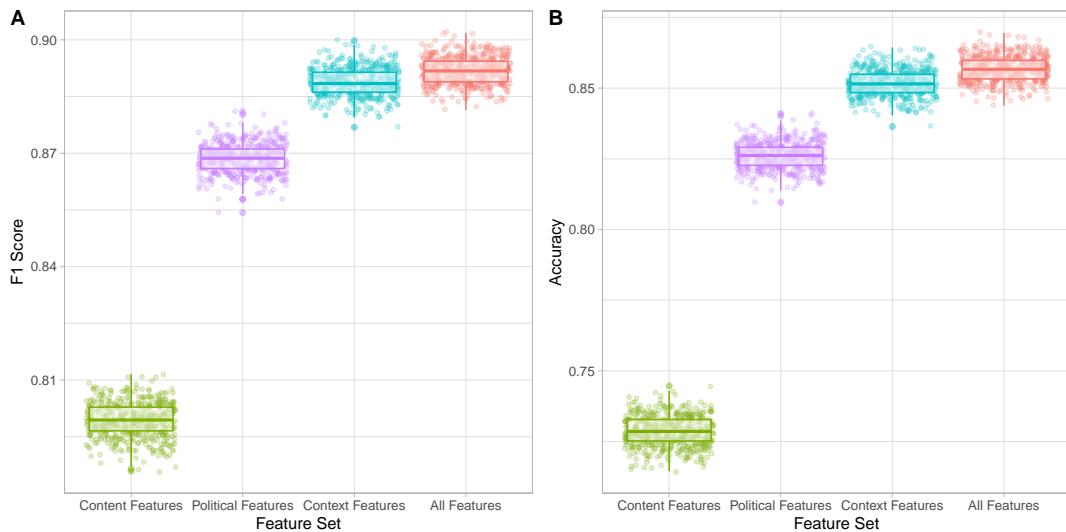
### A.1 Alternate Evaluation Metrics

A.1.1 *Misleadingness Classification.* We also evaluate an RF model predicting the misleadingness classification of tweets using F1 score and Accuracy as alternate metrics. Figure 9 shows these results. According to both metrics, "Content features" are the worst performing. Accuracy follows the same pattern as AUC, with "Political Features" as second worst, followed by "All Features", and then "Context Features" being the best. However, on F1 score, the "All Features" is second worst, followed by "Context", and then "Political" being the best.

A.1.2 *Helpfulness Classification.* We also evaluate an RF model predicting the helpfulness classification of notes using F1 score and Accuracy as alternate metrics. Figure 10 shows these results. Both metrics follow the same pattern as AUC, with "Content Features" being worse, then "Political Features", "Context Features", and "All Features", and being the best.



**Figure 9: A comparison of RF models predicting misleadingness classification of tweets, trained with different sets of feature using (A) F1 Score and (B) Accuracy as Evaluation Metrics**



**Figure 10: A comparison of RF models predicting helpfulness classification of notes, trained with different sets of feature using (A) F1 Score and (B) Accuracy as Evaluation Metrics**

## B LOGISTIC REGRESSION, FULL RESULTS

Results for a model predicting misleadingness can be found in Table 8; results for a model predicting helpfulness can be found in Table 9.

**Table 8: Full logistic regression output for predicting misleadingness classifications of tweets. Table rendered via [29]**

Constant	2.017*** (0.423)
Note writer Follower Count	-0.00000* (0.00000)
Note writer Statuses Count	0.00000 (0.00000)
Note writer Gender	0.319 (0.257)
Tweeter Follower Count	-0.00000* (0.000)
Tweeter Statuses Count	0.00000 (0.00000)
Tweeter Gender	0.151 (0.188)
Note writer Age	0.025 (0.089)
Tweeter Age	-0.059 (0.094)
Tweet Length	0.002 (0.001)
Tweet Sentiment	0.177 (0.144)
Tweet FK Score	-0.001 (0.002)
Tweet URL Count	-0.017 (0.131)
Note writer Partisanship Score	0.181 (0.126)
Tweeter Partisanship Score	-0.119 (0.120)
Note writer Partisanship Score X Tweeter Partisanship Score	-1.254*** (0.136)

*Note:*

\*p&lt;0.05; \*\*p&lt;0.01; \*\*\*p&lt;0.001

**Table 9: Full logistic regression output for predicting helpfulness classifications of notes.**

Constant	0.430 (0.294)
Note writer Follower Count	-0.000 (0.00000)
Note writer Statuses Count	0.00000 (0.00000)
Note writer Gender	-0.083 (0.102)
Tweeter Follower Count	-0.000** (0.000)
Tweeter Statuses Count	0.00000 (0.00000)
Tweeter Gender	0.016 (0.075)
Rater Follower Count	0.00000 (0.00000)
Rater Statuses Count	0.00000 (0.00000)
Rater Gender	-0.118 (0.164)
Note writer Age	-0.110* (0.047)
Tweeter Age	-0.038 (0.047)
Rater Age	-0.052 (0.040)
Tweet Length	-0.0004 (0.0005)
Note Length	0.002*** (0.0005)
Tweet Sentiment	0.058 (0.066)
Note Sentiment	0.182* (0.085)
Tweet FK Score	-0.001 (0.001)
Note SK Score	-0.00001 (0.0003)
Note URL Count	0.494*** (0.106)
Tweet URL Count	-0.048 (0.062)
Note writer Partisanship Score	-0.246** (0.075)
Rater Partisanship Score	0.206** (0.072)
Tweeter Partisanship Score	0.120* (0.050)
Note writer Partisanship Score X Rater Partisanship Score	1.268*** (0.074)
Note writer Partisanship Score X Tweeter Partisanship Score	-0.054 (0.045)
Rater Partisanship Score X Tweeter Partisanship Score	-0.517*** (0.058)
Note writer Partisanship Score X Rater Partisanship Score X Tweeter Partisanship Score	-0.073 (0.052)

*Note:*

\*p&lt;0.05; \*\*p&lt;0.01; \*\*\*p&lt;0.001