Measuring the Misinformation Ecosystem: Audiences, Platforms, and Storytellers

A Proposal Presented

by

Shan Jiang

to

Khoury College of Computer Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Computer Science

Northeastern University

Boston, Massachusetts

October 2020

Abstract of the Proposal

Measuring the Misinformation Ecosystem:

Audiences, Platforms, and Storytellers

by

Shan Jiang

Doctor of Philosophy in Computer Science

Northeastern University, October 2020

Dr. Christo Wilson, Advisor

Misinformation, broadly defined as any false or inaccurate information, has been proliferating on social media. This proliferation has been raising increasing societal concerns about its potential consequences, e.g., polarizing the public and eroding trust in institutions. Existing surveys and experiments across disciplinary have investigated the misinformation problem from multiple perspectives, ranging from the socio-psychological foundations of audiences' susceptibility to algorithmic solutions aiding platforms' intervention on the spread of misinformation. Yet, a large-scale empirical study is still in need to comprehensively understand how different players behave and interact in the misinformation ecosystem.

To this end, the goal of this thesis is to study the misinformation ecosystem by measuring the behaviors of its three key players: audiences, platforms, and storytellers.

Audiences receive and respond to misinformation, and therefore their behaviors are potentially influenced by such falsehood or inaccuracies. The first part of the thesis investigates if and how audiences respond differently under misinformation. This part starts with an unsupervised exploration of user comments to misinformation posts on social media, where I observe significantly distinctive linguistic patterns when audiences comment on fabricated stories than truthful ones, e.g., increased signals suggesting their awareness of misinformation and extensive usage of angry emojis and swear words. In light of this exploration, I then refocus on measuring to what extend audiences disbelieve or believe in these stories. Applying supervised classifiers trained to identify (dis)beliefs, I estimate 12%/15% of audiences express disbelief, and 26%/20% of them express belief in true/false information.

Platforms play an essential role in how misinformation reaches its audiences. The second part of the thesis examines a specific practice of platforms' operations - content moderation, the AI-human

hybrid process of removing toxic content to maintain community standards. Using YouTube as a lens, this part investigates how misinformation and partisanship of a video interact with its comment moderation practice. I observe that, though not disclosed, videos containing verifiably false content are moderated more heavily for their comments, especially when the comments are posted after a fact-check. Additionally, I find no evidence to support allegations of political bias in this practice, when justifiable factors (e.g., hate speech) are controlled.

Storytellers generate misinformation from skewed facts and fabricated stories and then release them onto platforms. The third part of the thesis is proposed to measure the strategies of storytellers. I plan to extract phrases indicating how information is manipulated (e.g., digitally synthesized) from fact-checking articles, and structurize these phrases to systematically understand the story-making strategies of misinformation.

Altogether, my work presents an overview of the misinformation ecosystem to date, as well as methodologies and tools for the measurement. The empirical findings in the thesis are derived from computational approaches on observational data, therefore are reproducible from released repositories and applicable to future research. Ultimately, I hope that my research helps the public to understand misinformation and regain trust in authentic content online.

Chapter 1

Introduction

Misinformation is broadly defined as any false or inaccurate information. It takes many forms, ranging from unintentional poor journalism [81] to deliberate hoaxes [40, 41], propaganda [6, 46, 64, 76], disinformation [39, 76], and recently (and controversially) "fake news" [12, 81].

The online information ecosystem was and remains ground-zero where misinformation proliferates. During the 2016 US presidential election cycle, researchers estimated that "fake news" accounted for 6% of all news consumption [27], and 44% of Americans age 18 or older visited at least one untrustworthy website [28]. To date, misinformation has been documented across the globe, e.g., in Africa [82], Asia [38], and Europe [19]. As a countermeasure from online platforms, Facebook and Twitter have banned hundreds of pages and tens of thousands of accounts, respectively, linked to the Russian Internet Research Agency for generating and promoting misinformation [61, 72]. Yet, Misinformation continues to be posted on social media by politicians, partisan pundits, and even ordinary users [78].

The proliferation of misinformation has been raising increasing societal concerns about its potential consequences. In the political context, fabricated stories and partisan opinions may polarize the public [42], alter voters' perceptions about candidates [1, 15], and erode trust in institutions [9], therefore posing a threat to the democracy [31, 51].

Existing surveys and experiments across disciplinary have investigated the misinformation problem from multiple perspectives, ranging from the socio-psychological foundations of audiences' susceptibility [22, 54, 67, 80] to algorithmic solutions aiding platforms' intervention on the spread of misinformation [18, 63, 64, 73, 76, 77, 79]. Yet, a large-scaled empirical study is still in need to comprehensively understand how different players behave and interact in the misinformation ecosystem.

In this thesis, I aim to study the misinformation ecosystem by measuring the behaviors of its three key players:

- *Audiences*, who receive and respond to misinformation.
- *Platforms*, through which misinformation reaches its audiences.
- *Storytellers*, who generate misinformation.

In particular, I approach this study with computational methods on observational data, and publicly release corresponding datasets and code repositories to make results reproducible. These resources can be found at: `https://misinfo.shanjiang.me`.

1.1 Audiences' Response

Audiences receive and respond to misinformation, and therefore their behaviors are potentially influenced by this falsehood or inaccuracies. The first part of the thesis starts by exploring if and how this misinformation affects its audiences. Although scholars are still in debate of whether misinformation impacted the outcome of the 2016 US presidential election in whole [1, 28], exposure to misinformation may still harm audiences by promoting partisanship, reducing trust in civic institutions, and discouraging reasoned conversation [5, 21]. Research suggests that audiences are indeed vulnerable to misinformation because of psychological and sociological predispositions [22, 54,67,80]. Furthermore, misinformation often uses inflammatory and sensational language [64,76,77] that can alter audiences' emotions, which are a core component of how they perceive their political world [47], and can sometimes affect their perceived bias of information [83].

As a means to combat misinformation, journalists conduct research with evidence and logical reasoning to determine the veracity and correctness of factual claims made in public, and publish fact-checking articles (or fact-checks) on their news outlets, e.g., a tweet posted by Donald Trump claiming that Barack Obama was born in Kenya was later fact-checked by both Snopes and PolitiFact and found to be false [17, 49]. These fact-checks are later utilized in various ways by social media platforms, e.g., Facebook and Google have both deployed systems that integrate fact-checking services [10, 26]. Additionally, social media users may post links to facts as a way to independently debunk misinformation. These facts can originate from different sources, ranging from first-hand experiences to scientific studies (including fact-checks).

However, this reliance on fact-checking raises a parallel question of whether and how people respond to fact-checking itself. Some studies have found that fact-checking has corrective effects on audiences' beliefs [20, 29, 62, 84], while others found that it has minimal impact [44, 57] and sometimes even "backfires" on its audience [57–59]. In fact, the work of Snopes and PolitiFact has itself become politicized by those who view their work as biased, and this has led to attempts to discredit fact-checks [52, 65, 69].

To explore audiences' response to misinformation and fact-checks, I look at linguistic signals in user comments on social media in the presence of misinformation and fact-checks. I collect a dataset of 5,303 social media posts with 2,614,374 user comments from Facebook, Twitter, and YouTube, and associate these posts to fact-checks from Snopes and PolitiFact to obtain veracity rulings (i.e., from true to false). Then, I build a emotional and topical lexicon, named *ComLex*, using a hybrid method of natural language processing (NLP) techniques and human validation. This lexicon is later used to analyze data and test hypotheses. Overall, this part investigates the following research questions (RQs):

- **RQ1.1**, *do linguistic signals in user comments vary in the presence of misinformation?* As post veracity decreases, social media users express more misinformation-awareness signals, as well as different emotional and topical signals, e.g., extensive use of emojis and swear words, less discussion of concrete topics, and decreased objectivity.

- **RQ1.2**, *do linguistic signals in user comments vary after a post is fact-checked?* There are signals indicating positive effects after fact-checking, such as more misinformation-awareness and less doubtful signals. However, there are also signals indicating potential "backfire" effects, such as increased swear word usage.

This exploration suggests that audiences do respond differently, as expressed in their comments, to misinformation. In light of this exploration, I then refocus on measuring a specific signal in audiences' response: *belief*.

Belief is an important signal of audiences' response, as the consequences of misinformation are mostly framed under the audiences' susceptibility to misinformation, i.e., the public is unable, or disinclined, to distinguish truth from fiction. This narrative naturally needs further investigation and quantification. Recent surveys from the Reuters Institute and Pew Research Center reported that audiences are indeed aware of the misinformation problem, and (dis)believe certain information sources (e.g., news outlets, politicians) more than others [2, 55]. However, these studies are small-

scale in nature, and thus unable to quantitatively measure to what extent do audiences (dis)believe in (mis)information.

Complementary to these surveys, I propose an observational approach as an alternate lens through which to interrogate the audiences' (dis)belief in (mis)information, which leverages user comments (collected above) as a proxy for assessing audiences' response. The language used in comments in response to claims can express signals of the users' (dis)belief, therefore, if modeled properly, these comments can be used to measure the prevalence of expressed (dis)belief at scale.

To model (dis)belief expressed in user comments, I start by collecting a small sample of tweets that comment on fact-checked claims, and then manually annotate each tweet with disbelief and belief labels. Using this dataset, I experiment with several NLP techniques. I first conduct an exploratory analysis using lexicon-based methods, which reveals differences in word usage (e.g., falsehood awareness signals, positive and negative emotions) in tweets expressing (dis)belief verses others. Next, I experiment with classification models, including linear models with lexicon-derived features, as well as state-of-the-art neural transfer-learning models (e.g., BERT [13], XLNet [85], and RoBERTa [45]). Then, I develop a domain-specific thresholding strategy for classifiers to make unbiased predictions compared to human experts. Under chosen thresholds, the best-performing classifier achieves macro-F_1 scores around 0.86 for predicting disbelief and 0.80 for belief. Next, I aim to measure expressed (dis)belief at scale by applying the trained classifier. I run the classifier on the large, unlabeled dataset collected above, and analyze the estimated prevalence of expressed (dis)belief. Overall, this part investigates the following RQs:

- **RQ1.3**, *do audiences believe in misinformation, and if so, to what extent?* For true/mixed/false claims on social media, 12%/14%/15% of comments express disbelief and 26%/21%/20% of comments express belief, suggesting (optimistically) increased disbelief and decreased belief as information veracity decrease, yet (pessimistically) considerable suspicions on truthful information.

- **RQ1.4**, *does the prevalence of expressed (dis)belief in misinformation vary over time?* There is an extremely slight time effect of misinformation-awareness, where disbelief increases 0.001% and belief decreases 0.002% per day after a false claim is published.

- **RQ1.5**, *does the prevalence of expressed (dis)belief in misinformation vary after fact-checking?* Controlling for the time effect, disbelief increases 5% and belief decreases 3.4% after claims

are fact-checked, suggesting a positive effect of fact-checks on altering the prevalence of (dis)belief.

1.2 Platforms' Moderation

Platforms play an essential role in how misinformation reaches its audience. The second part of the thesis examines the behaviors of platforms. Besides the misinformation problem [11], social media platforms have also been subject to heightened levels of controversy and scrutiny for other issues, e.g., violent hate speech [60] and partisanship [1].

The solution promulgated by social media platforms for these problems is an increase in content moderation. In terms of mechanisms, the major platforms have committed to hiring tens of thousands of new human moderators [43], investing in more artificial intelligence to filter content [23], and partnering with fact-checking organizations to identify misinformation [25]. In terms of policy, the platforms are updating their community guidelines with expanded definitions of what they believe constitutes hate speech, harassment, etc [16, 74, 86].

Although content moderation is not specifically designed to filter out misinformation, Chapter **??** suggests that misinformation does alter audiences' comments and increasing their usages of swear words and others that might violate community guidelines, and misinformation content could draw more attention from moderators due to this increased violation rate. Therefore, it is worth investigating how misinformation and content moderation interact in practice.

Another issue raised by this increased reliance on content moderation is a backlash from ideological conservatives, who claim that social media platforms are biased against them and are censoring their views [37, 75]. Two US House Committees have held hearings on content moderation practices to "specifically look at concerns regarding a lack of transparency and potential bias in the filtering practices of social media companies (Facebook, Twitter and YouTube)" [7, 14]. In June 2019, the "Ending Support for Internet Censorship Act" was introduced into the US Senate to limit immunity granted by Section 230 of the Communications Decency Act to "encourage providers of interactive computer services to provide content moderation that is politically neutral" [30]. These concerns are driven by multiple factors, including anecdotal reports that: Facebook's Trending News team did not promote stories from conservative media outlets [56], Twitter "shadow banned" conservative users [53], fact-checking organizations are biased [65], and selective reporting by partisan news agencies [3].

However, there is no scientific evidence that social media platforms' content moderation practices exhibit systematic partisan bias [34, 70, 71]. On the contrary, there are many cases where ideologically liberal users were moderated, although these cases have received less attention in the media [48]. It is possible that moderation only appears to be biased because political valence is correlated with other factors that trigger moderation, such as bullying, calls to violence, or hate speech [24]. Further, there is evidence suggesting that users tend to overestimate bias in moderation decisions [71].

In this study, I use YouTube as a lens and aim to disentangle these issues by investigating how partisanship and misinformation in videos affect the likelihood of comment moderation. Specifically, I examine four hypotheses related to four attributes of YouTube videos and comments: the leaning of partisanship (i.e., left or right), the magnitude of partisanship (i.e., center or extreme), the veracity of the content (i.e., true or false), and whether a comment was posted before or after the video was fact-checked. For each variable, I start with the null hypotheses $\mathbf{H}_0$ that the variable has no effect on comment moderation, and then use two formal criteria (i.e., *independence* and separation [4]) to collect evidence on rejecting the null hypotheses.

To investigate these hypotheses, I refine the dataset collected above to 84,068 comments posted across 258 YouTube videos, and associate them to partisanship labels from existing research [66] and misinformation labels from Snopes or PolitiFact [36]. I first test for independence and find that all of the hypothesized variables significantly correlate with the likelihood of comment moderation. Although this seems to suggest a political bias against right-leaning content, I argue that such bias is misperceived as it ignores other confounding variables that are justified and potentially contribute to moderation decisions, such as social engagement (e.g., views and likes) [50] and the linguistics in comments (e.g., hate speech) [8, 71]. Therefore, I re-analyze our dataset using a causal propensity score model to test the separation hypotheses when potential confounds are controlled. Overall, this part investigates the following RQs:

- **RQ2.1**, *does the political leaning of a video affect the moderation decision of its comments?* No significant difference is found for comment moderation on left- and right-leaning videos.
- **RQ2.2**, *does the extremeness of a video affect the moderation decision of its comments?* Comments on videos from ideologically extreme channels are $\sim$50% more likely to be moderated than center channels.
- **RQ2.3**, *does the veracity of content in a video affect the moderation decision of its comments?* Comments on true videos are $\sim$60% less likely to be moderated than those on false videos.

- **RQ2.4**, *does the fact-check of a video affect the moderation decision of its comments?* Comments posted after a video is fact-checked are $\sim$20% more likely to be moderated than those posted before the fact-check.

I approach these hypotheses using an empirical method for auditing black-box decision-making processes [68] based on publicly available data on YouTube. Neither I, nor the critics, have access to YouTube's internal systems, data, or deliberations that underpin moderation decisions. Instead, I aim to highlight the difference in *perceived* bias when analyzing available data using correlational and causal models, and further, foster a healthier discussion of algorithmic and human bias in social media.

1.3 Storytellers' Strategies

Storytellers generate misinformation from skewed facts and fabricated stories and then release them onto platforms. In the third part of the thesis, I propose to measure the strategies of storytellers.

From storytellers' perspectives, misinformation can be generated in numerous ways, e.g., fabricating or manipulating content, making false context or connection, and mis-spreading satire or parody [81]. However, these strategies are hitherto theorized, and there is no empirical study measuring these strategies in real world.

To systematically study misinformation storytellers' strategies, I plan to extract phrases indicating how misinformation is generated from fact-checks, a specialized news format that tend to share certain common structured information, i.e., the claim, the claimant and the verdict [32]. When reasoning about the veracity of a claim, fact-checks often writes how a false claim is made, e.g., "this photo is digitally synthesized" or "the numbers do not match with official content". These short phrases summarize the storymaking strategies of such misinformation.

I plan to use *rationalized* NLP models to extract these phrases. Rationalized NLP models aim to make a prediction along with rationales of why the prediction is made. In this context, the phrases "this photo is digitally synthesized" and "the numbers do not match with official content" are rationales of a "false" verdict from a fact-check.

There are some collateral RQs I plan to investigate in this part. For example, as these phrase are extracted from NLP models, their semantic embeddings can also be exported. These embeddings can be used for hierarchical clustering to structurally understand the storymaking strategies of

misinformation. Additionally, these strategies can be mapped to the dataset collected above to measure the prevalence of different strategies.

1.4 Timeline of Milestones

The research questions introduced in § 1.1 have been investigated in two of my papers published at CSCW 2018 [36] and ICWSM 2020 [33].

The research questions introduced in § 1.2 have been investigated in two of my papers published at ICWSM 2019 [34] and AAAI 2020 [35].

The research questions proposed in § 1.3 are work in progress that is aiming to be submitted to ICWSM 2021 on Jan 15, 2021.

After the submission of § 1.3, I plan to integrate it into my thesis and defense in July 2021.

Bibliography

[1] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[2] J. Anderson and L. Rainie. The future of truth and misinformation online. *Pew Research Center*, 2017.

[3] K. Arceneaux, M. Johnson, and C. Murphy. Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics*, 74(1):174–186, 2012.

[4] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[5] N. Berman. The victims of fake news, 2017.

[6] E. L. Bernays. *Propaganda*. Ig publishing, 1928.

[7] M. Bickert, J. Downs, and N. Pickles. Facebook, google and twitter: Examining the content filtering practices of social media giants. House Judiciary Committee, 7 2018.

[8] E. Chandrasekharan, M. Samory, S. Jhaver, H. Charvat, A. Bruckman, C. Lampe, J. Eisenstein, and E. Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *PACM on HCI*, 2(CSCW):32, 2018.

[9] G. L. Ciampaglia, A. Mantzarlis, G. Maus, and F. Menczer. Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine*, 39(1), 2018.

[10] J. Constine. Facebook tries fighting fake news with publisher info button on links, 10 2017.

[11] J. Constine. Facebook reveals russian troll content, shuts down 135 ira accounts. Tech Crunch, 3 2018.

[12] N. A. Cooke. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The Library Quarterly*, 87(3):211–221, 2017.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 2019.

[14] J. Dorsey. Twitter: Transparency and accountability. House Energy and Commerce Committee, 9 2018.

[15] R. Epstein and R. E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *PNAS*, 112(33), Aug. 2015.

[16] Facebook. Community standards, 2018.

[17] R. Farley. Trump said obama's grandmother caught on tape saying she witnessed his birth in kenya, 7 2011.

[18] 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: Short Papers-Volume 2*, pages 171–175. Association for Computational Linguistics, 2012.

[19] R. Fletcher, A. Cornia, L. Graves, and R. K. Nielsen. Measuring the reach of ?fake news? and online disinformation in europe. *Reuters institute factsheet*, 2018.

[20] K. Fridkin, P. J. Kenney, and A. Wintersieck. Liar, liar, pants on fire: How fact-checking influences citizens? reactions to negative advertising. *Political Communication*, 32(1):127–151, 2015.

[21] U. Friedman. The real-world consequences of "fake news", 12 2017.

[22] M. Gentzkow, J. M. Shapiro, and D. F. Stone. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1, pages 623–645. Elsevier, 2015.

[23] S. Gibbs. Google says ai better than humans at scrubbing extremist youtube content. The Guardian, 8 2017.

[24] T. Gillespie. There's a reason that misleading claims of bias in search and social media enjoy such traction. Medium, 8 2018.

[25] A. Glaser. Youtube is adding fact-check links for videos on topics that inspire conspiracy theories. Slate, 8 2018.

[26] Google. Google fact checks feature, 2018.

[27] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425), 2019.

[28] A. Guess, B. Nyhan, and J. Reifler. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 us presidential campaign. *European Research Council*, 2018.

[29] K. Haglin. The limitations of the backfire effect. *Research & Politics*, 4(3):2053168017716547, 2017.

[30] J. Hawley. Ending support for internet censorship act, 2019.

[31] J. L. Hochschild and K. L. Einstein. *Do facts matter?: Information and misinformation in American politics*, volume 13. University of Oklahoma Press, 2015.

[32] S. Jiang, S. Baumgartner, A. Ittycheriah, and C. Yu. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proc. of WWW*, 2020.

[33] S. Jiang, M. Metzger, A. Flanagin, and C. Wilson. Modeling and measuring expressed (dis)belief in (mis)information. In *Proc. of ICWSM*, 2020.

[34] S. Jiang, R. E. Robertson, and C. Wilson. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proc. of ICWSM*, 2019.

[35] S. Jiang, R. E. Robertson, and C. Wilson. Reasoning about political bias in content moderation. In *Proc. of AAAI*, 2020.

[36] S. Jiang and C. Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *PACM on HCI*, 2(CSCW), November 2018.

[37] B. Kamisar. Conservatives cry foul over controversial group's role in youtube moderation. The Hill, 3 2018.

[38] K. Kaur, S. Nair, Y. Kwok, M. Kajimoto, Y. T. Chua, M. Labiste, C. Soon, H. Jo, L. Lin, T. T. Le, et al. Information disorder in asia and the pacific: Overview of misinformation ecosystem in

australia, india, indonesia, japan, the philippines, singapore, south korea, taiwan, and vietnam. *SSRN*, 2018.

[39] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, pages 591–602. International World Wide Web Conferences Steering Committee, 2016.

[40] D. Lazer, M. Baum, N. Grinberg, L. Friedland, K. Joseph, W. Hobbs, and C. Mattsson. Combating fake news: An agenda for research and action. *Harvard Kennedy School, Shorenstein Center on Media, Politics and Public Policy*, 2, 2017.

[41] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[42] M. S. Levendusky. Why do partisan media polarize viewers? *American Journal of Political Science*, 57(3):611–623, 2013.

[43] S. Levin. Google to hire thousands of moderators after outcry over youtube abuse videos. The Guardian, 12 2017.

[44] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131, 2012.

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019.

[46] C. Lumezanu, N. Feamster, and H. Klein. # bias: Measuring the tweeting behavior of propagandists. In *Sixth International AAAI Conference on Weblogs and Social Media*, 2012.

[47] G. E. Marcus. The sentimental citizen: Emotion in democratic politics. *Perspectives on Politics*, 2002.

[48] M. Masnick. Internet content moderation isn't politically biased, it's just impossible to do well at scale. Techdirt, 8 2018.

[49] D. Mikkelson. Barack obama birth certificate: Is barack obama's birth certificate a forgery?, 8 2011.

[50] M. Mondal, L. A. Silva, and F. Benevenuto. A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, pages 85–94. ACM, 2017.

[51] S. Morgan. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, 3(1), 2018.

[52] NewsBusters. Don't believe the liberal "fact-checkers"!, 2018.

[53] C. Newton. Why twitter should ignore the phony outrage over "shadow banning". The Verge, 7 2018.

[54] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175, 1998.

[55] R. K. Nielsen and L. Graves. "news you don't believe": Audience perspectives on fake news. *Reuters Institute*, 2017.

[56] M. Nunez. Former facebook workers: We routinely suppressed conservative news. Gizmodo, 5 2016.

[57] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

[58] B. Nyhan and J. Reifler. Does correcting myths about the flu vaccine work? an experimental evaluation of the effects of corrective information. *Vaccine*, 33(3):459–464, 2015.

[59] B. Nyhan, J. Reifler, and P. A. Ubel. The hazards of correcting myths about health care reform. *Medical care*, 51(2):127–132, 2013.

[60] A. Olteanu, C. Castillo, J. Boy, and K. R. Varshney. The effect of extremist violence on hateful speech online. In *Proc. of ICWSM*, 2018.

[61] B. Popken. Twitter deleted 200,000 russian troll tweets, 2 2018.

[62] E. Porter, T. J. Wood, and D. Kirby. Sex trafficking, russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news. *Journal of Experimental Political Science*, pages 1–6, 2018.

[63] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.

[64] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, 2017.

[65] V. Richardson. Conservative project seeks to fact-check the fact-checkers accused of liberal bias, 3 2018.

[66] R. E. Robertson, D. Lazer, and C. Wilson. Auditing the personalization and composition of politically-related search engine results pages. In *Proc. of The Web Conference (WWW)*, 2018.

[67] R. J. Robinson, D. Keltner, A. Ward, and L. Ross. Actual versus assumed differences in construal: "naive realism" in intergroup perception and conflict. *Journal of Personality and Social Psychology*, 68(3):404, 1995.

[68] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry*, pages 1–23, 2014.

[69] M. Shapiro. Running the data on politifact shows bias against conservatives, 12 2016.

[70] Q. Shen and C. Rose. The discourse of online content moderation: Investigating polarized user responses to changes in reddit?s quarantine policy. In *Proc. of ALW3 ACL*, 2019.

[71] Q. Shen, M. Yoder, Y. Jo, and C. Rose. Perceptions of censorship and moderation bias in political debate forums. In *Proc. of ICWSM*, 2018.

[72] S. Sheth. Facebook takes down over 200 accounts and pages run by the ira, a notorious russian troll farm, 4 2018.

[73] 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*, 19(1):22–36, 2017.

[74] Twitter. Rules and policies, 2018.

[75] N. Usher. How republicans trick facebook and twitter with claims of bias. The Washington Post, 8 2018.

[76] S. Volkova and J. Y. Jang. Misleading or falsification? inferring deceptive strategies and types in online news and social media. In *The 2018 Web Conference Companion (WWW 2018 Companion)*, Lyon, France, 4 2018.

[77] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 647–653, 2017.

[78] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[79] W. Y. Wang. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 422–426, 2017.

[80] A. Ward, L. Ross, E. Reed, E. Turiel, and T. Brown. Naïve realism in everyday life: Implications for social conflict and misunderstanding. *Values and knowledge*, 1997.

[81] C. Wardle. Fake news. it's complicated. *First Draft News*, 2017.

[82] H. Wasserman and D. Madrid-Morales. An exploratory study of ?fake news? and media trust in kenya, nigeria and south africa. *African Journalism Studies*, 40(1), 2019.

[83] B. E. Weeks. Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication*, 65(4):699–719, 2015.

[84] T. Wood and E. Porter. The elusive backfire effect: mass attitudes? steadfast factual adherence. *Political Behavior*, pages 1–29, 2016.

[85] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. of NeurIPS*, 2019.

[86] YouTube. Community guidelines, 2018.