Analyzing Biases in Perception of Truth in News Stories and Their Implications for Fact Checking

Mahmoudreza Babaei, Juhi Kulshrestha, Abhijnan Chakraborty,
Elissa M. Redmiles, Meeyoung Cha, and Krishna P. Gummadi

Abstract—Misinformation on social media has become a critical problem, particularly during a public health pandemic. Most social platforms today rely on users' voluntary reports to determine which news stories to fact check first. Despite the importance, no prior work has explored the potential biases in such a reporting process. This work proposes a novel methodology to assess how users perceive truth or misinformation in online news stories. By conducting a large-scale survey (N=15,000), we identify the possible biases in news perceptions and explore how partisan leanings influence the news selection algorithm for fact-checking. Our survey reveals several perception biases or inaccuracies in estimating the truth level of stories. The first kind, called the total perception bias (TPB), is the aggregate difference in the ground truth and perceived truth level. The next two are the false-positive bias (FPB) and false-negative bias (FNB), which measure users' gullibility and cynicality of a given claim. We also propose ideological mean perception bias (IMPB), which quantifies a news story's ideological disputability. Collectively, these biases indicate that user perceptions are not correlated with the ground truth of new stories; users believe some stories to be more false and vice versa. This calls for the need to fact-check news stories that exhibit the most considerable perception biases first, which the current voluntary reporting does not offer. Based on these observations, we propose a new framework that can best leverage users' truth perceptions to (1) remove false stories, (2) correct misperceptions of users, or (3) decrease ideological disagreements. We discuss how this new prioritizing scheme can aid platforms to significantly reduce the impact of fake news on user beliefs.

Index Terms—Perception of News, Fake News Detection, Fact Checking, Online Misinformation, Perception Bias

1 Introduction

Policymakers, technologists and media watchdog groups have frequently criticized social media sites (e.g., Facebook and Twitter) for allowing misinformation to spread unchecked on their platforms [1]. Such unabated fake news spread has been linked to foreign meddling in political elections [2], [3], riots and mass displacements [4], and even loss of human lives [5].

To counter the propagation of fake news, prior research works have attempted to develop tools to automatically detect fake news, by identifying different linguistic features employed by the fake news creators [6], by analyzing the propagation patterns of different news stories in social media [7], or by checking a new content against a database of known fake and real news [8], [9]. Despite these advances, fully automated fake news detection mechanisms are yet to replace human fact-checkers due to their limitations in adapting to dynamic news contexts without human supervision [10], and for their lack of responsibility [11]. Though social media outlets have adopted a mix of machine learning methods to assist human decisions, it is still the human fact-checkers who assign the final labels [12].

For this purpose, the platforms have partnered with independent fact-checking outlets like Snopes, PolitiFact, FullFact and FactCheck, who follow principled methods (e.g., Poynter's Code of Principles[1]) to fact-check stories [12]. Stories deemed likely false by fact-checkers are then ranked lower in users' news feeds or timelines, significantly limiting their future views [13], [14].

However, fact-checking by human experts is a highly resource-constrained process. Even after automatically removing near-duplicates to already fact-checked content [12], it is not possible to check every new information getting circulated on social media. Given the rate at which new information is generated, and as more platforms and users rely on fact-checking systems, the decision on which stories fact-checkers should review first becomes a critical issue [15]. Thus, the most pertinent question that emerges in this context is *how should the platform decide which news stories are check-worthy?* Social media sites currently encourage their users to report any news they encounter and perceive to be fake [16], [17]. Stories reported as fake by numerous people are then prioritized for validation. In essence, to counter the proliferation of fake news, *social media platforms are relying on their users' perceptions of the truthfulness of news* to prioritize stories for fact-checking.

Despite this reliance on user perceptions, no prior study has focused on understanding how the crowd perceives truth in news stories and how these perceptions affect the detection and possible correction of online falsehoods. In this work, we perform an in-depth analysis of users' truth

- *M. Babaei, E. M. Redmiles, and K. P. Gummadi are with Max Planck Institute for Software Systems, Germany.*
- *J. Kulshrestha is with University of Konstanz, Germany.*
- *M. Cha is with Institute for Basic Science and Korea Advanced Institute of Science and Technology, both in South Korea.*
- *A. Chakraborty is with Indian Institute of Technology Delhi, India.*

[1] https://ifcncodeofprinciples.poynter.org

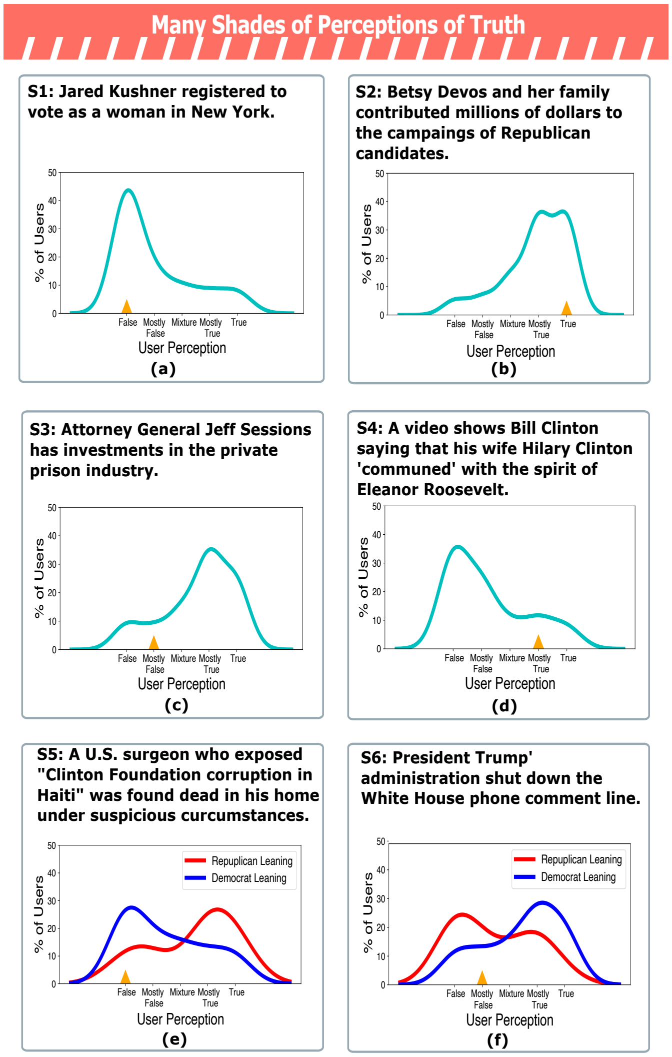


Fig. 1. **Ground truth and perceived truth levels for six different news stories. Here, ground truth level (shown as orange triangles on the x-axis) of each news story is obtained from Snopes, and the perceived truth levels are inferred by gathering the truth perceptions of 100 surveyed users.**

perceptions of individual news stories – not news outlets as in some related work [18] – by designing and validating a novel truth perception test. Using this test, we solicit users' truth perceptions for 150 stories that have already been fact-checked, allowing us to compare the users' perceptions to a known *ground truth level* determined by fact-checkers.

Our comparison of users' perceptions of truth and actual ground truth reveals several discrepancies. To illustrate them, consider the following six stories:

(S1) Jared Kushner registered to vote as a woman in New York — *Fact-checked as False*

(S2) Betsy DeVos and her family contributed millions of dollars to the campaigns of Republican candidates — *Fact-checked as True*

(S3) Attorney General Jeff Sessions has investments in the private prison industry — *Fact-checked as Mostly False*

(S4) A video shows Bill Clinton saying that his wife Hillary Clinton 'communed' with the spirit of Eleanor Roosevelt — *Fact-checked as Mostly True*

(S5) A U.S. surgeon who exposed "Clinton Foundation corruption in Haiti" was found dead in his home under suspicious circumstances — *Fact-checked as False*

(S6) President Trump's administration shut down the White House phone comment line — *Fact-checked as Mostly False*

Figure 1 shows users' truth perceptions for these six stories, along with their fact-checked ground truth levels, as determined by Snopes. The difference between the ground truth and perceived truth levels highlights the need to account for different perception biases.

First, the majority of users correctly inferred the truthfulness of stories $S1$ and $S2$. Since story $S1$ is perceived to be false by most users, the claim will be reported by many and thus likely to be prioritized by the social media platforms for fact-checking. However, we assert that there is *little* to be gained by fact-checking stories whose truth value is already correctly judged by the crowds, just as there is little use in fact-checking claims by news satire outlets like The Daily Show and The Onion.

On the contrary, the figure shows biases in users' truth perceptions for stories $S3$ and $S4$, with significant differences between the truth levels perceived by users and the ground truth. $S3$ reveals *gullibility* of users, where people over-estimate the truth level of the story (*i.e.,* false positive bias), whereas $S4$ reveals users' *cynicality* – people under-estimate the truth level of the story (*i.e.,* false negative bias). Interestingly, $S4$ is more likely to be reported by users and fact-checked with higher priority than S_3. In fact, on today's social media platforms, the higher the false-positive bias in the perceptions of a story, the less likely it is to be reported and become a subject for fact-checking. Worse, currently these platforms do not have mechanisms to reassure users about the credibility of a true story like S_4 that many users mistakenly perceive as false (*i.e.,* high false-negative bias), even after the story is fact-checked.

Figures 1(a-d) also highlight disagreements between users about the truthfulness of individual stories. These disagreements are highly correlated with their political leaning. Figures 1(e) and 1(f) show that users with different political ideologies (*e.g.,* Democrat and Republican-leaning users) indeed perceive truth in news stories differently. People are more likely to trust stories that confirm their political beliefs, while they are more likely to distrust stories that contradict their beliefs. Story $S6$, which attacks Donald Trump's administration, is 'Mostly False' as determined by the expert fact-checkers. However, most users who identify themselves as Democrats perceive this story to be accurate, while most Republican users label it as false. On the other hand, the story $S5$, which raises questions against Hillary Clinton, is 'False' according to the expert fact-checkers. Most Republican-leaning users perceive it to be accurate, and Democrat users perceive it as false. These examples highlight the pitfalls of ignoring biases in truth perceptions when using them to prioritize stories for fact-checking.

In this paper, we propose a framework (shown in Figure 2) for social media platforms to prioritize stories for fact-checking by effectively leveraging users' truth perceptions to satisfy three important objectives:

- **O1. Removing false news from circulation.** False stories need to be fact-checked with higher priority to restrict their circulation on social media

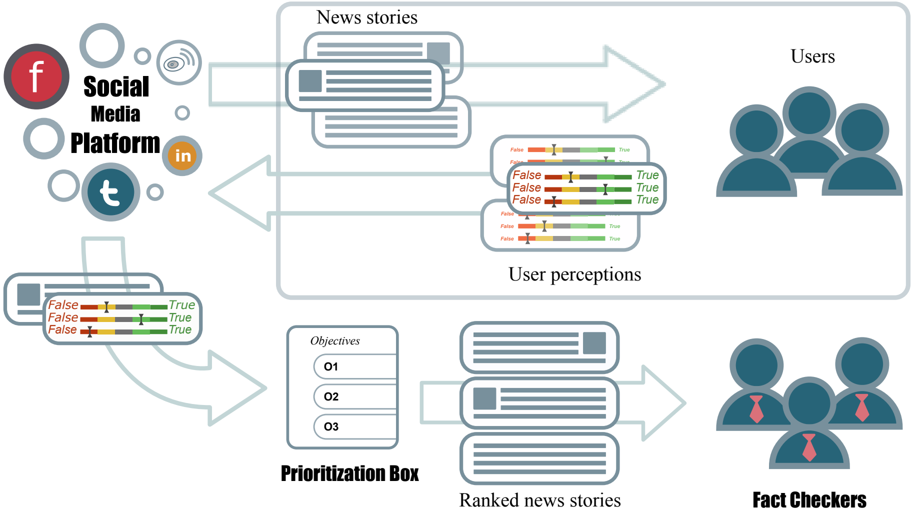


Fig. 2. **Overview of our proposed framework for social media platforms to prioritize stories for fact-checking by leveraging users' truth perceptions. The platforms should first gather users' perceptions of truth for the news stories widely shared on the platform, using our proposed Truth Perception Test (RQ1). Then they can pass these news stories, along with the users' truth perceptions, to the prioritization box (RQ2) and specify the prioritization objective. The prioritization box would then output a ranked list of news stories, based on the platform's chosen objective, which should be sent to the fact-checkers (RQ3).**

platforms. Intuitively, this objective has been the primary focus of social media platforms. Since the truth values of stories are not known beforehand, prior research efforts [6], [9], [19] have focused on automatically detecting potentially false stories. Such potential false stories can then be prioritized for fact-checking.

- **O2. Correcting the misperception of users.** While most of the prior work has argued for the removal of false stories from social media platforms, legal experts and free speech campaigners have compared it to censorship [20]. To address such concerns, social media platforms may want to prioritize and fact check stories for which users' perceived truth levels are far from their ground truth levels, and flag these stories rather than removing them altogether.
- **O3. Decreasing the disagreement between different users' perceptions of truth.** For society to have fruitful debates in the public sphere, it is essential to set a common ground for different sections of the society. To ensure a common ground, the platforms should identify topics that incur a significant degree of disagreement and then prioritize them for fact-checking to let people know the *objective* truth value of the stories. In our experiments, such stories have a high variance in truth perceptions reported by different users, especially when they have different ideological leanings.

Given this context, we focus on the following three research questions in this paper:

- **RQ1:** How can we collect users' perceptions of truth in news stories in a robust manner?
- **RQ2:** How do the three objectives for fact-checking compare to one another? Can they be satisfied simultaneously?
- **RQ3:** If a platform chooses an objective for prioritizing stories for fact-checking, how can the objective be implemented by leveraging users' perceptions of truth in news stories?

Our contributions: While answering these questions, we make three primary contributions in this paper.[2]

First is the methodological contribution. We develop a new method for assessing users' truth perceptions of news stories (N=15,000). Our test asks users to *rapidly* assess (*i.e.*, at the rate of a few seconds per story) how truthful or untruthful the claims in a news story are. We conducted our truth perception tests online and gathered truth perceptions of 100 Amazon Mechanical Turk (AMT) workers [22] for every story. With different experiments, we prove the robustness and efficacy of the proposed test.

Second is the empirical findings. Our exploratory analyses of users' truth perceptions yield several interesting findings. For instance, (i) for many stories, the collective wisdom of the crowd (average truth rating) differ significantly from the actual truth of the story, *i.e.*, the wisdom of crowds is inaccurate; (ii) across different stories, we find evidence for both false-positive perception bias (*i.e.*, a gullible user perceiving the story to be more true than it is in reality) and false negative perception bias (*i.e.*, a cynical user perceiving a story to be more false than it is in reality); and (iii) users' political ideologies (*e.g.*, whether they support democrats vs. republicans) influence their truth perceptions for the most controversial stories (*i.e.*, stories with high variance in truth perception between users).

Third is the practical suggestions derived from data. Our predictive analysis of users' perception biases reveals the limitations of current strategies for selecting a small set of news stories to fact check, based on how many users report the story as fake. We provide a proof of concept simulation of how our truth perception test, coupled with a

[2] This work is an extension of the abstract published at [21].

SVM classifier can be used to achieve the three goals stated earlier for prioritizing stories for fact-checking. We hope that this study will spawn future research on the design of mechanisms to signal the fact-checked label to the users such that they are receptive to them. Such mechanisms, aided by the understanding of users' truth perceptions, should be more helpful in curtailing the spread of fake news.

2 Related Work

Over the recent years, a growing amount of efforts have been put into detecting false information by analyzing large-scale digitally logged user behavioral and social network data on the web. False news include two kinds of information types. The first kind is *misinformation* (*i.e.,* a piece of information that happens to be wrong). The second information type is *disinformation* (*i.e.,* a piece of information that is intentionally manipulated to be wrong). In this section, we briefly discuss the literature on false news in light of its information types and detection methods.

2.1 Rumor Detection

Researchers have long investigated online *rumors*, a term to describe claims that are yet to be verified as 'true' [23]. Based on the theoretical studies on characterizing online rumor behaviors [24]–[27], computer science researchers have developed rumor detection algorithms using features across multiple categories. Machine learning models have been tested based on features describing linguistic characteristics and diffusion patterns of rumors [19], [28], [29]. Kwon *et al.* [7] compared classification capabilities across such multiple feature categories and built an algorithm that achieves a competitive accuracy at an early stage of rumor spreading. Another line of studies proposed deep learning approaches to detect rumors without a labor-intensive feature engineering. Ma *et al.* [30] proposed a RNN-based algorithm to learn sequential information on online rumor spreading. From experiments on Twitter and Weibo, their approach outperformed existing feature-based algorithms and further tackled early detection problems. Other newly proposed deep learning models combine temporal activity patterns of spreaders and source characteristics into existing features. In particular, a model called CSI [9] showed the state-of-the-art performance in detecting rumors on social media.

2.2 Identification of Clickbaits

Presentation of false news often takes a sensational form to attract readers by using a psychological technique known as the curiosity gap [31], [32]. Such sensational stories are known as clickbait, and recent studies have focused on identifying *clickbait articles*, where news headlines and the associated body text have a discordance relationship [33]. Chakraborty *et al.* [31] developed an SVM model that predicts clickbait articles based on linguistic patterns. On the same dataset, another group suggested a neural network approach that measures textual similarities between the headline and the first paragraph [34]. Further research works have attempted *stance detection* in different news articles [35], [36], trying to correlate the usage of clickbait articles and the ideological leaning of the media houses.

2.3 Fake News Detection

Detecting fake news is a challenging task even for human evaluators. A crowdsourced study has shown that human annotators gain marginal improvements (66%) over random guesses (50%) [37]. Such findings justify the need for an automated fact-checking system that can process and remember more information than human evaluators. Such automated methods to detect fake news can employ different approaches based on the content, source, or the propagation pattern of fake news.

2.3.1 Content-based methods

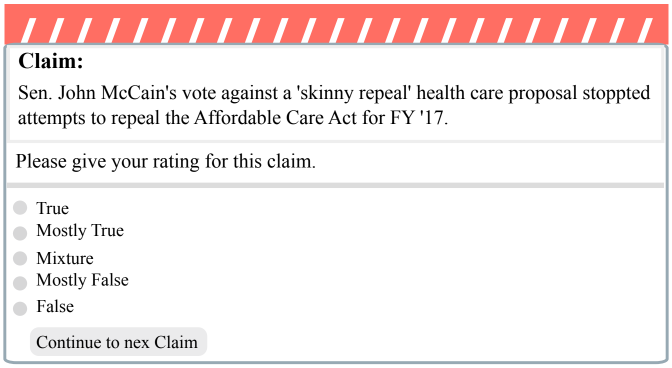
One way to assess the authenticity of news is to evaluate its content, such as text or images. Traditional machine learning frameworks used a set of manually selected features at various language levels such as lexicon, syntax, semantic, and discourse-level to detect fake news [38]–[40]. Later, by embedding text [41] and images as news content to word-level [42] or pixel matrix, well-trained neural network models have been used to extract latent textual and visual features of news content, a given news is classified as true or fake news. Studies as in [43] propose a neural network-based model to automatically find mismatches in text and image of an online post (i.e., clickbait detection).

2.3.2 Propagation-based methods

Malicious spreaders can easily manipulate the content-based methods that are being used for detecting fake news. Thus, several studies focus on other methods; for example, [44] claim that fake news has different patterns compared to true news, such as having high informality and diversity as well as being more emotional. Authors in [45] observed that fake news spreads through social media with different patterns compared to true news. Several cascade features such as cascade size, cascade breadth, cascade depth, structural virality (Average distance among all pairs of nodes in a cascade), node degree, spread speed, and cascade similarity are used to classify the news as fake or true [45]–[47]. Authors in [48] and [49] developed recursive neural networks based on news cascades to classify the news. Many recent works have developed several models, like SIRS, SIS, SEIR, epidemic SEIR, and *etc.*to model rumor spreading in online social networks in which the goal is to detect and eliminate fake news [50]–[53].

2.3.3 Source-based methods

Some approaches attempt to detect fake news by focusing on the credibility of its source, covering not only the sources that create and publish the news but also the sources that spread the news stories [54]–[56]. Assessing a few outlets' credibilities, such as traditional mass media or popular news publishers in social media, might be useful. Sitaula *et al.* [57] constructed the collaboration network of news authors in which they show that the networks are homogeneous, meaning that the fake-news authors are more densely connected. True news authors are also strongly connected, while there is a weak connection across the groups. Multiple independent efforts also show the credibility of news sources; examples include Media Bias/Fact Check (`www.mediabiasfactcheck.com`) or



Claim:

Sen. John McCain's vote against a 'skinny repeal' health care proposal stoppted attempts to repeal the Affordable Care Act for FY '17.

Please give your rating for this claim.

○ True
○ Mostly True
○ Mixture
○ Mostly False
○ False

Fig. 3. **An example of the survey question that we used for performing Truth Perception Tests for the news claims in our dataset.**

NewsGuard (`www.newsguardtech.com`, which provide the list of news sources with their credibility based on a different point of view such as political leaning. However, every social media user can be a source of news, and they might be malicious users (who intentionally spread fake news the same as bots) or vulnerable normal users (who spread fake news unintentionally without recognizing the falsehood). Several works detect malicious bots using groups of features such as network, user, friend, temporal, content, sentiment [58], [59].

2.3.4 Reliance on human fact-checking

Despite such advancements in automated fake news detection methods, in the context of misinformation research and practice taken by major platforms, algorithmic scores on information veracity remain as "suggestions" and the final labels on true or false are still made by the human fact-checkers (see [12]–[14] for Facebook, [14], [60] for Instagram, and [17], [61] for Twitter). Facebook in their fake new detection policy page [13] mentions how they label false information relying on third-party fact-checkers. Facebook partners with almost 35 fact-checking outlets in 24 countries [62]. Similarly, Twitter also claimed that they rely on trusted partners to identify content that is likely to result in offline harm [17], [61]. Researchers have called for Twitter and all other platforms to reach out more to fact-checkers and work with them for these kinds of actions [63]. However, fact-checking by human experts is a highly resource-constrained process. Full-Fact, one of the Facebook fact-checkers partners, claimed that the process becomes slower because Facebook provides them a list of news that users flagged, and many of them may be just opinions or not harmful [16]. In this paper, we focus on the pitfalls of such prioritization strategies for fact checking. In the next section, we introduce the data and methodology used for such prioritization.

3 Designing Truth Perception Tests

In this section, we address the first research question (RQ1) of how one can design a test that allows to measure users' perceptions of truth in a robust manner.

3.1 Methodology

As mentioned in the introduction, we designed *Truth Perception Tests* (TPTs) that can be used to assess how users implicitly perceive truth in news stories — *i.e., Perceived Truth Level (PTL)*. We perform TPTs as online surveys. While we did not limit our respondents to a specific time frame, we strongly encouraged them to respond rapidly by giving them the following instructions at the start of the test: *"Please do not conduct any web search or use any online/offline resources for verifying or validating the claim presented to you. Please use your best judgment (your instinctive gut based guess within a few seconds) to label the claims."* On average, our respondents gave their truth perception responses for each claim within 10 seconds.

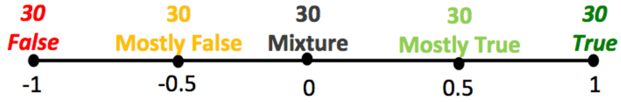


Fig. 4. **Mapping truth labels of news stories on a scale between -1.0 and + 1.0. The number of stories collected for each ground truth label are also indicated.**

To gather truth perceptions, we showed respondents a news claim and asked them to label the claim as either 'True', 'Mostly True', 'Mixture', 'Mostly False' and 'False', as shown in the example depicted in Fig. 3. We mapped these five perceived truth level (PTL) choices to a scale between -1.0 and +1.0. By aggregating the answers given by each user u for a news story S, $PTL_u(S)$, we compute the aggregate perceived truth level, $PTL(S)$, of the story as follows:

$$\text{PTL(S)} = \frac{1}{N}\sum_{u=1}^{N} PTL_u(\text{S}) \tag{1}$$

where N is the total number of users whose truth perceptions for the story S are being aggregated.

3.2 Data Collection

The news claims utilized in the TPTs were drawn from news stories that had been professionally fact-checked by Snopes and thus we know their ground truth level. Snopes uses the same set of labels that we used as our answer choices to categorize news stories: 'False', 'Mostly False', 'Mixture', 'Mostly True', and 'True'. Again, we mapped these truth categories on a scale between -1.0 and +1.0, as shown in Fig. 4. In January 2018, from the claims labeled under the Politics topic category by Snopes, we selected 30 recently fact checked news stories for each truth category to get a total of 150 stories. The ground truth level for each story S, $GTL(S)$, is given by the value of the truth category assigned by Snopes for that story.

We ran our validated truth perception tests on Amazon Mechanical Turk (MTurk), collecting a total of 15,000 responses. Each MTurk worker saw 50 claims and no worker could take the survey more than once. Any Mturk worker over the age of 18 who resided in the US was eligible to participate in our survey.

3.3 Test Design Validation

To ensure that our TPTs are maximally robust to variations in deployment and a broad set of potential survey biases,

Surveys	χ^2 dependency of Dist-ANS and Acc	Correlation of TBT
MTurk Masters & MTurk naive	χ-value:0.0 p-value=1.0	0.90
MTurk naive & SSI workers	χ-value:0.0 p-value=1.0	0.89
7-pt scale & 6-pt scale	χ-value:0.0 p-value=1.0	0.94
7-pt scale & 5-pt scale	χ-value:0.0 p-value=1.0	0.98
5-pt scale & incentive and 5-pt scale	χ-value:0.0 p-value=1.0	0.85
7-pt scale & incentive and 7-pt scale	χ-value:0.0 p-value=1.0	0.92

TABLE 1

Survey variation effects: We evaluate the similarity between the distribution of answers to each survey using a χ^2 test of independence. χ-values and p-values for all tests are close to 0 and 1 respectively. The first two rows depict the results on sample effects. The subsequent two rows show the Answer choice effect. The bottom two rows correspond to Satisficing and Incentive effects.

we conducted multiple micro-experiments. In these micro-experiments we evaluated how, if at all, different test designs may influence our results. Specifically, we evaluated three types of effects: Sample Effects, Answer Choice Effects, Satisficing and Incentive Effects.

3.3.1 Sample Effects

Literature on survey methodology [64]

reports that expert respondents may answer certain survey questions differently than naive respondents. Additionally, demographic composition of the survey sample is known to affect the generalizability of results [65]. Therefore, to account for such sample effects, we compare two survey variations: we run the test (i) using Amazon Mechanical Turk (MTurk) Masters (i.e., expert participants) [22] versus naive MTurk workers, both from the US, and, (ii) using a census-representative sample of participants recruited by Survey Sampling International[3] (SSI participants) versus the MTurk Masters.

We evaluate the similarity between the distribution of answers to the survey variations using a χ^2 test of independence. Table 1 shows that the χ-values and p-values for all tests are close to 0 and 1 respectively, which depicts that both distributions of answering and accuracy of judgments are independent (i.e., they fail to reject the null hypothesis H0) of types of survey respondents (first and second rows in Table 1). The last column in Table 1 shows that there is a high correlation between Total Perception Bias (TPB) of claims of different surveys with different worker samples. Similar values of TPB for particular claims in different surveys confirm that our measure is robust against the sample effects.

Note that, **Total Perception Bias (TPB)** of a story S captures the total error (gullibility or cynicality) in the users' perceptions of truth levels of the story, and is given by

$$\text{TPB(S)} = \frac{1}{N}\sum_{u=1}^{N}|PTL_u(\text{S}) - GTL(\text{S})| \quad (2)$$

where N is the total number of users whose truth perceptions of the story S are being aggregated.

3.3.2 Answer Choice Effects

It has been reported previously [66] that Likert scale length (e.g., even or odd numbers of answer choices, where scales with an odd number of answer choices include a "middle" neutral option), may effect the strength of participants' responses. We compared the effects of using a 6 and 7 point Likert item scale. Additionally, the text labels of the Likert answer choices may also affect respondents' answers to survey questions[4]. To examine this effect, we compared the effect of using the Snopes' labels (see Fig. 3) with an alternate 7 point scale ("I can confirm it to be true", "Very likely to be true", "Possibly true", "Can't tell", "Possibly false", "Very likely to be false", and "I can confirm it to be false") and 6 point scale (which excluded the "Can't tell" option from the 7 point scale). We evaluated the answer choice effects by comparing 6, 7 point scale with Snopes' 5 point scale.

Table 1 (third and fourth rows) depicts that both distributions of answering and accuracy of judgments are independent (fail to reject H0) of the types of answer choices in the surveys. A significantly high correlation between TPB of claims of different surveys with different answer choices is shown in the last column.

3.3.3 Satisficing and Incentive Effects

Satisficing [68] is a commonly observed survey response effect in which respondents select what they consider to be the minimum acceptable answer, without fully considering their true feelings. Surveys such as our TPTs may be at particular risk of satisficing because they encourage quick responses. Thus, we explored the effect of incentivizing participants to provide correct answers to evaluate whether satisficing may be affecting our test results.

To investigate the impact of satisficing and incentives, we designed a survey in which we gave respondents incentives for answering correctly. At the beginning of the survey, we told the participants: "In addition to the amount promised for the task, for each of your judgements which CORRECTLY matches the actual truth status of the claims, we will pay you 5 cents as a bonus. For example, if you judge a claim to be 'True', or 'Mostly True', and the claim is actually true, then you'll get 5 cents for the claim. Similarly, to get the bonus for an actual false claim, it should be judged by you as 'False' or 'Mostly False'. Finally if you judged the claim as 'Mixture' and the claim actually is mixture or mostly true/mostly false you will earn bonus." To ensure that participants do not use online or offline resources to estimate the truthfulness of the claims we showed a timer in each page and told them: "If your judgment for each question takes more than 15 seconds then there would not be any bonus, even if you answer the question correctly."

To test whether incentivizing has any effect, we compare the results of the incentivized survey with the unincentivized one. The last two rows in Table 1 shows that incentivizing does not affect the survey results. We found no statistically significant differences across the survey variations for the proportion of correct answers. Additionally, we

[3] https://surveysampling.com

[4] Keeping text labels on Likert item points has been advocated as the best practice in multiple prior works [67], hence we do not examine the omission of text labels.

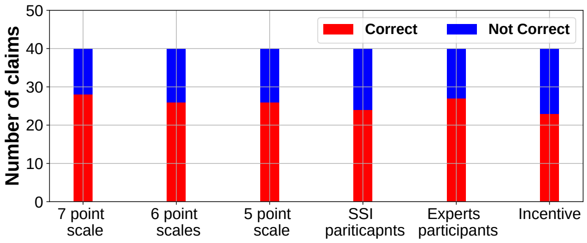


Fig. 5. **Accuracy of judgments (wisdom of crowd) for different design surveys.**

observed statistically significant high-correlation between our proposed measure of TPB, computed for our survey variations, with the Pearson correlation coefficients ranging from 0.90 to 0.96. Figure 5 depicts that the wisdom of the crowd (accuracy of judging by users) is very similar across different survey variations.

Figure 6 summarizes the results of comparing TPB test variations, which show very similar results across variants. We thus conclude that our test is relatively robust and consequently useful for application in industry settings and future research on content misperceptions.

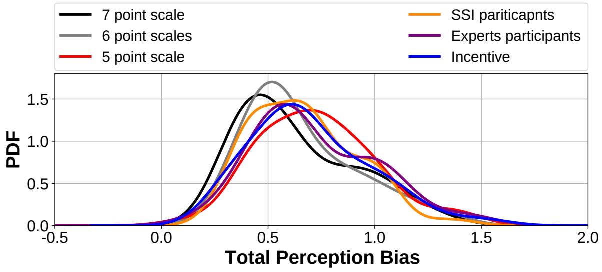


Fig. 6. **Total Perception Bias (TPB) for different design surveys.**

3.3.4 *Limitations*

While we validate our truth perception tests extensively to ensure they are robust against design variations, our method does have some limitations. When users encounter and flag false news stories on the social media platforms, they are not only exposed to the claim or headline, but also to the source of the article, the images from the article, summary snippet or text of the article, and additional context for instance likes or shares for the story etc. Our controlled experiments do capture the effect of the claim (or headline) of the news stories on the users, but they do not capture the effects of other factors as yet, and a promising direction of future work would be to design controlled experiments to measure the impact of the other factors.

4 Comparing the Prioritization Objectives

This section addresses the second research question (RQ2) of how the three objectives of fact-checking compare to one another. As discussed earlier, social media sites today prioritize stories based on the number of reports they receive from users flagging a piece of content as false. This approach assumes that false stories will receive more reports from users than true stories and hence will be fact checked with higher priority than true stories. Figure 7 depicts the perceived truth levels of the six stories mentioned in Introduction, versus their ground truth levels as determined by Snopes.

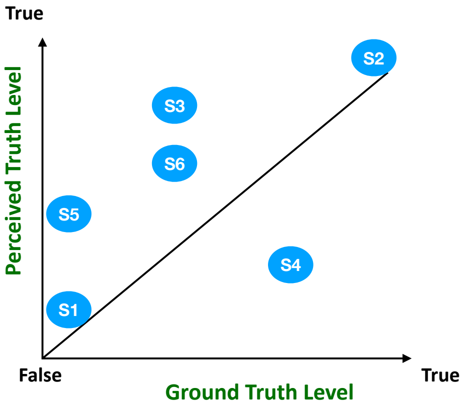


Fig. 7. **Sample stories from Fig. 1. Perceived Truth Level is determined by averaging truth perceptions of 100 AMT workers. Ground Truth Level is determined by Snopes.**

Under the current strategy, the priority order would be S1, S4, S5, S6, S3, and then S2, while the desired ranking to satisfy the first objective (removing false news stories by ranking according to GTL) would be S1, S5, S6, S3, S4, and then S2. Thus the current strategy does not satisfy this objective satisfactorily. Based on our analysis of users' truth perceptions, we identified two additional objectives for fact checking stories: O2 (correcting users' misperceptions) and O3 (decreasing disagreement among users). Thus, we also seek to evaluate: *Does the current strategy satisfy O2 or O3? Are these three objectives (O1, O2 and O3) compatible and can one strategy address them simultaneously?*

For comparisons, we prioritize the six stories according to every objective. For objective O2, we use the metric of Total Perception Bias (TPB) to rank the stories, where TPB captures the aggregate deviation of perceived truth level (aggregated over N users) from the ground truth level of a story S that we discussed earlier. To rank stories according to O3, we can either rank news stories using Disputability (*i.e.,* the variance in the individual truth perceptions of users) or according to Ideological Mean Perception Bias (IMPB) which captures the difference in truth perceptions of different ideological groups (Democrats and Republican in this case), given by

$$IMPB(S) = |MPB_{Dem}(\text{S}) - MPB_{Rep}(\text{S})| \qquad (3)$$

where, $MPB(S)$ = PTL(S) - GTL(S), measures the error in the collective perceptions of users in assessing the truth level of a story.

When we rank the example stories based on the three objectives using these metrics, we get different priority orders:

- Priority order to satisfy O1: S1,S5,S6,S3,S4,S2
- Priority order to satisfy O2: S4,S5,S3,S6,S2,S1
- Priority order to satisfy O3 (Disputability): S3,S5,S6,S2,S1,S4
- Priority order to satisfy O3 (IMPB): S5,S6,S1,S4,S3,S2

	O1 and O2	O1 and O3	O2 and O3
Spearman's ρ	0.31	-0.05	-0.01

TABLE 2
Correlation between rankings to satisfy different objectives.

Moreover, when we consider the full dataset of all 150 news stories and rank them according to each objective, we observe little correlation between the rankings according to these different objectives. Table 2 presents the Spearman's rank correlation coefficient ρ[5] between stories ranked by different objectives. While we can observe some association of ranks between O1 and O2, there is almost no association (ρ close to 0) between the other pairs of objectives. Thus, we can conclude that *these three objectives are incompatible, and can not be satisfied simultaneously*.

Thus, platform providers must chose one objective over the others to prioritize stories. The choice of objective will mean that an entirely different set of potentially "fake" news stories will remain unverified. To illustrate this effect, Fig. 8 displays the top five news stories, ranked by each objective. Thus, special care needs to be taken by the platforms to finalize the design of their fact checking exercise.

5 Operationalizing Different Objectives using Truth Perceptions

This section investigates the final research question (RQ3) of how social media platforms may prioritize stories for fact-checking based on their users' perceptions of truth in news stories.

5.1 Decreasing disagreement among different users' truth perceptions (O3)

We start by describing the easiest objective to operationalize (O3). The goal is to prioritize stories that have the highest disagreement in user truth perceptions. We quantify the disagreement in users' perceptions as the *disputability* of news stories, *i.e.*, the variance in the individual truth perceptions of users. The platform can collect users' truth perceptions and rank the stories according to their disputability to satisfy O3.

If the ideological leanings of the users assessing the stories are known then the stories which have a maximum disagreement in the perceptions of users with different ideologies can be prioritized. We capture such differences in assessment as the Ideological Mean Perception Bias (IMPB) of a story, as defined earlier. Most social media platforms, such as Facebook and Twitter, have detailed information about their users via users' explicit inputs or behavior on these platforms, including information on their potential ideological leaning. Therefore, platforms could compute the IMPB of a story to assist in fact checking prioritization.

Further, even in the absence of such information about the ideological leanings of users, it is possible to achieve O3. We found that the disputability of stories is moderately correlated (Pearson Correlation: 0.38) with IMPB. Thus, prioritizing stories by disputability also prioritizes stories with higher variation in perception between users with different ideological leanings.

5.2 Correcting the misperception of users (O2)

To correct users' misperceptions, we need to quantify the extent to which users incorrectly perceive the truth of a story. To do so, we use the previously defined Total Perception Bias (TPB) metric to measure the aggregated error (gullibility or cynicality) in users' perceptions of a story S. Ranking stories by TPB prioritizes misperceived stories: stories where users' perceived truth levels (PTL) differ widely from the ground truth level (GTL) of the story. However, to compute TPB, we must know GTL, which is not available in practice. Here, we propose an alternative approach: training a supervised learning classifier that classifies a story as having either high or low TPB. To design such a classifier, we need the GTL of a small set of stories that have been labeled as high or low TPB for generating the training data. Then, TPB can be predicted for a larger set of stories for which GTLs may not be known.

As an illustration, we construct a classifier to predict the TPB values for the 150 stories we studied in this work. We label a news story to have 'High TPB' if it has a TPB value above the median TPB value, or 'Low TPB' if it has a value lower than the median. We split our dataset of 120 claims, and consider 80% of the data (96 claims) as the training dataset and the remaining 20% (24 claims) as the test dataset. Using this ground truth dataset, we train four types of classifiers (Linear SVM, Naive Bayes, Logistic Regression, and Random Forest). Our feature set includes the mean, median, variance, and skew of perceptions of users with different demographic features such as 'Political Ideology', 'Age', 'Gender', 'Education', 'Employment', 'Income', and 'Marital status' (Table 3 shows the distribution of different demographic attributes). Applying feature ranking with recursive feature elimination, we observed that the best set of features includes 'Political Ideology', and 'Income'.

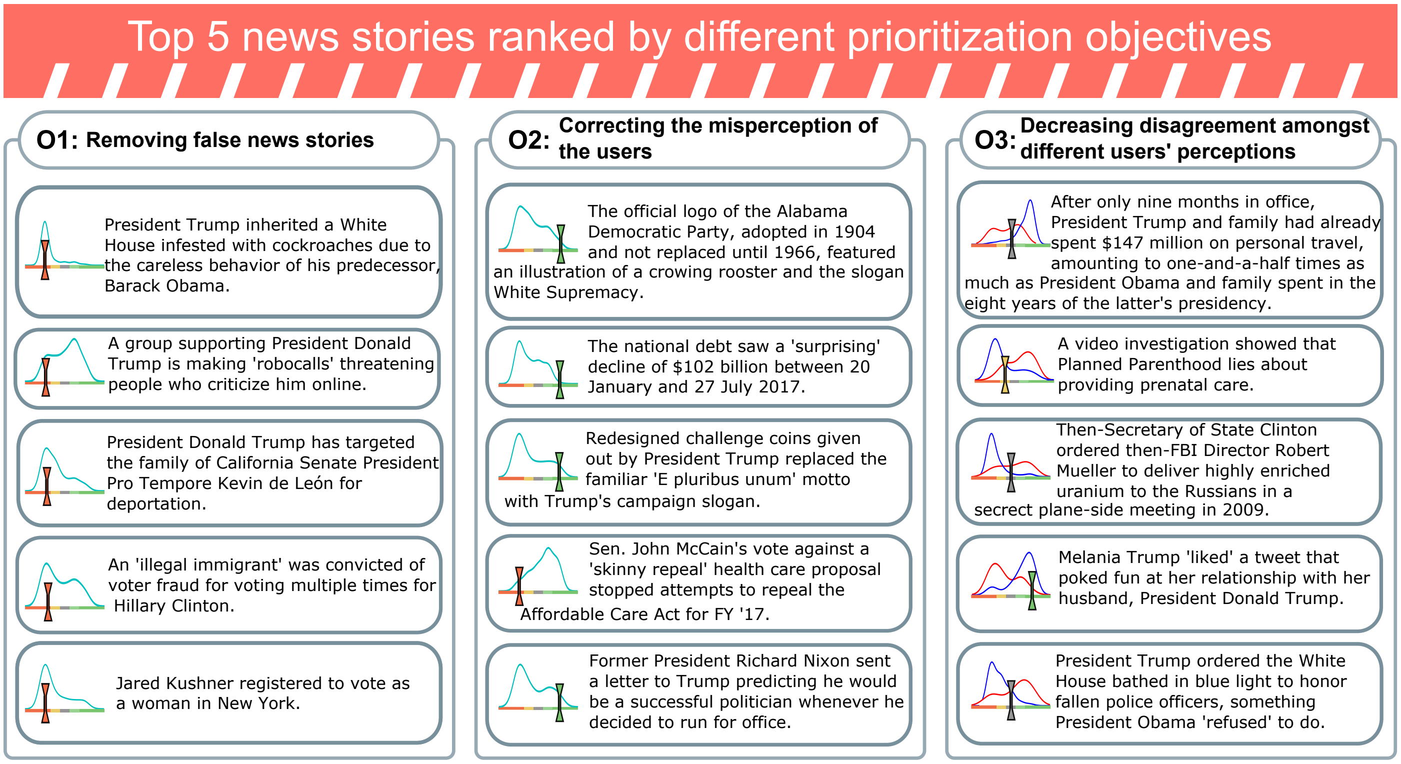
For evaluation, we use 5-fold cross-validation. In each test, the original sample is partitioned into 5 sub-samples, out of which 4 are used as training data, and the remaining one is used for testing the classifier. The process is then repeated 5 times, with each of the 5 sub-samples used exactly once as the test data, thus producing 5 results. The entire 5-fold cross-validation was then repeated 20 times with different seeds used to shuffle the original dataset, thus producing 100 different results. The results reported are average accuracies across these 100 runs, along with the 90% confidence interval.

We observe an average prediction accuracy of 82% (using Linear SVM & Random Forest classifiers), with 90% confidence interval of 0.09%, illustrating the potential for satisfying O2 given a small ground truth dataset.

In the second column of Table 4, we depict the performance as the average accuracy across the 100 runs along with the 90% confidence interval of the four types of supervised classifiers for our prediction task, using the best set of features (including 'Political ideology' and 'Income') determined by feature ranking with recursive feature elimination. As shown in the table, we achieve maximum accuracy of 82%.

Note that our prediction algorithm for TPB of news stories is based only on users' truth perceptions and their basic demographic attributes. We believe, predictive performance

[5] The value of ρ ranges between $+1$ (positive correlation) and -1 (negative correlation), with 0 denoting no correlation.



Top 5 news stories ranked by different prioritization objectives

O1: Removing false news stories

President Trump inherited a White House infested with cockroaches due to the careless behavior of his predecessor, Barack Obama.

A group supporting President Donald Trump is making 'robocalls' threatening people who criticize him online.

President Donald Trump has targeted the family of California Senate President Pro Tempore Kevin de León for deportation.

An 'illegal immigrant' was convicted of voter fraud for voting multiple times for Hillary Clinton.

Jared Kushner registered to vote as a woman in New York.

O2: Correcting the misperception of the users

The official logo of the Alabama Democratic Party, adopted in 1904 and not replaced until 1966, featured an illustration of a crowing rooster and the slogan White Supremacy.

The national debt saw a 'surprising' decline of $102 billion between 20 January and 27 July 2017.

Redesigned challenge coins given out by President Trump replaced the familiar 'E pluribus unum' motto with Trump's campaign slogan.

Sen. John McCain's vote against a 'skinny repeal' health care proposal stopped attempts to repeal the Affordable Care Act for FY '17.

Former President Richard Nixon sent a letter to Trump predicting he would be a successful politician whenever he decided to run for office.

O3: Decreasing disagreement amongst different users' perceptions

After only nine months in office, President Trump and family had already spent $147 million on personal travel, amounting to one-and-a-half times as much as President Obama and family spent in the eight years of the latter's presidency.

A video investigation showed that Planned Parenthood lies about providing prenatal care.

Then-Secretary of State Clinton ordered then-FBI Director Robert Mueller to deliver highly enriched uranium to the Russians in a secrect plane-side meeting in 2009.

Melania Trump 'liked' a tweet that poked fun at her relationship with her husband, President Donald Trump.

President Trump ordered the White House bathed in blue light to honor fallen police officers, something President Obama 'refused' to do.

Fig. 8. **The top 5 ranked news stories prioritized according to the three objectives of social media platforms for selecting stories for fact checking. The low overlap between the three ranked lists highlights the complementary nature of the objectives.**

Demographic attribute	Attribute values
Political ideology	Conservative (21%), Moderate (24%), Liberal (55%)
Age	18-24 (7.5%), 25-34 (36%), 35-44 (30%), 45-54 (15%), 55-64 (8.5%), 65-74 (3%)
Gender	Female (48%), Male (52%)
Education degree	College graduate bs/ba or other 4year degree (43%), Postgraduate training or professional schooling after college (10%) Toward a masters degree or PhD law or medical school (12%), Some college associate degree no 4 year degree (16%), High school graduate grade 12 or certificate (12%), Technical trade or vocational school after high school (7%)
Employment	In full-time work permanent (59%), In full-time work temp contract (2%), Retired (1%), Unemployed (4%), In part-time work permanent (4%), In part timework temp contract (3%), Part-time work part-time student (4%), Self-employed (23%)
Income	Under 10000 (6%), 10000-20000 (18%), 20001-30000 (22%), 30001-40000 (14%), 40001-50000 (9%), 50001-60000 (12%), 60001-70000(9%), 70001-100000 (7%), 100001-150000 (2%), 150001 or more (1%)
Marital status	Married (29%),, Living with partner (12%), Divorced (9%), Widowed & Separated (2%), Single (48%)

TABLE 3

Demographic attributes of the Amazon Mechanical Turk workers who participated in the Truth Perception Test. Values in the parenthesis express the demographic distribution.

could be further improved by including more detailed demographic and behavioral features, typically available to the social media platforms.

5.3 Removing false news from circulation (O1)

Finally, to operationalize O1, social media platforms need to prioritize false stories for fact-checking. We examined two methods that leverage the users' truth perceptions (PTL) to estimate the ground truth levels of news stories. For both the methods, we need a labeled ground truth dataset, so we label all the stories annotated to be 'True' or 'Mostly True' by Snopes to be 'True', while labeling all stories annotated to be 'False' or 'Mostly False' by Snopes as 'False'. Ignoring the stories labeled 'Mixture', we were left with a labeled dataset of 60 'True' stories and 60 'False' stories.

We first took a "wisdom of crowds" approach and estimated the GTL using the average PTL value for the 100 workers who assessed the story. We considered stories with a positive average PTL to be 'True', while negative ones to be 'False'. We observed that we correctly assess the truth labels for 67% of stories in our ground truth labeled dataset. Additionally, when we rank stories by PTL and GTL, respectively, we observe a moderate ranking correlation of 0.4.

Alternatively, similar to O2, we trained supervised classifiers to predict the truth value ('True' or 'False') of a story. Using the same set of classifiers, feature set and experimental setup as O2, we achieve an average accuracy of 70% (using Linear SVM & Random Forest classifiers) across the 100 runs, with a 90% confidence interval of 0.7%. In first column of Table 4, we depict the performance as the average accuracy across the 100 runs along with the 90% confidence

	Predicting GTL	Predicting TPB
Linear SVM	0.7±0.007	0.82±0.009
Naive Bayes	0.67±0.008	0.78±0.008
Logistic Regression	0.68±0.010	0.79±0.008
Random Forest	0.7±0.009	0.82±0.010

TABLE 4

Prediction results using different types of supervised methods for the two tasks of predicting GTL and TPB. Performance of each classifier is reported as the average accuracy across the 100 runs along with the 90% confidence intervals.

interval of the four types of supervised classifiers for our prediction task.

Operationalizing O1 proved to be very challenging, as also demonstrated by the amount of prior research on automatically identifying "fake" news stories in recent times [6], [8], [9], [9], [23], [24], [28], [37], [43], [69]–[71]. While we only achieve limited success in operationalizing O1, further improvements could be potentially made in the future, if we can gather more information such as the network structure [23], [37], [71], [72] or engagement of users while sharing the news [9], [37], [72].

6 Conclusion

In this paper, we thoroughly examined how users perceive truth in news stories by developing novel and robust truth perception tests, where users are asked to rapidly assess how true or false the claims in a news story are. We validated our tests against deployment variations and common survey biases such as sample effects, answer choice effects, and satisficing and incentive effects. For our dataset of 150 news claims collected from Snopes.com, we performed our truth perception tests online on the AMT platform to collect users' perceptions of truth in news stories (N=15,000).

Leveraging the users' truth perceptions, we propose a novel framework for prioritizing stories for fact checking, with three potential, competing objectives: (i) removing false news stories from circulation, (ii) correcting the misperception of the users, and (iii) decreasing the disagreement between different users' perceptions of truth. Using a combination of user perceptions elicited using our truth perception tests, users' demographic features, and supervised machine learning methods, we provide operationalization strategies that utilize users' truth perceptions to achieve the above objectives for prioritizing stories for fact-checking.

We believe that our findings will help inform the design of mechanisms for selecting stories to fact check. They can aid social media platform providers and fact-checking organizations to combat fake news more efficiently.

Acknowledgements

K. P. Gummadi acknowledges support from ERC Advanced Grant "Foundations for Fair Social Computing" (No. 789373). M. Cha was supported by the Institute for Basic Science (IBS-R029-C2) and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. NRF-2017R1E1A1A01076400). E. M. Redmiles acknowledges the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and from a Facebook Fellowship for supporting part of her time on this project.

References

[1] T. Barrabi, "Facebook, twitter face congressional hearings on political bias, fake news," https://www.foxbusiness.com/technology/facebook-twitter-face-congressional-hearings-on-political-bias-fake-news, 2018.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, 2017.

[3] S. Valenzuela, D. Halpern, J. E. Katz, and J. P. Miranda, "The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation," *Digital Journalism*, vol. 7, no. 6, pp. 802–823, 2019.

[4] A. Taub and M. Fisher, "Where countries are tinderboxes and facebook is a match," https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html, 2018.

[5] L. Hogan and M. Safi, "Revealed: Facebook hate speech exploded in myanmar during rohingya crisis," https://www.theguardian.com/world/2018/apr/03/revealed-facebook-hate-speech-exploded-in-myanmar-during-rohingya-crisis, 2018.

[6] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[7] S. Kwon, M. Cha, and K. Jung, "Rumor detection over varying time windows," *PloS one*, vol. 12, no. 1, p. e0168344, 2017.

[8] S. Kumar, R. West, and J. Leskovec, "Toward Automated Fact-Checking: Detecting Check-Worthy Factual Claims by ClaimBuster," in *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2017.

[9] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A Hybrid Deep Model for Fake News Detection," in *ACM International Conference on Information and Knowledge Management (CIKM)*, 2017.

[10] L. Graves, "Understanding the promise and limits of automated fact-checking," Tech. Rep., 2018.

[11] B. Nyhan, "Why the fact-checking at facebook needs to be checked," https://www.nytimes.com/2017/10/23/upshot/why-the-fact-checking-at-facebook-needs-to-be-checked.html, 2017.

[12] Facebook, "How is facebook addressing false information through independent fact-checkers?" https://www.facebook.com/help/1952307158131536, 2021.

[13] T. Lyons, "Hard questions: What's facebook's strategy for stopping false news?" https://newsroom.fb.com/news/2018/05/hard-questions-false-news, 2018.

[14] C. Edgecomb, "Facebook and instagram starting to identify and label 'fake news'," https://www.impactplus.com/blog/facebook-and-instagram-starting-to-identify-and-label-fake-news-before-it-goes-viral, 2019.

[15] S. Park, J. Y. Park, H. Chin, J.-h. Kang, and M. Cha, "An experimental study to understand user experience and perception bias occurred by fact-checking messages," in *International World Wide Web Conference (WWW)*, 2021.

[16] D. Funke and A. Mantzarlis, "Report on the facebook third party fact checking programme," https://fullfact.org/media/uploads/tpfc-q1q2-2019.pdf, 2019.

[17] J. Clayton, "Twitter pilot to let users flag 'false' content," https://www.bbc.com/news/technology-55806002, 2021.

[18] G. Pennycook and D. G. Rand, "Fighting misinformation on social media using crowdsourced judgments of news source quality," *Proceedings of the National Academy of Sciences*, vol. 116, no. 7, pp. 2521–2526, 2019.

[19] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *IEEE International Conference on Data Mining (ICDM)*, 2013.

[20] J. Henley, "Global crackdown on fake news raises censorship concerns," https://www.theguardian.com/media/2018/apr/24/global-crackdown-on-fake-news-raises-censorship-concerns, 2018.

[21] M. Babaei, A. Chakraborty, J. Kulshrestha, E. Redmiles, M. Cha, and K. Gummadi, "Analyzing biases in perception of truth in news stories and their implications for fact checking," in *ACM Conference on Fairness, Accountability, and Transparency*, 2019.

[22] "Get better results with less effort with mechanical turk masters, http://mechanicalturk.typepad.com/blog/2011/06/get-betterresults-with-less-effort-with-mechanical-turk-masters-.html," The Mechanical Turk Blog 2011.

[23] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[24] O. Oh, M. Agrawal, and H. R. Rao, "Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises." *Mis Quarterly*, vol. 37, no. 2, 2013.

[25] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Transactions on Computational Social Systems*, vol. 2, no. 3, pp. 99–108, 2015.

[26] A. E. Fard, M. Mohammadi, Y. Chen, and B. Van de Walle, "Computational rumor detection without non-rumor: A one-class classification approach," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 830–846, 2019.

[27] A. Louni and K. Subbalakshmi, "Who spread that rumor: Finding the source of information in large online social networks with probabilistically varying internode relationship strengths," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 2, pp. 335–343, 2018.

[28] Z. Zhao, P. Resnick, and Q. Mei, "Enquiring minds: Early detection of rumors in social media from enquiry posts," in *International World Wide Web Conference (WWW)*, 2015.

[29] P. Shiralkar, A. Flammini, F. Menczer, and G. L. Ciampaglia, "Finding streams in knowledge graphs to support fact checking," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 859–864.

[30] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting Rumors from Microblogs with Recurrent Neural Networks," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[31] A. Chakraborty, B. Paranjape, S. Kakarla, and N. Ganguly, "Stop clickbait: Detecting and preventing clickbaits in online news media," in *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2016.

[32] A. Chakraborty, R. Sarkar, A. Mrigen, and N. Ganguly, "Tabloids in the era of social media? understanding the production and consumption of clickbaits in twitter," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, no. CSCW, pp. 1–21, 2017.

[33] Y. Chen, N. J. Conroy, and V. L. Rubin, "Misleading online content: Recognizing clickbait as false news," in *Proceedings of the ACM Workshop on Multimodal Deception Detection*, 2015.

[34] M. M. U. Rony, N. Hassan, and M. Yousuf, "Diving Deep into Clickbaits: Who Use Them to What Extents in Which Topics with What Effects?" *arXiv preprint arXiv:1703.09400*, 2017.

[35] F. Ribeiro, L. Henrique, F. Benevenuto, A. Chakraborty, J. Kulshrestha, M. Babaei, and K. Gummadi, "Media bias monitor: Quantifying biases of social media news outlets at large-scale," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, 2018.

[36] A. Chakraborty, S. Ghosh, N. Ganguly, and K. P. Gummadi, "Editorial versus audience gatekeeping: analyzing news selection and consumption dynamics in online news media," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 4, pp. 680–691, 2019.

[37] S. Kumar, R. West, and J. Leskovec, "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes," in *International World Wide Web Conference (WWW)*, 2016.

[38] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2012, pp. 171–175.

[39] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104*, 2017.

[40] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *International Conference on Knowledge Discovery & Data Mining (KDD)*, 2016.

[41] P. K. Verma, P. Agrawal, I. Amorim, and R. Prodan, "Welfake: Word embedding over linguistic features for fake news detection," *IEEE Transactions on Computational Social Systems*, 2021.

[42] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[43] "automatically detecting image–text mismatch on instagram with deep learning."

[44] C. Jin, P. Netrapalli, and M. I. Jordan, "Accelerated gradient descent escapes saddle points faster than gradient descent," in *Conference On Learning Theory*. PMLR, 2018, pp. 1042–1085.

[45] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[46] S. Castelo, T. Almeida, A. Elghafari, A. Santos, K. Pham, E. Nakamura, and J. Freire, "A topic-agnostic approach for identifying fake news pages," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 975–980.

[47] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st international conference on data engineering*. IEEE, 2015, pp. 651–662.

[48] T. Bian, X. Xiao, T. Xu, P. Zhao, W. Huang, Y. Rong, and J. Huang, "Rumor detection on social media with bi-directional graph convolutional networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 01, 2020, pp. 549–556.

[49] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks." Association for Computational Linguistics, 2018.

[50] G. Shrivastava, P. Kumar, R. P. Ojha, P. K. Srivastava, S. Mohan, and G. Srivastava, "Defensive modeling of fake news through online social networks," *IEEE Transactions on Computational Social Systems*, vol. 7, no. 5, pp. 1159–1167, 2020.

[51] J. Huang and Q. Su, "A rumor spreading model based on user browsing behavior analysis in microblog," in *2013 10th International Conference on Service Systems and Service Management*. IEEE, 2013, pp. 170–173.

[52] S. Dong, Y.-B. Deng, and Y.-C. Huang, "Seir model of rumor spreading in online social network with varying total population size," *Communications in Theoretical Physics*, vol. 68, no. 4, 2017.

[53] Y. Yao, X. Xiao, C. Zhang, C. Dou, and S. Xia, "Stability analysis of an sdilr model based on rumor recurrence on social media," *Physica A: Statistical Mechanics and its Applications*, vol. 535, p. 122236, 2019.

[54] K. Shu, D. Mahudeswaran, S. Wang, and H. Liu, "Hierarchical propagation networks for fake news detection: Investigation and exploitation," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 2020, pp. 626–637.

[55] J. Zhang, L. Cui, Y. Fu, and F. B. Gouza, "Fake news detection with deep diffusive network model," *arXiv preprint arXiv:1805.08751*, 2018.

[56] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD Explorations Newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[57] N. Sitaula, C. K. Mohan, J. Grygiel, X. Zhou, and R. Zafarani, "Credibility-based fake news detection," in *Disinformation, Misinformation, and Fake News in Social Media*.

[58] C. Cai, L. Li, and D. Zeng, "Detecting social bots by jointly modeling deep behavior and content information," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1995–1998.

[59] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A new approach to bot detection: striking the balance between precision and recall," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 533–540.

[60] C. Newton, "The verge interview with facebook and instagram," https://www.theverge.com/2019/10/21/20925204/facebook-2020-election-interference-prevention-tools-policy-false-misinformation, 2019.

[61] Y. roth and N. Pickles, "Twitter product blog on "updating our approach to misleading information"," https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html, 20120.

[62] D. Funke and A. Mantzarlis, "We asked 19 fact-checkers what they think of their partnership with facebook. here's what they told us." https://www.poynter.org/fact-checking/2018/we-asked-19-fact-checkers-what-they-think-of-their-partnership-with-facebook-heres-what-they-told-us/, 2018.

[63] A. Matthews, "How does twitter's tweet labeling work?" https://www.dw.com/en/how-does-twitters-tweet-labeling-work/a-53622684, 2019.

[64] E. Peer, J. Vosgerau, and A. Acquisti, "Reputation as a sufficient condition for data quality on amazon mechanical turk," *Behavior research methods*, vol. 46, no. 4, pp. 1023–1031, 2014.

[65] AAPOR, "Research synthesis: Aapor report on online panels," *Public Opinion Quarterly*, vol. 74, no. 4, pp. 711–781, 2010.

[66] E. M. Redmiles, Y. Acar, S. Fahl, and M. L. Mazurek, "A summary of survey methodology best practices for security and privacy researchers," Tech. Rep., 2017.

[67] J. A. Krosnick and L. R. Fabrigar, "Designing rating scales for effective measurement in surveys," *Survey measurement and process quality*, pp. 141–164, 1997.

[68] J. A. Krosnick, S. Narayan, and W. R. Smith, "Satisficing in surveys: Initial evidence," *New directions for evaluation*, vol. 1996, no. 70, pp. 29–44, 1996.

[69] S. Chopra, S. Jain, and J. M. Sholar, "Towards Automatic Identification of Fake News: Headline-Article Stance Detection with LSTM Attention Models," 2017.

[70] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "On the Benefit of Combining Neural, Statistical and External Features for Fake News Identification," *arXiv preprint arXiv:1712.03935*, 2017.

[71] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, vol. 10, no. 6, 2015.

[72] J. Kim, B. Tabibian, A. Oh, B. Schölkopf, and M. Gomez-Rodriguez, "Leveraging the crowd to detect and reduce the spread of fake news and misinformation," *arXiv preprint arXiv:1711.09918*, 2017.

Mahmoudreza Babaei is a postdoctoral researcher in the Max Planck Institute for Human-Development in the center of the human and machine group (CHM). He received his Ph.D. in Computer Science from the Max Planck Institute for Software Systems (MPI-SWS) in the Social Computing Research group with Prof. Krishna P. Gummadi. His research interests include the Social Networks, Data Mining, Graph Theory, and Natural Language process.

Juhi Kulshrestha is an Assistant Professor for Computational Social Science in the Department of Politics and Public Administration at the University of Konstanz. Prior to joining the University of Konstanz, she was a postdoctoral researcher in the Computational Social Science department at GESIS - Leibniz Institute for the Social Sciences. She received her Ph.D. in Computer Science from the Max Planck Institute for Software Systems (MPI-SWS) in the Social Computing Research group with Prof. Krishna P. Gummadi. Her research focuses on studying how users consume news and information on the Web and in evaluating the role played by automated retrieval algorithms, like search and recommendation systems, in shaping users' online information diets.

Abhijnan Chakraborty is an Assistant Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology (IIT) Delhi. Prior to joining IIT Delhi, he was a post-doctoral researcher at the Max Planck Institute for Software Systems (MPI-SWS). He received the Ph.D. degree in Computer Science from the Indian Institute of Technology (IIT) Kharagpur where he was jointly advised by Prof. Niloy Ganguly, IIT Kharagpur and Prof. Krishna Gummadi, MPI-SWS. Before joining PhD, he has worked at Microsoft Research India in Bangalore for two years. His research focuses on Social Computing, Information Retrieval and Fairness in Machine Learning.

Elissa M. Redmiles Elissa M. Redmiles is a faculty member and research group leader of the Safety & Society group at the Max Planck Institute for Software Systems. She has additionally served as a consultant and researcher at multiple institutions, including Microsoft Research, Facebook, the World Bank, the Center for Democracy and Technology, and the University of Zurich. Dr. Redmiles uses computational, economic, and social science methods to understand users' security, privacy, and online safety-related decision-making processes. Her work has been featured in popular press publications such as the New York Times, Scientific American, Wired, and CNET and has been recognized with multiple Distinguished Paper Awards at USENIX Security and research awards including a Facebook Research Award and the John Karat Usable Privacy and Security Research Award. Dr.Redmiles received her B.S. (Cum Laude), M.S., and Ph.D. in Computer Science from the University of Maryland.

Meeyoung Cha is an associate professor at the Korea Advanced Institute of Science and Technology (KAIST) in South Korea. Her research is on data science with an emphasis on modeling socially relevant information propagation processes. Her work on misinformation, poverty mapping, fraud detection, and long-tail content has gained more than 16,000 citations. Meeyoung has worked at Facebook's Data Science Team as a Visiting Professor and is a recipient of the Korean Young Information Scientist Award and AAAI ICWSM Test of Time Award. She is currently jointly affiliated as a Chief Investigator at the Institute for Basic Science (IBS) in Korea.

Krishna P. Gummadi Krishna Gummadi is a Scientific Director and Head of the Networked Systems research group at the Max Planck Institute for Software Systems (MPI-SWS) in Germany. He also holds an honorary professorship at the University of Saarland. He received his Ph.D. (2005) and B.Tech. (2000) degrees in Computer Science and Engineering from the University of Washington and the Indian Institute of Technology, Madras, respectively.

Krishna's research interests are in the measurement, analysis, design, and evaluation of complex Internet-scale systems. His current projects focus on understanding and building social computing systems. Specifically, they tackle the challenges associated with (i) assessing the credibility of information shared by anonymous online crowds, (ii) understanding and controlling privacy risks for users sharing data on online forums, (iii) understanding, predicting and influencing human behaviors on social media sites (e.g., viral information diffusion), and (iv) enhancing fairness and transparency of machine (data-driven) decision making in social computing systems. He received an ERC Advanced Grant in 2017 to investigate "Foundations for Fair Social Computing" (No. 789373).