Measuring the Misinformation Ecosystem: Audiences, Platforms, and Storytellers

A Dissertation 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

To my family.

Contents

List of Figures

List of Tables

Acknowledgments

Abstract of the Dissertation

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-scaled 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 to 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 [228] to deliberate hoaxes [114, 115], propaganda [19, 129, 184, 222], disinformation [109, 222], and recently (and controversially) "fake news" [38, 228].

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 [74], and 44% of Americans age 18 or older visited at least one untrustworthy website [75]. To date, misinformation has been documented across the globe, e.g., in Africa [229], Asia [100], and Europe [60]. 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 [175, 201]. Yet, Misinformation continues to be posted on social media by politicians, partisan pundits, and even ordinary users [224].

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 [117], alter voters' perceptions about candidates [1, 45], and erode trust in institutions [34], therefore posing a threat to the democracy [84, 148].

Existing surveys and experiments across disciplinary have investigated the misinformation problem from multiple perspectives, ranging from the socio-psychological foundations of audiences' susceptibility [67, 156, 193, 227] to algorithmic solutions aiding platforms' intervention on the spread of misinformation [56, 179, 184, 203, 222, 223, 225]. 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, 75], exposure to misinformation may still harm audiences by promoting partisanship, reducing trust in civic institutions, and discouraging reasoned conversation [18, 63]. Research suggests that audiences are indeed vulnerable to misinformation because of psychological and sociological predispositions [67, 156, 193, 227]. Furthermore, misinformation often uses inflammatory and sensational language [184, 222, 223] that can alter audiences' emotions, which are a core component of how they perceive their political world [133], and can sometimes affect their perceived bias of information [230].

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 [51, 139]. 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 [36, 73]. 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 [62, 77, 176, 231], while others found that it has minimal impact [119, 159] and sometimes even "backfires" on its audience [159–161]. 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 [153, 188, 197].

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 [6, 157]. 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 [42], XLNet [236], and RoBERTa [126]). 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 [37], social media platforms have also been subject to heightened levels of controversy and scrutiny for other issues, e.g., violent hate speech [162] 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 [118], investing in more artificial intelligence to filter content [68], and partnering with fact-checking organizations to identify misinformation [71]. In terms of policy, the platforms are updating their community guidelines with expanded definitions of what they believe constitutes hate speech, harassment, etc [48, 217, 237].

Although content moderation is not specifically designed to filter out misinformation, Chapter 3 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 [99, 219]. 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)" [20, 44]. 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" [82]. These concerns are driven by multiple factors, including anecdotal reports that: Facebook's Trending News team did not promote stories from conservative media outlets [158], Twitter "shadow banned" conservative users [154], fact-checking organizations are biased [188], and selective reporting by partisan news agencies [7].

However, there is no scientific evidence that social media platforms' content moderation practices exhibit systematic partisan bias [93, 199, 200]. On the contrary, there are many cases where ideologically liberal users were moderated, although these cases have received less attention in the media [136]. 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 [70]. Further, there is evidence suggesting that users tend to overestimate bias in moderation decisions [200].

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 [15]) 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 [191] and misinformation labels from Snopes or PolitiFact [94]. 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) [147] and the linguistics in comments (e.g., hate speech) [29, 200]. 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 ∼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 ∼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 [195] 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 [228]. 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 [91]. 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 Outline

The remainder of the thesis is organized as follows: § 2 introduces the background of the thesis and positions it around related areas, § 3 measures audiences' response to misinformation and answers **RQ1.1**-**RQ1.5**, and § 4 investigates platforms' moderation practice on misinformation and answers **RQ2.1**-**RQ2.4**.

Chapter 2

Background

This Chapter introduces the background of the thesis and positions it around related areas: misinformation and its consequence (§ 2.1), content moderation and its controversy (§ 2.2), and, in terms of methodology, NLP for social science (§ 2.3).

2.1 Misinformation and Its Consequences

The misinformation problem is in nature interdisciplinary, therefore drawing researchers from different areas, e.g., computer, social, and political science. In this section, I introduce the background of misinformation and its consequences.

2.1.1 Foundations of Misinformation

As "misinformation" (broadly construed) takes many forms, ranging from unintentional poor journalism to deliberate hoaxes and propaganda [19, 129, 184, 222], there is currently no agreement upon terminology across communities for such false and inaccurate information. In general, there are two criteria that separate existing terminology: *veracity* and *intentionality* [203]. Some scholars prefer to use "misinformation" to broadly refer to all false and inaccurate information regardless of intent [41, 75, 86, 119, 230], while others prefer the more modern (but polarizing) term "fake news" [1, 114, 115, 203]. Other scholars restrict "misinformation" to unintentional inaccuracies, and use "disinformation" for deliberate deception [109, 228]. "Propaganda" typically refers to intentional and strictly political information [19, 129], although its veracity may vary from untruths to true but manipulative information.

In this thesis, I adopt the term "misinformation" as it is inclusive and not heavily politicized.

The examination of misinformation has a long history of research. The psychological foundations are rooted in people's individual vulnerabilities. One theory that explains susceptibility to misinformation is *naïve realism*, where people tend to believe that their perceptions of reality are accurate, and views that disagree with their perceptions are uninformed, irrational, and biased [65, 193, 227]. Another theory called *confirmation bias* shows that people prefer to accept information that confirms their existing beliefs [156]. Sociological theories including *social identity theory* [25, 212] and *normative influence theory* [9] also suggest that social acceptance and affirmation are essential for people's identity and self-esteem. This causes people to choose "socially safe" options when responding to and spreading information by following the norms of their established ideological groups, regardless of the information veracity. Finally, economic theory posits that "fake news" occurs when a news publishers values short-term expansion of its customer base more than its long-term reputation, coupled with news consumers that prefer information that confirms their preexisting false beliefs [67].

Audiences' vulnerability to misinformation affects their behavior and communication. For example, in-lab experiments have shown that exposure to biased information online [190] may significantly impact voting behavior [45, 46], while naïve information sharing may promote homophilous "echo chambers" of information [41, 122, 123].

In contrast to the above theories, there is a growing body of empirical research on people's ability to identify misinformation. Surveys that have asked people how much trust they place in different news media outlets have found that people do perceive specific outlets as biased (e.g., InfoWars) and thus do not trust news from these sources [102, 143].

Another line of work measured the spread and impact of misinformation, finding that "fake news" spread faster than truthful news [224], and that a large fraction of "fake news" are spread by "bots" [57, 196]. Misinformation is especially (and alarmingly) easy to be spread during crises, because people attempt to complete partial information using their natural sense-making processes [86], although such misinformation can sometimes be self-corrected by the crowd [8].

Early computational work is focused on the algorithmic model and detection of misinformation [203]. These studies are generally divided into two categories. The first category analyzes *text content* to assess veracity. Some researchers use the claims included in text to do automatic fact-checking by comparing the consistency and frequency of claims [14, 132], or by attempting to infer a claim from existing knowledge graphs [35, 81, 202, 226, 233]. Others note that fake or hyper-partisan news publishers usually have malicious intent to spread misinformation as widely as possible, which causes them to adopt a writing style that is inflammatory and sensational. Such

stylistic features can distinguish false from truthful news [56, 179, 184, 222, 223, 225].

Therefore, it is worth investigating whether the inflammatory content and sensational writing style that is sometimes characteristic of misinformation affects the emotional and topical signals that people express in their social media comments, as previous research has shown that linguistic signals, e.g., usage of emojis, can be used to infer people's actual emotional states [55, 101, 128, 189, 239].

The second category of detection algorithms leverage *social context* to predict misinformation, i.e., users' different behaviors when reacting to false or truthful news. These behaviors including different stances and discussed topics than the original threads [96, 182, 211], as well as different propagation network structures between fake and truthful news [76, 95]. Recent work has also proposed tools that actively solicit and analyze "flags" on misinformation from users [216]. Therefore, it is possible that the linguistic signals expressed in users comments can help to detect misinformation as well.

Taken together these related studies, I propose **RQ1.1**, *do linguistic signals in user comments vary in the presence of misinformation?* This RQ have substantial implications on the design of social computing systems. If user comments on misinformation significantly deviate from typical conversations (e.g., extensive usage of swear words), they could easily deteriorate into trolling [32], harassment [169], or hate speech [147]. Understanding and detecting the linguistic variants present in these comment threads may help when implementing intervention and moderation systems [69, 90].

2.1.2 Fact-Checking as an Intervention

Fact-checking is 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-checks on their news outlets.

There is a line of research focusing on the effects of fact-checking. Many in-lab experiments have examined the effects of fact-checking on human behaviors, but unfortunately they reveal drastically different behaviors in different contexts. A fact-check against a false rumor that the flu vaccine gave people the flu significantly reduced people's belief in the rumor, but also reduced some people's willingness to vaccinate because of side effects [160, 161]. However, later research failed to duplicate the results [77]. This phenomenon is called the "backfire" effect, where attempting to intervene against misinformation only entrenches the original, false belief further [159].

Even without the backfire effect, there are several experiments that found that fact-checking has limited corrective effects [119, 159]. However, others found that people are willing to accept fact-checking even when the information challenges their ideological commitments [62, 176, 231]. These

studies suggest that context is an important variable when examining the effect of fact-checking, as studies under different conditions often generate different results that cannot be generalized.

Major fact-checking organizations include Snopes [140], Politifact [198], and FactCheck.org [89]. These websites use facts and evidence to determine the veracity and correctness of factual claims in news articles, political speeches, social media posts, etc. In general, their verdicts have a very high degree of agreement with each other [4, 5]. However, the corrective effects of these websites has not been investigated in detail. Previous research has shown that fact-check articles posted on social media are likely to get more exposure when shared by a friend instead of strangers [78, 134], but that including biographical information about the author of the fact-check in the article itself reduces the effectiveness [65]. On online platforms, alert messages and tags that warn users to the presence of misinformation can help reduce the influence of this content [46, 173].

These studies suggest that context is an important variable when examining the effect of fact-checking, as studies under different conditions often generate different results that cannot be generalized. Furthermore, recent studies have proposed integrating fact-checking results [108] or bias warnings [46] into social computing systems.

Building on this line of work, I propose **RQ1.2**, *do linguistic signals in user comments vary after a post is fact-checked?* This RQ can shed light on recent discussion on whether and how to integrate fact-checks into social computing systems [108], e.g., Google search [91], YouTube [73], Facebook [36].

2.1.3 Belief and Disbelief in Misinformation

The consequences of misinformation are framed under the public's susceptibility to misinformation. This susceptibility is supported by existing psychological and sociological theories discussed in § 2.1.1: Naïve realism [227] and confirmation bias theory [156] from psychology suggested that people tend to believe in information that resonates with their pre-existing (yet potentially false) beliefs. Social identity [206] and normative influence theory [104] from sociology suggested that people tend to follow the norms of their established ideological groups when responding to information, and spread their beliefs in "socially safe" information, often regardless of its veracity.

On the empirical side, a report from the Pew Research Center provided evidence for these theories by conducting a survey about trust in news outlets across the ideological spectrum. It found a significant correlation between the self-reported trust and the ideological proximity between the audience and the news outlet, e.g., the liberal audience tended to trust the New York Times

while conservative audiences did not, and vice-versa for Fox News [143]. More recent reports from the Reuters Institute [157] and Pew Research Center [6] surveyed in more depth about the socio-psychological mechanisms behind (dis)belief and (mis)information, and reported that the public is indeed aware of the misinformation problem. Despite the valuable evidence offered, these qualitative and experimental studies are small-scale, and they required direct interactions with the participants, therefore potentially suffering from the Hawthorne Effect where participants modified their behaviors under their awareness of being surveyed [137].

Quantitative research on this topic is relatively limited. In the following § 3.3, I analyze social media comments in response to misinformation using an unsupervised approach, and showed that certain linguistic signals suggesting (dis)belief (e.g., "fake", "dumbest") were distributed differently in response to claims with differing veracity. In § 3.5.1, I further verify that these signals do indeed correlate with the likelihood to express (dis)belief, but they are insufficient predictors to judge if a comment expresses (dis)belief. Therefore, I propose to specifically measure belief and disbelief in misinformation, and ask **RQ1.3**, *do audiences believe in misinformation, and if so, to what extent?*

In light of existing studies that leverage the "wisdom of the crowd" for misinformation detection, as discussed in § 2.1.1, I hypothesize that audiences can gradually realize the truth after a claim is made and then lose trust in false claims over time, and propose the next **RQ1.4**, *does the prevalence of expressed (dis)belief in misinformation vary over time?*

Finally, continuing the discussion on the effect of fact-checking in § 2.1.1, I propose **RQ1.5**, *does the prevalence of expressed (dis)belief in misinformation vary after fact-checking?* This RQ measures the effect of fact-checking from the perspective of audiences' belief and disbelief.

2.2 Content Moderation and Its Controversy

Content moderation is an AI-human hybrid process of removing toxic content from social media to promote community health. In this section, I introduce the background of content moderation and its controversy.

2.2.1 Platforms and Community Guidelines

The content moderation practices of social media platforms are guided by their *community guidelines*, which explain the types of content they prohibit [48, 217, 237].

In the case of YouTube, it lists rules for: nudity or sexual content, harmful or dangerous content, hateful content, violent or graphic content, harassment and cyberbullying, etc [237]. Once content on YouTube (e.g., a video or comment) is judged to violate the guidelines, it is taken down, i.e., *moderated*. There are multiple reasons why a comment could be moderated on YouTube. A comment may be reviewed by patrolling YouTube moderators, or a comment may be flagged by YouTube users and then reviewed by the YouTube moderators [118]. Additionally, a comment may be removed by the corresponding video uploader, or by the commenter themselves [237]. Besides these human efforts, YouTube also uses algorithms that automatically flag and moderate inappropriate content [68]. In general, the mechanisms that lead to comment moderation are convoluted. Therefore, I view the internal YouTube system as a black-box, and focus on the moderation decision instead.

2.2.2 Effects of Content Moderation

Content moderation has been shown to have positive effects on social media platforms. A study that investigated Reddit's ban of the r/fatpeoplehate and r/CoonTown communities found that the ban expelled more "bad actors" than expected, and those who stayed posted much less hate speech than before the ban [28]. A study that interviewed users of Twitter's "blocklist" feature discussed how it can be used to prevent harassment [90].

However, content moderation systems have also raised concerns about bias and efficacy. Human moderators have been shown to bring their own biases into the content evaluation process [43] and automated moderation algorithms are prone to false positives and negatives [221]. These moderation strategies are also brittle: a study on Instagram found that users in pro-eating disorder communities invented variations of banned tags (e.g., "anorexie" instead of "anorexia") to circumvent lexicon-based moderation [27].

Researchers have also studied the community norms behind moderation from a linguistic perspective. A study on Reddit used 2.8M removed comments to identify macro-, meso-, and micro-norms across communities [29]. A study on the Big Issues Debate group of Ravelry found that comments expressing unpopular viewpoints were more likely to be moderated, but that this effect is negligible when compared to the total level of moderation [200]. These studies highlight the role of linguistics on the task of comment moderation, which sheds light on the importance of controlling for linguistics when investigating bias in moderation practices.

2.2.3 Bias of Human and Algorithms

Content moderation, and other online social systems of an opaque nature, has raised investigation on whether they exhibit bias against specific groups [195]. Studies have found gender and racial bias on hiring sites [31], freelance markets [79], ridesharing platforms [92], and online writing communities [54]. In the case of ideological groups, it has been reported that social media platforms such as Facebook are inferring users' ideologies to target them with political ads [205], while search engines may create "filter bubbles" that isolate users from ideologically opposing information [85, 121].

However, research on ideological bias in online contexts has sometimes led to surprising conclusions. Facebook researchers found that the partisan bias of content appearing in the Newsfeed was due more to homophily than algorithmic curation [13]. A study on Google Search also found that the partisan bias of search results was dependent largely on the input query rather than the self-reported ideology of the user [190].

As for content moderation, there have been several allegations that social media platforms are censoring or biased against political conservatives [99,219]. In August 2018, the 45th President of the United States stated that tech companies "are totally discriminating against Republican/Conservative voices", though no evidence was offered to back the claim [151]. Therefore, I first investigate the veracity of these allegations and propose two RQs, **RQ2.1**, *does the political leaning of a video affect the moderation decision of its comments?* **RQ2.2**, *does the political extremeness of a video affect the moderation decision of its comments?* These two RQs investigate the impact of two key measures of partisanship, its leaning (left or right) and extremeness (extreme or center).

Content moderation is not specifically designed for the misinformation problem, however, as I will show in § 3.3, misinformation does alter audiences' comments and increasing their usages of swear words and others that might violate community guidelines, and therefore misinformation content could draw more attention from moderators due to this increased violation rate. A news from Facebook also suggested that their policies were updated "to remove misinformation that has the potential to contribute to imminent violence, physical harm, and voter suppression." [49] Therefore, it is worth investigating how misinformation and content moderation interact in practice. Using YouTube videos and comments as a lens, I propose two RQs, **RQ2.3**, *does the veracity of content in a video affect the moderation decision of its comments?* And continuing on the effect of fact-checking, **RQ2.4**, *does a fact-check of a video affect the moderation decision of its comments?*

2.3 Natural Language Processing for Social Science

NLP methods have been increasingly used for social science research on text data. In this section, I introduce the background of NLP for social science.

2.3.1 Bag-of-Words and Lexicons

In the realm of computational social science, automatically *scoring* text is a common prerequisite for hypothesis testing. Existing studies that used language as a signal mostly adopted a simple, straightforward scoring method that leveraged unigram-based bag-of-words (BoW) models [66, 85]. In short, this method counts word occurrence in text and maps words to pre-defined dictionary categories, e.g., the word "bad" to the category "negative".

Using such dictionary categories is one of the traditional ways to perform computational analysis of text corpora [98, 167]. Originally, these techniques focused on *sentiment analysis*, with only positive and negative sentiment labels on words. Over time, researchers built more fine-grained lexicons for more sophisticated emotions and topics.

There are several existing lexicons that are commonly used to perform text analysis. The most extensively used and validated lexicon is Linguistic Inquiry and Word Count (LIWC) [171, 214], which contains both emotional, topical, and syntactic categories. An alternative for LIWC is Empath [53], which is an automatically generated lexicon that was later manually validated by crowdsourced workers. Empath has strong correlations with LIWC within overlapping categories of words. NRC Word-Emotion Association Lexicon (EmoLex) [145, 146] is another human curated lexicon that is structured around Plutchik's wheel of emotions [174]; it includes eight primary emotions (anger, anticipation, joy, trust, fear, surprise, sadness, and disgust) and two additional classes for all positive and negative emotions. Other lexicons include the General Inquirer (GI) [207] which has more topics than LIWC but fewer emotions, and Affective Norms for English Words (ANEW) [21] and SentiWordNet [12, 47] which have more emotions than LIWC.

Although the psychological foundations of the above lexicons are solid, they are extracted from general text, and usually do not perform well when analyzing text from specific contexts [120]. In the case of social media, existing lexicons such as NRC Hashtag Emotion Lexicon (HashEmo) [144] and others [16, 124] are mostly automatically generated and not manually validated.

An alternative approach to perform text analysis is to *learn* a lexicon for a specific domain. Recently, one extensively used method is to learn vector representations of word embeddings [130,

141, 172] and then use unsupervised learning to cluster words [53]. This methods has been used by some studies in the misinformation domain to analyze stylistic features of articles [184, 223].

Existing lexicons are insufficient for my research as they offered a limited number of word categories. Therefore, I construct a new context-specific lexicon with emotional and topical categories for user comments on fact-checked social media posts, and also use EmoLex and LIWC throughout § 3 as supporting evidence to validate my findings. I also present a performance evaluation between lexicons in terms of predictive ability in § 3.2.3.

2.3.2 Sequence and Neural Models

BoW and lexicons have limited applicability for tasks that require a higher accuracy, as this method ignores the dependency among words. Therefore, more targeted analysis, e.g., identifying expressed (dis)belief, requires models to comprehend the entire text sequence as a whole instead of averaging signals of unigrams.

Modeling a specific task as a sequence classification problem, the *score* of text is the native output of probabilistic classifiers [234]. Recent solutions for solving the sequence classification problem use neural architectures [110, 240] and pre-trained transfer-learning models [42, 126, 236].

Specific applications of the sequence classification problem are defined within domain-specific datasets. There are several existing NLP tasks that are related to my task. Stance detection, for example, aims to determine the for-or-against stance in comments for a two-sided argument (e.g., marijuana, gay marriage) [80, 97], and, in the political context, it often overlaps with ideology identification [181]. Classifications of other creative languages such as sarcasm [72], satire [24], irony [50], and humor [235] also share certain commonalities with my task.

Chapter 3

Audiences

In this chapter, I measure audiences' response to misinformation and investigate the following RQs:

- **RQ1.1**, *do linguistic signals in user comments vary in the presence of misinformation?*
- **RQ1.2**, *do linguistic signals in user comments vary after a post is fact-checked?*
- **RQ1.3**, *do audiences believe in misinformation, and if so, to what extent?*
- **RQ1.4**, *does the prevalence of expressed (dis)belief in misinformation vary over time?*
- **RQ1.5**, *does the prevalence of expressed (dis)belief in misinformation vary after fact-checking?*

3.1 Audiences' Comments to Misinformation - an Unlabeled Dataset

RQ1.1 and **RQ1.2** require an unsupervised exploration of audiences' response, I first collect an unlabeled dataset of audiences' comments to misinformation. In the section, I discuss how this dataset is collected and give an overview of the data.

3.1.1 Data Collection from Fact-Checks and Social Media

The interaction between social media and fact-checking websites is shown in Figure 3.1. Politicians, news organizations, or other individuals publish posts on social media websites such as Twitter, Facebook, YouTube, etc. Some of these posts are selected for fact-checking by specialized journalists at websites such as Snopes and PolitiFact, who then publish articles containing evidence for or against

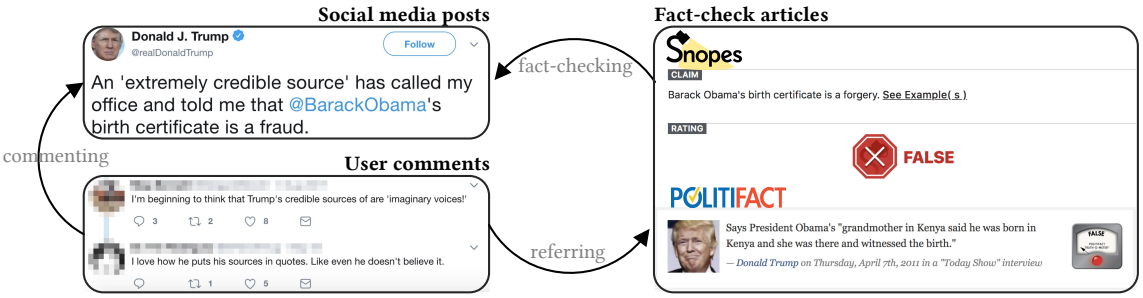

Figure 3.1: **Interaction between social media and fact-checking websites.** Following the publication of a post on Twitter, Facebook, YouTube, etc., Snopes and PolitiFact fact-check it and rate its veracity. Meanwhile, users comment on the post and sometimes refer to fact-check articles once they are released.

the claims and reasoning within the posts, as well as a veracity ruling for the posts. Meanwhile, users may comment on the posts, which sometimes refer to the fact-check articles.

To gather this data (i.e., posts and their associated comments and fact-check articles), I use the fact-checking websites Politifact and Snopes as starting points. I choose PolitiFact and Snopes because **a)** they are both confirmed by the International Fact-Checking Network (IFCN) to be non-partisan, fair, and transparent fact-checking agencies; and **b)** they list their sources and rulings in a structured format that is easy to automatically parse. I crawled all the fact-check articles from Politifact and Snopes, and then filtered this set down to articles that point specifically to social media posts on Facebook, Twitter, and YouTube (e.g., the one from Figure 3.1). I extracted the unique post ID[1] and veracity rating from these articles. Finally, I used the Facebook, Twitter, and YouTube platform APIs to crawl all of the user comments on the fact-checked posts by leveraging their unique IDs.

3.1.2 Overview of Data

Overall, I collected 14,184 fact-check articles from Politifact and 11,345 from Snopes, spanning from their founding to January 9, 2018. After filtered out all articles whose sources were not from Facebook, Twitter, or YouTube, my dataset contained 1,103 social media posts from Facebook, 2,573 from Twitter, and 2,753 YouTube videos.

Note that PolitiFact and Snopes have different ruling criterion and therefore different textual descriptions for post veracity. To make them comparable, I translated their descriptions to a scale

[1] Although the post ID formats for Facebook, Twitter, and YouTube are not the same, they are all structured and relatively easy to automatically parse.

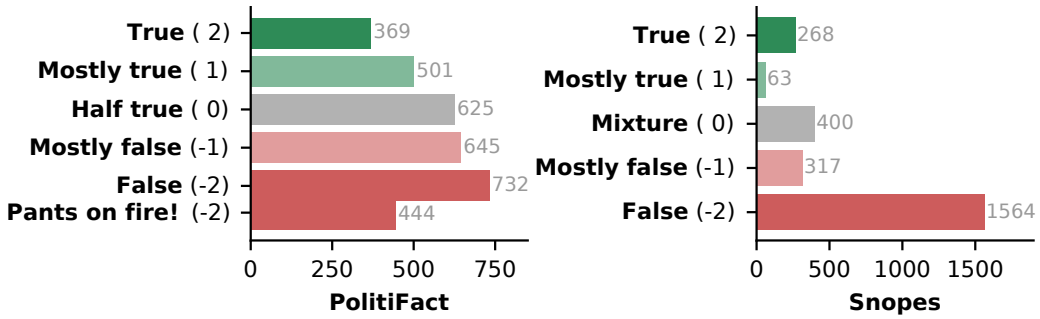

Figure 3.2: **Distribution of veracity for posts from PolitiFact and Snopes.** I map textual descriptions of veracity to ordinal values. I ignore descriptions that cannot be categorized such as *full flop, half flip, no flip* from PolitiFact and *legend, outdated, unproven, undetermined, research in progress, miscaptioned, misattributed, correct attribution, not applicable, etc.* from Snopes.

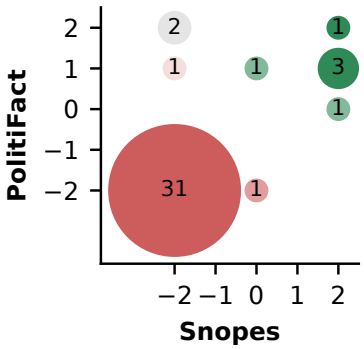

Figure 3.3: **Veracity of posts fact-checked by both PolitiFact and Snopes.** The veracity rulings are strongly correlated ($\rho = 0.671^{***}$).

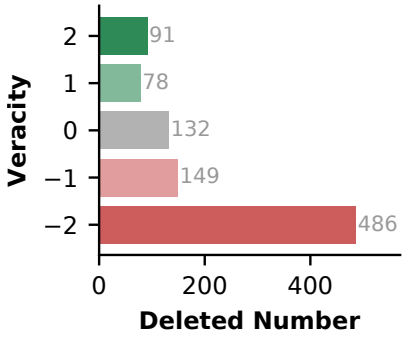

Figure 3.4: **Distribution of veracity for deleted posts.** The likelihood of post deletion is negatively correlated with the veracity of posts ($r_{pb} = -0.052^{***}$).

from -2 to 2 using the mapping shown in Figure 3.2. I view *pants on fire!* and *false* as -2 for PolitiFact, and ignore descriptions that cannot be categorized such as *full flop, half flip, no flip* from PolitiFact and *legend, outdated, unproven, undetermined, research in progress, miscaptioned, misattributed, correct attribution, not applicable,* etc. from Snopes. After mapping and removing descriptions that cannot be categorized, I kept 5,303 posts. 41 of 5,303 (0.77%) of the mapped posts were checked by both PolitiFact and Snopes, and their veracity rulings from the two websites are strongly correlated (Spearman $\rho = 0.671^{***}$)[2] as shown in Figure 3.3, which is consistent with previous observations [4, 5].

Finally, I collected user comments on the 5,303 fact-checked social media posts using their

[2] $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

respective platform APIs. I note that 1,659 (31%) of the posts were no longer available because they were either deleted by the platform or by their authors, of which 1,364 (82%) had veracity ≤ 0. This finding may be attributable to platforms' efforts to fight misinformation [58, 210]. In addition, there were 757 posts with zero comments. From the remaining posts I collected 1,672,687 comments from Facebook, 113,687 from Twitter, and 828,000 from YouTube.

Before moving on, I take a deeper look at the deleted posts. The distribution of their veracity is shown in Figure 3.4. I observe that the likelihood of post deletion increases significantly as veracity decreases (Point Biserial $r_{pb} = -0.052^{***}$). This means that, overall, my dataset is missing some deeply misleading and/or untrue posts and their associated comments. These omissions will make my model under-estimate the effect of misinformation and fact-checking. Therefore, my statistics should be viewed as conservative lower bounds on the linguistic variants in user comments in the presence misinformation and fact-checking.

I were careful to obey standard ethical practices during my data collection. I only used official APIs from the social networks to collect data, I did not make any "sock puppet" accounts, and I rate limited my crawlers. All of the posts and associated comments are publicly accessible, and my dataset does not contain any posts or comments that were deleted or hidden by their authors prior to my crawl in January 2018. The datasets that I plan to publish are fully anonymized, i.e., all user IDs are removed.

3.2 Lexicon Construction for Linguistic Signals

Using the collected dataset, I build a new lexicon called *ComLex* based on the corpus of user comments. In this section, I discuss how I constructed the lexicon, and then present three complementary validation tests based on (1) human raters, (2) comparisons with two representative lexicons from prior work, and (3) re-evaluation of datasets used in prior work.

3.2.1 Building ComLex via Clustering Word Embeddings

I generate ComLex using a combination of learning word embeddings and unsupervised clustering. I first build a corpus of user comments by applying standard text preprocessing techniques using *NLTK* [127], including tokenization, case-folding, and lemmatization. Importantly, I choose not to remove punctuation and non-letter symbols because such symbols may carry meanings for my task, such as exclamation "!" and smile ":)". This also allow me to keep emojis, which are important

"words" for my analysis because they enable users' to express emotional signals, sometimes even more significantly than with text [2, 55]. In addition, I replaced all URLs that link to a Snopes or PolitiFact webpages with the special tokens *snopesref* or *politifactref*. This enables me to group all fact-checked posts from Snopes and PolitiFact together, respectively, and later learn their semantics.

Next, I learn word embeddings from the clean corpus, i.e., transform words into vector space to provide numerical representations of each word in the corpus. To do this, I use *gensim* [187] to learn *Word2Vec* [141] representations, and use a 100-dimension vector to represent each word. To avoid noise, I only kept words that appear ≥ 100 times in the corpus. Subsequently, I apply spectral clustering [155] to divide my vectors into 300 disjoint clusters, with each cluster contains words with similar semantics. Finally, I manually examined each cluster and provide a suitable name and additional descriptive information for it. The final, labeled clusters of words are ComLex.

For each cluster in a given lexicon (e.g., ComLex, EmoLex, or LIWC), I compute a statistic for each user comment based on the word frequencies in each cluster. I then normalize these statistics by the total word frequencies in a cluster. My analytical sections mainly focus on the statistics from ComLex, but I also provide results from EmoLex and LIWC as support.

3.2.2 Human Evaluation of ComLex

To validate the robustness of my lexicon, I designed a survey that included two rating questions:

> **Semantic closeness**, *how closely, in terms of semantics, are words in each cluster related to each other?* Please provide a rating from 1 to 5 with 1 being not related and 5 being extremely related for each word cluster. *e.g., "apple, banana, peach, grape, cherry" should be considered extremely related (5) since they are all fruits; "apple, sky, happy, tomorrow, birds" should be considered not related (1).*
>
> **Information accuracy**, *how accurately do the name and additional information describe the word cluster?* Please provide a rating from 1 to 5 with 1 being not accurate and 5 being extremely accurate for each word cluster. *e.g., "fruit" should be considered extremely accurate (5) for a cluster of apple, banana, peach, grape, cherry; "weather" should be considered not accurate (1).*

Each question asked for a rating on 5-point Likert scale, with descriptive adverbs chosen from [22]. The authors (*A* in Figures 3.5 and 3.6) took the survey first and gave ratings for all learned clusters. I then keep only the top 56 of 300 (18.7%) clusters with ratings $\geq$4 for both questions. After this filtering process, I then asked three independent raters (*R1*, *R2*, and *R3*) to take the survey to rate the remaining 56 clusters to ensure semantic closeness and accurate cluster names.

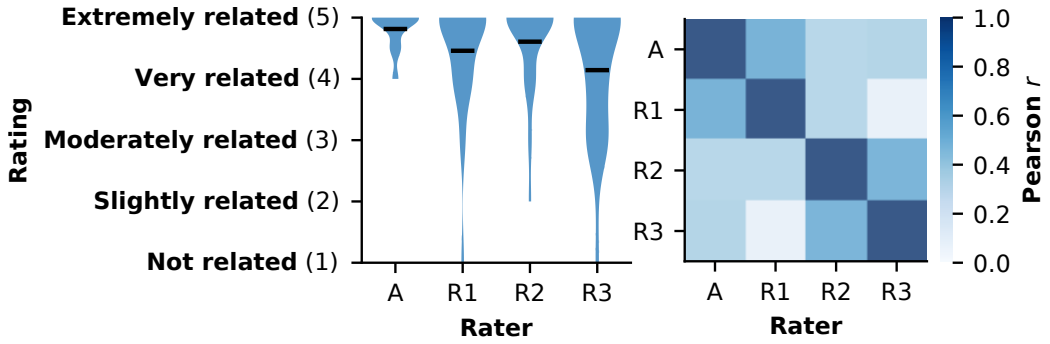

Figure 3.5: **Survey results for semantic closeness.** The left violin plot shows the rating distribution of four raters and the right heatmap shows the inter-rater correlations. Words in clusters are rated on average above "very related" ($\overline{\mu} = 4.506$) with moderate inter-rater agreement ($\overline{r} = 0.531$).

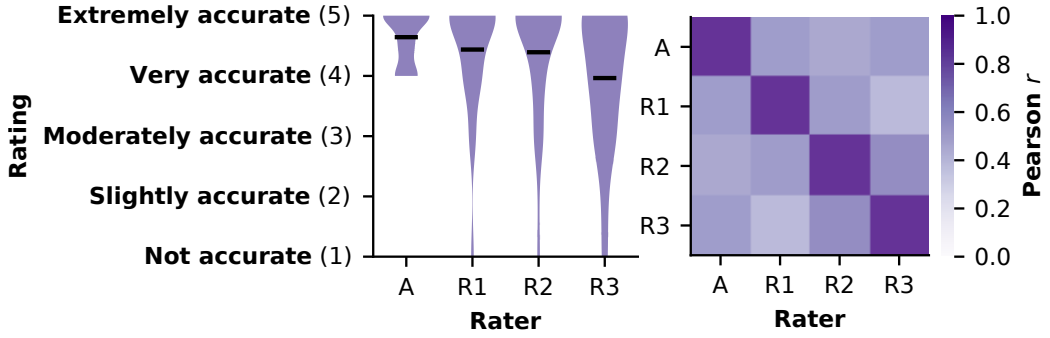

Figure 3.6: **Survey results for information accuracy.** The left violin plot shows the rating distribution of four raters and the right heatmap shows the inter-rater correlations. Cluster names and additional information are rated on average above "very accurate" ($\overline{\mu} = 4.359$) with strong inter-rater agreement ($\overline{r} = 0.675$).

Figure 3.5 shows the results of the first survey question on semantic closeness. The violin plot shows the distribution of four raters, among which the authors gave the highest average rating ($\mu_A = 4.814$) and R3 gave the lowest ($\mu_{R3} = 4.143$). Overall, words in clusters are rated above "very related" on average (mean average $\overline{\mu} = 4.506$), and the difference in μ among raters is significant (Kruskal-Wallis $H = 11.3^*$). The heatmap shows the inter-rater agreement represented by Pearson correlation, demonstrating moderate agreement among the raters on average (mean Pearson $\overline{r} = 0.531$).

Figure 3.6 shows the results of the second survey question on information accuracy. As shown in the violin plot, the authors gave the highest average rating ($\mu_A = 4.643$) and R3 gave the lowest ($\mu_{R3} = 3.964$). Overall, cluster names and additional information are rated above "very accurate" on average ($\overline{\mu} = 4.359$), and the difference in μ among raters is significant ($H = 10.8^*$). As shown

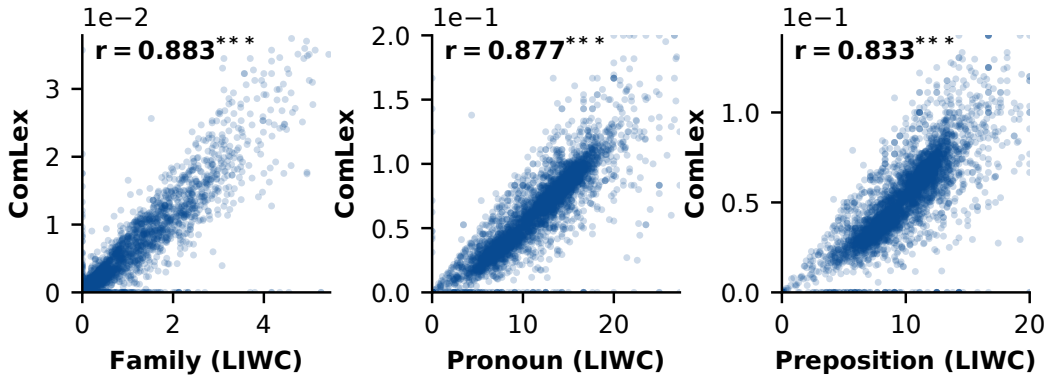

Figure 3.7: **Comparing ComLex with LIWC.** Each scatter plot shows the correlation of ComLex and LIWC for a similar word cluster. Selected clusters including *family* ($r = 0.883^{***}$), *pronoun* ($r = 0.877^{***}$) and *preposition* ($r = 0.833^{***}$) show very strong correlation.

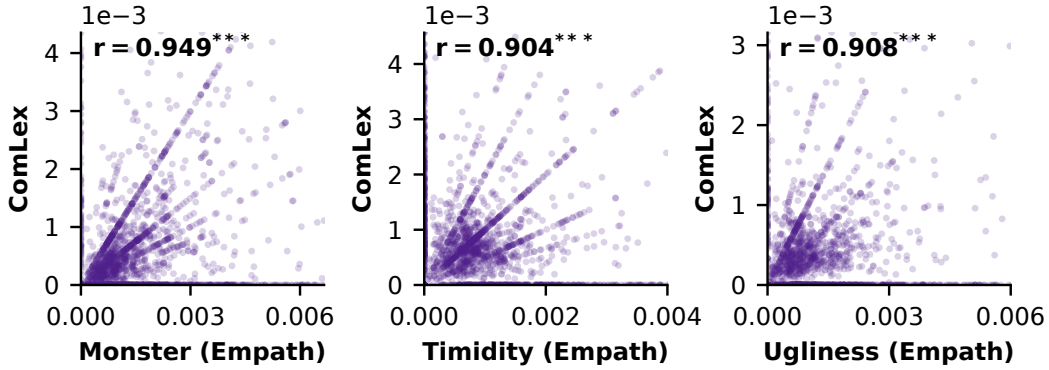

Figure 3.8: **Comparing ComLex with Empath.** Each scatter plot shows the correlation of ComLex and Empath for a similar word cluster. Selected clusters including *monster* ($r = 0.949^{***}$), *timidity* ($r = 0.904^{***}$) and *ugliness* ($r = 0.908^{***}$) show very strong correlation.

in the heatmap, the raters are strongly agreed with each other on average ($\overline{r} = 0.675$). These results show that ComLex is perceived as valid by humans.

3.2.3 Comparing ComLex with LIWC and Empath

Next, I compare ComLex with two existing lexicons: LIWC and Empath. LIWC is arguably the most extensively used lexicon, while Empath is generated in a similar manner to ComLex. I pair the statistics of user comments mapped using these lexicons and then select similar clusters to compare their correlation.

Figure 3.7 shows the comparison with LIWC. Each scatter plot shows the correlation of a similar word cluster between ComLex and LIWC. ComLex shows very strong correlation with LIWC in

Table 3.1: **Application of ComLex on related tasks.** The upper part of the table shows the performance of ComLex at detecting deception in hotel reviews. It outperforms human judges, GI, and LIWC, but is not as accurate as learned unigrams. The lower part of the table shows the performance of ComLex at detecting sentiment of movie reviews. It outperforms human judges and is nearly as accurate as learned unigrams.

Dataset	Lexicon	Model	Accuracy*
Hotel reviews [164]	Human judges		56.9% - 61.9%
	GI	SVM	73.0%
	LIWC		76.8%
	ComLex		**81.4%**
	Learned unigrams		88.4%
Movie reviews [168]	Human judges		58.0% - 69.0%
	ComLex	SVM	**72.3%**
	Learned unigrams		72.8%

**All accuracy data are drawn from the original papers [164, 168] except for ComLex.*

similar clusters such as *family* (Pearson $r = 0.883^{***}$), *pronoun* ($r = 0.877^{***}$) and *preposition* ($r = 0.833^{***}$).

Figure 3.8 shows the comparison with Empath. Again, ComLex shows very strong correlation with Empath in similar clusters such as*monster* ($r = 0.949^{***}$), *timidity* ($r = 0.904^{***}$) and *ugliness* ($r = 0.908^{***}$). This step shows that statistics derived from ComLex and LIWC/Empath are similar for overlapping word categories.

3.2.4 Application of ComLex on Related Tasks

Lastly, I test ComLex on previously released datasets to evaluate the generality and performance of ComLex when applied to related domains. In the following experiments, I run ComLex datasets of hotel and movie reviews, respectively, and build predictive models to evaluate the performance of ComLex. To compare my accuracy with the ones reported in the original papers, I adopt the same learning model (Support Vector Machine, SVM), and report the same evaluation metric (accuracy). Note that the datasets I choose have balanced binary labels, therefore accuracy is a reasonable metric for evaluation.

The first application uses a hotel dataset of 800 positive reviews [164], of which half are truthful reviews from TripAdvisor and half are deceptive reviews from Amazon Mechanical Turk. The task is to predict whether a review is truthful or deceptive. The original paper reported the accuracy of three human judges, the existing GI and LIWC lexicons, and domain-specific learned unigrams. I

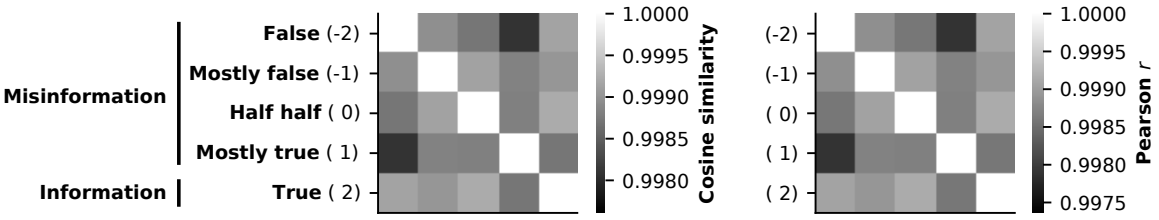

Figure 3.9: **Similarity matrix over veracity.** Heatmaps shows the similarity matrix over veracity using cosine similarity and Pearson correlation respectively. Using both measures, clear patterns of decreasing similarity are visible from -2 to 1, but the trend does not hold for 2.

run 10-fold cross validation using vectors mapped by ComLex and report my results in Table 3.1. I see that ComLex outperforms the human judges, GI, and LIWC, but not the learned unigrams.

The second application uses a movie dataset of 1,400 reviews [168], of which half are labeled as positive and half as negative. The task is to predict whether a review is positive or negative. The original paper reported the accuracy of three human judges and domain-specific learned unigrams. I run 10-fold cross validation using vectors mapped by ComLex and report my results in Table 3.1. Again, ComLex outperforms the human judges, and is nearly as accurate as the learned unigrams.

ComLex is generated using my dataset of user comments specifically on misinformation, yet it is essentially a lexicon of user comments in general, and leverages comments from multiple sources, i.e., Facebook, Twitter, and YouTube. This step demonstrates that ComLex can be broadly and flexibly applied to other related domains with reasonable performance.

3.3 Unsupervised Exploration of Linguistic Signals

In this section, I focus on linguistic signals in the presence of misinformation. Using my dataset and ComLex, I analyze how linguistic signals vary versus the veracity of social media posts. Considering that fact-checking articles may be a strong confounding variable in this analysis, I only examine user comments that were generated *before* the post was fact-checked.

3.3.1 Effect of Misinformation on Audiences' Response

Before I analyze specific linguistic clusters, I first take a look at the overall linguistic similarity between user comments on posts of varying veracity. To do this, I group user comments by the veracity (-2 to 2) of the post and compute the mean of all vectors in that veracity group. I then compute the cosine similarity and Pearson's r between different veracity groups.

As shown in Figure 3.9, there is a clear pattern from *false* (-2) to *mostly true* (-1): users' comments are self-identical (1.0), and the similarity gradually decrease as the comparisons become more distant (e.g., *false* versus *mostly true*). However, this pattern does not hold for comments on posts whose veracity is *true* (2). This observation holds regardless of whether cosine similarity or Pearson correlation is used to compute distance. This motivates me to split my research questions into different experiments by looking at the *degree* of misinformation and the *existence* of misinformation separately. In the following sections, I will first look at how linguistic signals vary versus the *degree* of misinformation by analyzing user comments from posts rated from -2 to 1, and then looking at how linguistic signals vary versus the *existence* of misinformation by comparing posts rated 2 to those rated < 2.

In this section, I examine **RQ1.1**, *do linguistic signals in user comments vary in the presence of misinformation?*, i.e., , whether there are differences in the emotional and topical signals expressed in user comments based on the degree of misinformation in the original post. I perform Spearman correlation tests between each word cluster's normalized frequency and each veracity value, and report significant results of ρ in Figure 3.10.

The first evidence for **RQ1.1** is that **the usage likelihoods for several word clusters that express misinformation-awareness are negatively correlated with veracity.** These clusters include verbs that describe fakes (*fake*, *mislead*, *fabricate*, etc., $\rho = -0.087^{***}$), and nouns for very fake content (*hoax*, *scam*, *conspiracy*, etc., $\rho = -0.045^{*}$) and somewhat fake content (*propaganda*, *rumor*, *distortion*, etc., $\rho = -0.046^{*}$), e.g., "this is fake news", "this is brainwash propaganda", etc. This means social media users are more likely to use these misinformation-aware words when commenting on posts that are ultimately proven to have low veracity. Combining these word clusters together, their mean values increase from 0.0025 to 0.0033 as veracity decreases from 1 to -2, i.e., on average, each word that identifies misinformation has a 9.7% greater chance of appearing in each user comments with one decrement in veracity. This observation is, in a different direction, supported by EmoLex where **trust declines as misinformation increases**. I observe positive correlations between veracity and word clusters that express *trust* (*accountable*, *lawful*, *scientific*, etc., $\rho = 0.063^{**}$). This means people are less likely to express trust when commenting on posts that are ultimately shown to have low veracity. In terms of effect size, the mean value of the trust category decreases from 0.0554 to 0.053 as veracity decreases from 1 to -2, i.e., on average, people are using 1.4% less of these words with each single decrement of veracity.

The second evidence for **RQ1.1** is that **the usage of emojis increases as misinformation increases**. I observe significant negative correlations for eight clusters of emoji, including *gesture*

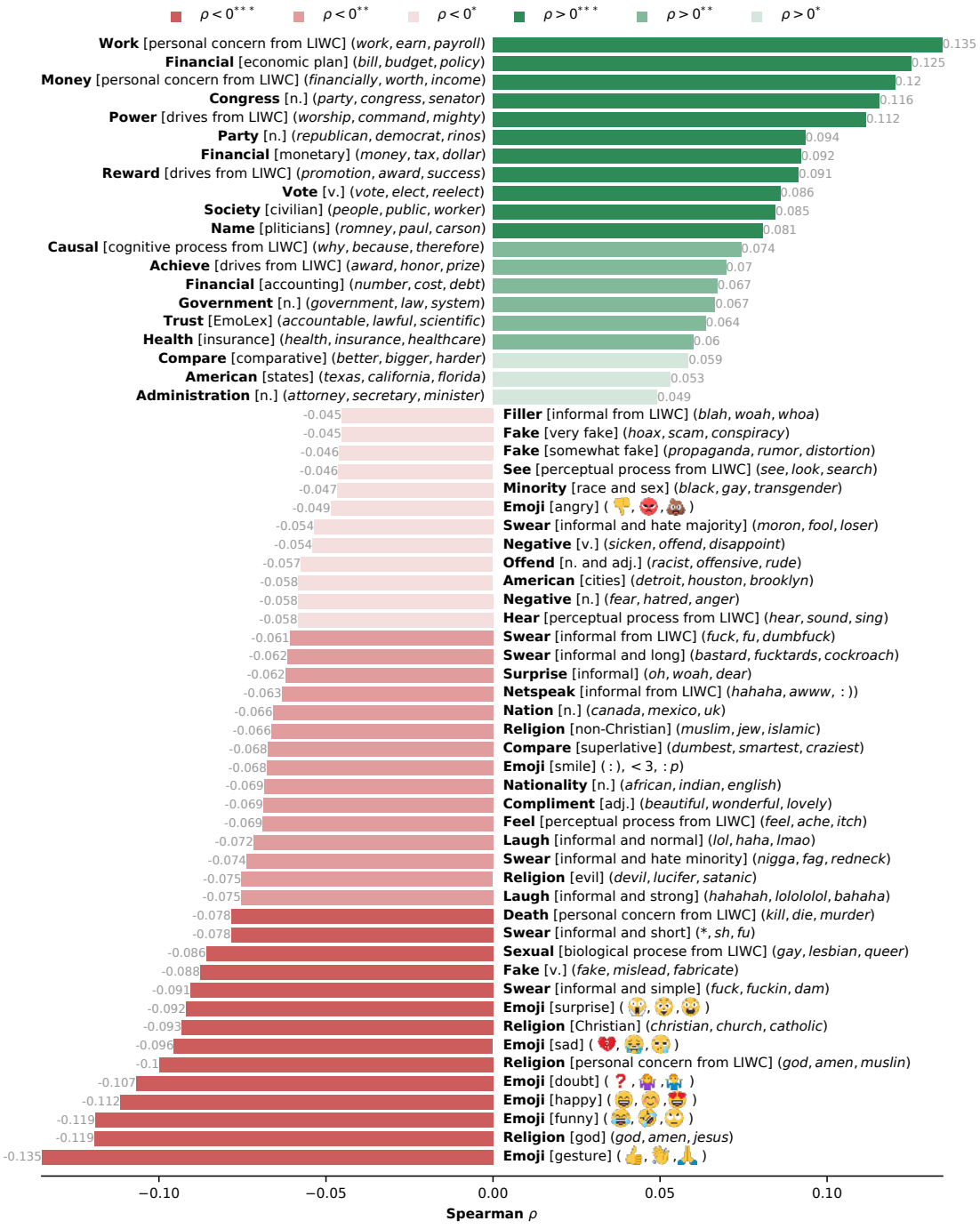

Figure 3.10: **Linguistic signals versus degree of misinformation.** Clusters with significance ρ are plotted, ranked by the sign and strength of correlation. A positive ρ indicates that the statistic increases with veracity, and vice versa. Clusters are labeled in the figure using the format: **name** [additional information] (*three example words*).

(👍, 👋, 🙏, etc., $\rho = -0.135^{***}$), *funny* (😂, 🤣, 🙄, etc., $\rho = -0.119^{***}$), *happy* (😁, 😊, 😍, etc., $\rho = -0.112^{***}$), *question* (❓, 🤷‍♀️, 🤷‍♂️, etc., $\rho = -0.107^{***}$), *sad* (💔, 😭, 😟, etc., $\rho = -0.096^{***}$), *surprise* (😱, 😳, 😬, etc., $\rho = -0.092^{***}$), and *angry* (👎, 👺, 💩, etc., $\rho = -0.049^{*}$), e.g., "so ridiculous 😂😂", "really? 😭", "i smell bull 💩", etc. This means people are more likely to use these emoji when commenting on posts that are ultimately proven to have low veracity. Combining these emoji clusters together, their mean values increase from 0.0015 to 0.005 as veracity decreases from 1 to -2, i.e., users are 49.4% more likely to use emojis with each single decrement in veracity value. Given the popularity of emojis [55, 128], I view them as important proxies for people's actual emotional state [101, 189, 239] when confronted with misinformation.

The third evidence for **RQ1.1** is that **the usage of swear words increases as misinformation increases**. I observe significant negative correlations for five clusters of swear words, including popular swear words (*fuck*, etc., $\rho = -0.091^{***}$), shortened or moderated swear words (*, *fu*, etc., $\rho = -0.078^{***}$), hateful terms against minority groups ($\rho = -0.074^{**}$), long and complicated swears (*bastard*, etc., $\rho = -0.062^{**}$), and belittling words (*moron*, *fool*, *loser*, etc., $\rho = -0.054^{*}$). This means people are more likely to swear or use hateful terms towards other users (including the author of the post) when commenting on posts that are eventually found to have low veracity. Combining these swear clusters together, their mean values increases from 0.0034 to 0.0046 as veracity decreases to -2, i.e., on average, users are using 16.3% more swear words with one decrement in veracity value. This observation is further supported by LIWC's swear word category (*fuck*, etc., $\rho = -0.061^{**}$). People associate swear words with their own emotional states, and these words affect the emotional states of others [183]. In my data, I observe an increasing amount of people using negative or offensive words in comments as veracity decreases and swear words increase. This includes negative correlations with a cluster of negative verbs (*sicken*, *offend*, *disappoint*, etc., $\rho = -0.054^{*}$) and another of offensive nouns and adjectives (*racist*, *offensive*, *rude*, etc., $\rho = -0.057^{*}$) with an effect size of 3.7%.

The fourth evidence for **RQ1.1** is that **discussion of concrete topics declines as misinformation increases**. I observe significant positive correlations for 12 clusters of words about concrete political topics, including financial clusters about economic plans (*bill*, *budget*, *policy*, etc., $\rho = 0.125^{***}$) and monetary issues (*money*, *tax*, *dollar*, etc., $\rho = 0.092^{***}$), and clusters about congress (*party*, *congress*, *senator*, etc., $\rho = 0.116^{***}$), party (*republican*, *democrat*, *rinos*, etc., $\rho = 0.094^{***}$), voting (*vote*, *elect*, *reelect*, etc., $\rho = 0.086^{***}$), society (*people*, *public*, *worker*, etc., $\rho = 0.085^{***}$), government (*government*, *law*, *system*, etc., $\rho = 0.067^{**}$), health (*health*, *insurance*, *healthcare*, etc., $\rho = 0.06^{**}$), administration (*attorney*, *secretary*, *minister*, etc., $\rho = 0.049^{*}$), and references

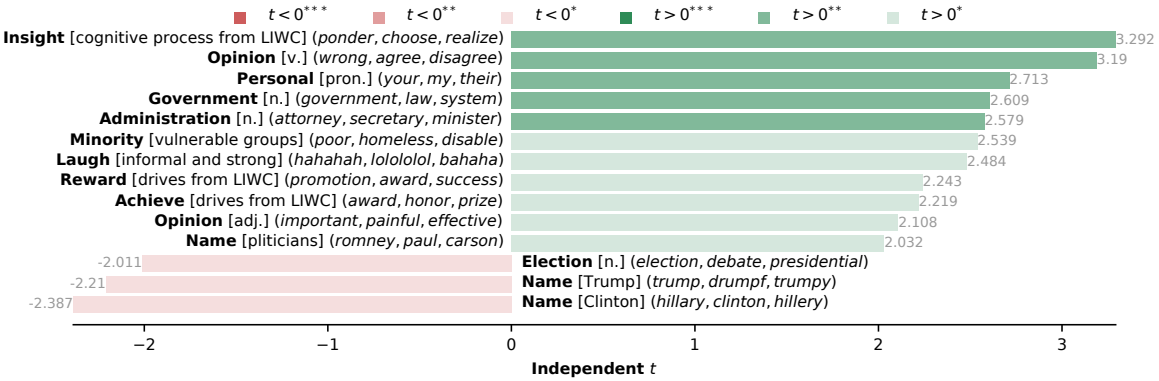

Figure 3.11: **Linguistic signals versus existence of misinformation.** Clusters with significance independent t are plotted, ranked by the sign and strength of difference. A positive t indicates that the mean of the statistic for accurate information is higher than misinformation, and vice versa.

to states ($\rho = 0.053^{*}$) and politicians ($\rho = 0.081^{**}$). This means people are more likely to talk about concrete topics on posts with higher veracity. Combining these clusters together, their mean value increases from 0.046 to 0.065 as veracity value increases from -2 to 1, i.e., on average, users are 12.2% more likely to use words in these clusters with one increment in veracity value. This observation is supported by LIWC's word categories involving concrete topics, including *work* ($\rho = 0.135^{***}$), *money* ($\rho = 0.120^{***}$), *power* ($\rho = 0.091^{***}$), and *achieve* ($\rho = 0.07^{**}$).

The fifth evidence for **RQ1.1** is that **objectivity declines as misinformation increases**. I observe that users are more likely to use superlatives (*dumbest*, *smartest*, *craziest*, etc., $\rho = -0.068^{**}$), e.g., "dumbest thing i've seen today", with an effect size of 25.5% with each single decrement in veracity value. At the same time, I observe that users are less likely to use comparatives (*better*, *bigger*, *harder*, etc., $\rho = 0.059^{*}$), e.g., "she would do better", with an effect size of 15.6% with each single decrement in veracity value. This observation is also supported by LIWC, where people use less causal inference (*why*, *because*, *therefore*, etc., $\rho = 0.074^{*}$) as misinformation increases. This implies that subjectivity increases and objectivity decreases as the veracity of the underlying post decreases. The relationship between subjectivity and objectivity has long been studied within the context of people's emotional states in sociology [112].

I now look at the differences in the emotional and topical signals of user comments in relation to the existence of misinformation (i.e., posts with veracity value 2 versus posts with value < 2). I report statistically significant independent t in Figure 3.11. These findings are similar to the evidences above, such as an increased likelihood of discussion about concrete political topics on true posts. This includes *government* ($t = 2.609^{**}$), *administration* ($t = 2.579^{**}$) and *minority* ($t = 2.539^{*}$).

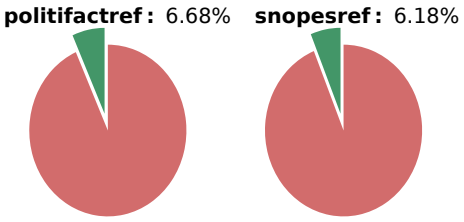

Figure 3.12: **Percentage of reference for PolitiFact and Snopes.** Each pie chart shows the percentage of posts that contains *politifactref* or *snopesref* over all posts checked by the website.

In terms of effect size, combining these clusters together, their mean is 0.0074 for misinformation and 0.01 for true posts, which represents a 35.1% difference. Similarly, I also observe that concrete topical categories from LIWC such as reward ($t = 2.243^{*}$) and achieve ($t = 2.219^{*}$) are significant.

Another supporting evidence for **RQ1.1** is **the increased likelihood of personal opinions on true posts**. I observe that users are more likely to express their opinions in a concrete manner, including opinionated adjectives (*important*, *painful*, *effective*, etc., $t = 2.108^{*}$), and personal opinions (*wrong*, *agree*, *disagree*, etc., $t = 3.190^{**}$), e.g., "this is important", "i agree with you", etc. These two clusters have a mean of 0.0036 for misinformation and 0.0049 for true posts, which represents a 36.1% difference. This is also supported by LIWC in its *insight* category, which is a subset of cognitive process ($t = 3.292^{**}$).

I also found that users are 43.1% less likely to mention the election ($t = -2.011^{*}$), Trump ($t = -2.210^{*}$), and Clinton ($t = -2.387^{*}$) when commenting on true posts. One possible explanation for this is that true posts invite discussion of more original and substantive topics, versus 2016 election coverage itself which was polarizing and prone to misinformation [1, 75].

3.3.2 Linguistic Signals after Fact-Checking

In this section, I focus on linguistic signals in the presence of fact-checking and analyze how they vary in users' comments. To motivate this analysis, I first examine the prevalence and semantics of references to fact-check articles. Note that I replaced any reference to PolitiFact and Snopes in user comments with special tokens *politifactref* and *snopesref*, respectively. I use these tokens for my analysis.

Figure 3.12 shows the prevalence of *politifactref* and *snopesref*. For all posts that were fact-checked by PolitiFact, 6.68% of them have at least one comment that mentioned PolitiFact. The number for Snopes is similar at 6.18%. This gives me an overview of the prevalence of direct

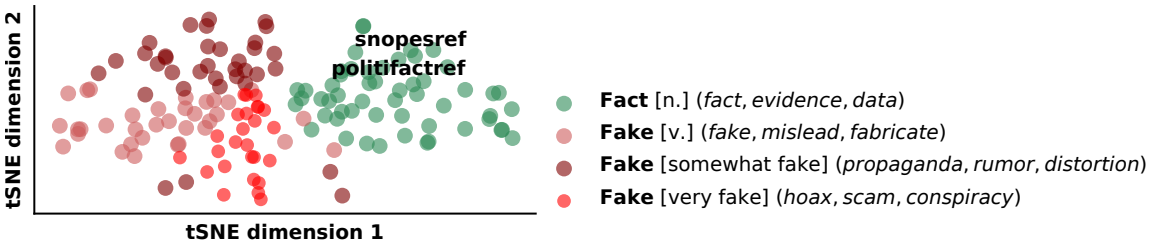

Figure 3.13: **Semantics of reference for PolitiFact and Snopes.** The learned embedding, which encodes the semantics of *politifactref* or *snopesref*, is plotted along with other words in *fact* and three *fake* clusters. Dimensions are reduced from 100 to 2 using t-SNE. References to PolitiFact and Snopes carry similar semantics as other words expressing *fact* in the right part of the figure, as oppose to words expressing *fake* in the left part of the figure.

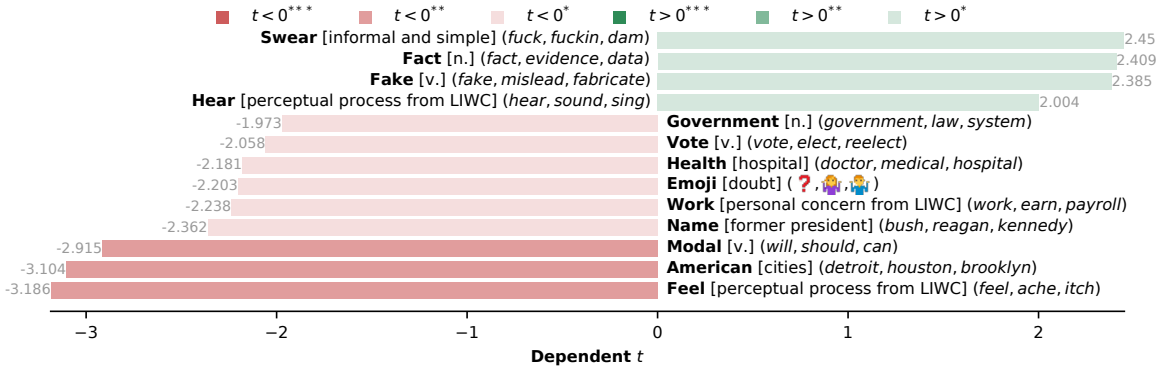

Figure 3.14: **Linguistic signals before and after fact-checking.** Clusters with significance dependent t are plotted, ranked by the sign and strength of difference. A positive t indicates that the mean of the statistic is higher after fact-checking than before, and vice versa.

references to PolitiFact and Snopes in the user comments.

Figure 3.13 shows the semantics of *politifactref* and *snopesref*. As before, I use a 100-dimensional vector to represent the semantics of each word. To visualize the proximity of word semantics, I used t-Distributed Stochastic Neighbor Embedding (t-SNE) [131] to reduce the dimensionality of each vector to 2-dimensional space. As shown in the figure, references to Snopes and PolitiFact have very similar semantics to words in the *fact* cluster (e.g., *fact*, *evidence*, *data*, *non-partisan*, etc.) as oppose to the words in three misinformation clusters (e.g., *fake*, *propaganda*, *hoax*, etc.). Also note that I observed references to other factual sources such as Wikipedia, Pew, Factcheck.org, etc. in the *fact* cluster, which suggests that within the context of user comments on social media, fact-checking websites and general purpose non-partisan websites are afforded a similar degree of trust by users.

I now analyze **RQ1.2**, *do linguistic signals in user comments vary after a post is fact-checked?*

i.e., how linguistic signals in users' comments vary before and after fact-checking. To do this, I split user comments into two groups (those written before a fact-check article was available for the given post, and those written after) and use them to perform dependent t tests. Figure 3.14 highlights the significant ($p < 0.05$) differences in emotions and topics before and after fact-checking for ComLex clusters.

The first evidence for **RQ1.2** is that **the usage likelihoods of several word clusters that express misinformation-awareness increase after a fact-check article is available**. The evidence for this claim includes an increase in factual references (*fact*, *evidence*, *data*, etc., $t = 2.409^*$) and verbs expressing deceit (*fake*, *mislead*, *fabricate*, etc., $t = 2.385^*$). Comments such as "check *snopesref* for the fact" and "according to *snopesref*, this is fake news" appear more frequently after the publication of fact-check articles. These two clusters have a mean of 0.0028 before fact-checking and 0.0042 after, which represents a 50% difference. This result suggests that social media users are aware of fact-checks, once they become available, and that this increases the likelihood of rational statements. This observation holds for the subset of posts with comments that explicitly link to PolitiFact or Snopes, e.g., the green boxes in Figure 3.12 (*fake*, $t = 2.224^*$; *fact*, $t = 2.441^*$).

The second evidence for **RQ1.2** is that **the usage likelihoods of word clusters expressing doubt decrease after a fact-check article is available**. Users' certainty increases after fact-checking, and this is reflected in the decreasing probability of using doubtful emojis (❓, 🤷, 🤷‍♂️, etc., $t = -2.203^*$), which have a mean of 0.0005 before fact-checking and 0.00025 after, which represents a 100% difference. Questions such as "is that true❓" and "is this a joke? 🤷" appear more frequently before the publication of corresponding fact-check articles.

The third evidence for **RQ1.2**, specifically about the backfire effect, is **the increase in swear words after a fact-check article is published**. I observe more swear word usage after fact-checking (*fuck*, *fuckin*, *dam*, etc., $t = 2.450^*$). In terms of effect size, the mean probability of this cluster is 0.0011 before and 0.0015 after fact-checking, which represents a 36.4% difference. However, I note that only one of five swear word clusters had significant differences before and after fact-checking, which suggests that the backfire effect in comments may be limited. Furthermore, I caution that the use of swear words is, at best, an indirect indicator of backfire: it suggests an increase in negative emotion from some users, and previous lab experiments have shown that this is symptomatic of a stubborn individual clinging to their original false beliefs [230].

Figure 3.15 shows three examples of backfire. Each presents the post veracity from a fact-check article and selected user comments exemplifying backfire effects. In all three examples, users referred to fact-checking websites and used swear words to expressed their dissatisfaction. These backfire

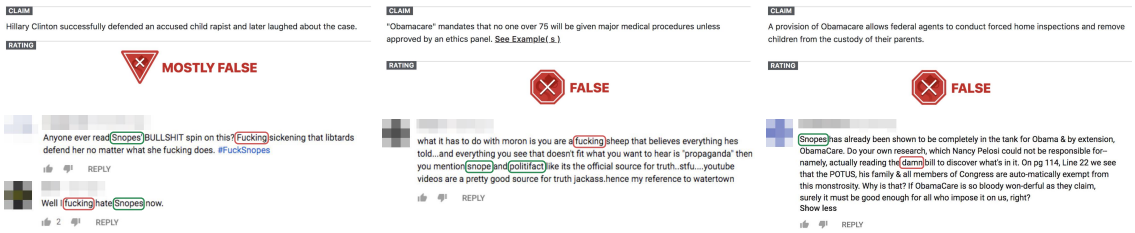

Figure 3.15: **Example comments of the backfire effect.** Three examples are given that include the post veracity from fact-check articles (top) and selected user comments indicating backfire effects (bottom). Words in green blocks (i.e., Snopes, PolitiFact) are identified as reference to fact-checking websites, while words in red blocks (i.e., fuck, damn) are mapped in the *swear* word cluster.

comments also tend to express doubt about the fact-checker themselves because the users perceive them to be biased and unreliable sources [153, 188, 197]. Note that these examples criticized Snopes or PolitiFact in whole rather than referring to individual fact-check articles.

3.4 Audiences' (Dis)belief to Misinformation - a Labeled Dataset

RQ1.3 to **RQ1.5** require a supervised measurement of audiences' (dis)belief. Therefore, I collect another labeled dataset of audiences' belief to misinformation. In the section, I discuss how this dataset is collected and annotated, and give an overview of the data and labels.

3.4.1 Another Data Collection from Fact-Checks and Social Media

I read through all of PolitiFact's fact-check articles written between January 1 to June 1, 2019 and manually found the ones whose claims originated from Twitter. I recorded the IDs of the tweets containing these claims. Using the above fact-checked tweets as seeds, I queried an archived 1% sample of the tweet stream [125] and found all *comments* to the seed tweets. In Twitter's terminology, these comments include "replies" and "retweets with comments" (i.e., quoted tweets) but excludes other retweets [218]. Note that I only keep comments whose text content is non-empty, as I aim to identify expressed (dis)belief using language features.

To filter out noise, I keep only the claims that I could link to >50 comments, which resulted in 18 claims with 6,809 comments. The short names of these claims are displayed as the x-tick labels in Figure 3.16. The full description of each claim and corresponding fact-check articles is available in my published dataset.

Representativeness. Although my archived 1% sample of the tweet stream has been shown to be representative of the Twitter ecosystem as a whole [150], this dataset is *not* a representative sample to understand the prevalence of (dis)belief at scale. This is due to **(a)** the narrow time period (i.e., half a year) of seed claims and comments, and **(b)** the omission of other mainstream social media platforms (e.g., Facebook, YouTube). While **(a)** is a common limitation on longitudinal validity in the literature [208], **(b)** is less commonly considered.

Taken together, these two issues mean that high-level statistics from this sample cannot be used to measure (dis)belief and test related hypothesis. Hence, I leverage a much larger dataset in § 3.6). However, this sample is useful to understand the *language* that people used to express their (dis)belief in response to (mis)information.

3.4.2 Annotation of (Dis)belief Labels

I annotate my unlabeled dataset of comments with belief and disbelief labels by recruiting a group of communication-majored undergrads and a faculty member from the communication department as the expert.

Task assignment. Annotating 6,809 tweets is a heavy task. To reduce the workload, I grouped these tweets by the initial claims and assigned each group of tweets to two independent human annotators. I trained the annotators, and then asked them to provide binary labels on each tweet in the given group: *disbelief* (i.e., if the person who wrote the comment *does not* believe the claim) and *belief* (i.e., if the person who wrote the comment *does* believe the claim). Note that these two labels are mutually exclusive but not necessarily complementary, i.e., I do not expect a tweet to show both belief and disbelief, but it can show neither.

Inter-annotator agreement. My task assignment strategy allows me to evaluate inter-annotator agreement at the individual group level. I use the inter-annotator percent agreement[3] (i.e., the number of agreed labels over the total count) for each group and each label, and show the results in Figure 3.16. Out of 36 evaluated groups/labels, 66.7% (24/36) are above 80% agreement, 88.9% (32/36) are above 70% agreement, and only two are below 60% agreement, suggesting a high level of agreement among annotators, especially for a relatively subjective task.

[3]Cohen's κ is not preferred here, as (dis)belief labels are, by my hypotheses, unevenly distributed and therefore κ's baseline agreement is irrelevant.

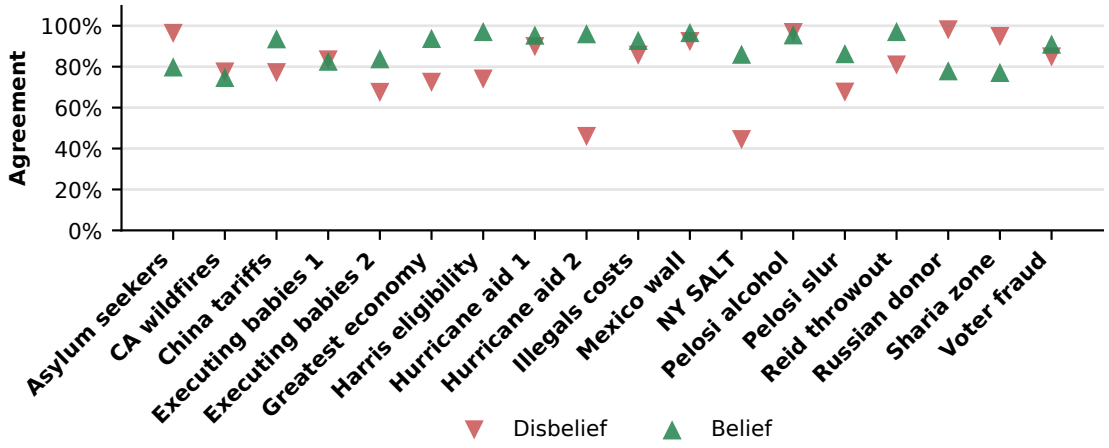

Figure 3.16: **Inter-annotator agreement per claim.** Out of 36 evaluated groups/labels, 66.7% are above 80% agreement and 88.9% are above 70% agreement.

Final labels. To obtain a final label for each tweet, a faculty member from the communication department read through all cases where two annotators disagreed and then provided a final judgement to break ties. This effectively makes my annotation process a majority vote among three members.

Note that there are two straightforward ways to formulate the (dis)belief labels: **(a)** a single-label quadruple-class formulation, where the four possible classes are: belief, disbelief, both, and neither; or **(b)** a double-label binary formulation, where one label is belief or not and the other is disbelief or not. Although these two formulations are equivalent here, **(b)** provides me with more flexibility for classification, as it is easy to threshold on each binary label and easy to analyze the performance tradeoff (as I discuss in § 3.5.2). Thus, I choose formulation **(b)** for the (dis)belief labels.

3.4.3 Overview of Data and Labels

Overall, out of 6,809 tweets, 2,399 (35.2%) are labeled as expressing disbelief, 1,282 (18.8%) are labeled as expressing belief, 3,128 (45.9%) are labeled as neither and none (0%) are labeled as both. Disbelief is over-represented in this sample (cf. the overall prevalence measured in § 3.6.1) as the 18 claims in the sample contain heavy misinformation.

The distribution of (dis)belief for each claim is shown in Figure 3.19 and Figure 3.20. There is large variation in expressed (dis)belief across the 18 claims, and the distributions of disbelief and belief are, as expected, negatively correlated (Pearson $r = -0.68^{***}$).

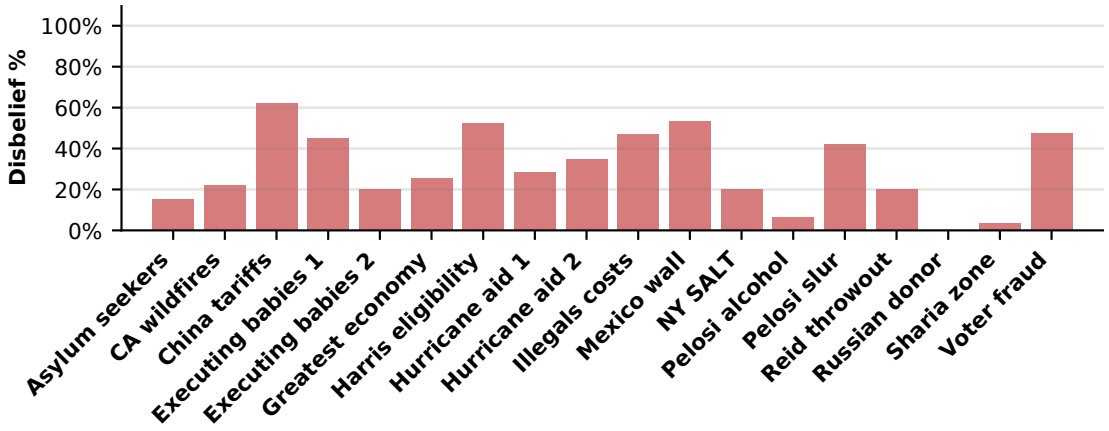

Figure 3.17: **Overview of the disbelief label per claim.** Disbelief distribution across 18 claims. The percentage of disbelief ranged from 0 to 62.4%, with a variance of 0.03.

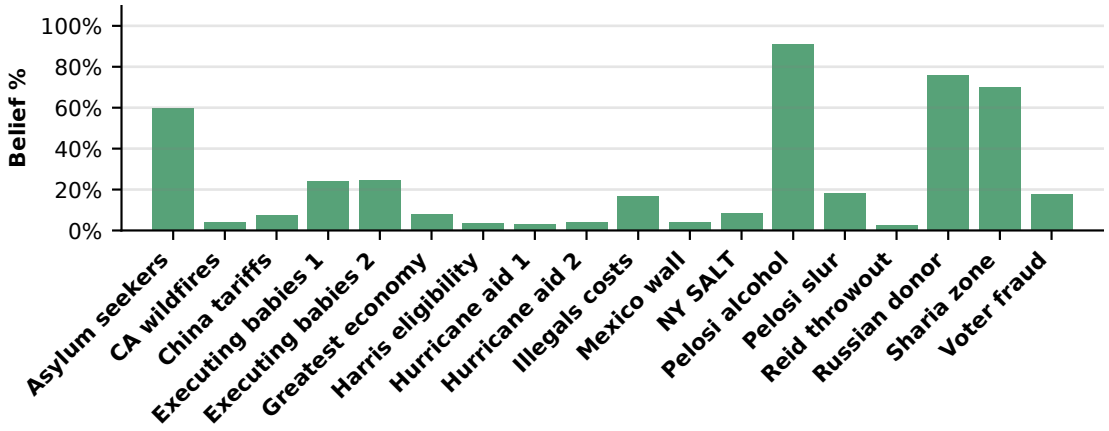

Figure 3.18: **Overview of the belief label per claim.** Belief distribution across 18 claims. The percentage of belief ranged from 2.8% to 91.1%, with a variance of 0.08.

3.5 Modeling (Dis)belief with Supervised Learning

Leveraging my labeled dataset, I first conduct a lexicon-based exploratory analysis of language used across tweets expressing belief and disbelief, and then experiment with NLP models to build classifiers.

3.5.1 Exploratory Analysis of Linguistic Signals

I start the modeling of (dis)belief by exploring the question *if tweets expressing (dis)belief use different language than the others, and if so, what are the differences?*

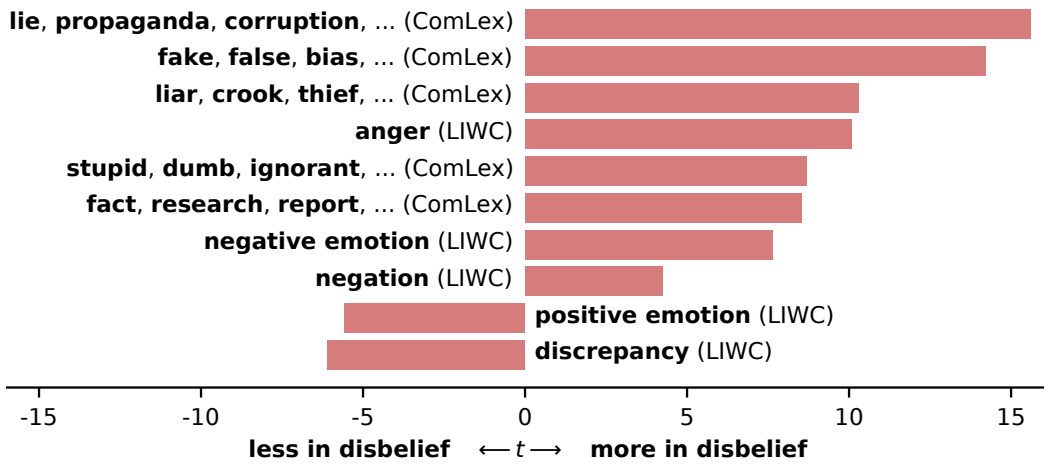

Figure 3.19: **Linguistic difference between tweets expressing disbelief and others.** Tweets expressing disbelief contains more falsehood awareness signals (e.g., "lie", "fake", "stupid") and negative emotions, and less positive emotions and discrepancy. Significance of difference is obtained by t-tests with $p < 0.01$ after Bonferroni correction.

I adopt a lexicon-based method to explore this question, and choose two lexicons: **(a)** LIWC [214], the most widely-used lexicon for understanding psychometric properties of language, containing generic emotional and topical word categories, e.g., "anger", "reward", "work"; and **(b)** ComLex [94], a more contextual lexicon built from social media comments to misinformation (§ 3.2), containing additional domain-specific categories, e.g., "fake", "fact", "hate speech".

Each word category in the lexicon contains a set of curated words that embody signals of the category (e.g., "sad" for "negative emotion"). Briefly, my method works as follows: I apply a lexicon on a tweet, which results in a frequency f_c for each category c in the lexicon, counting the overlap between words in the tweet and words in the corresponding category c. Then, at the dataset level, I compare the distributions of such frequency between tweets expressing (dis)belief and the others, by performing independent t-test for $\mathbb{E}(f_c)$. Significance is obtained by setting $p < 0.01$ after Bonferroni correction on the number of categories (392 total categories: 92 for LIWC and 300 for ComLex). Ten representative samples of significant categories with their t-values and category names[4] are shown in Figure 3.19 and Figure 3.20.

Figure 3.19 shows that tweets expressing disbelief contain more falsehood awareness signals, including referrals to falsehood "lie, propaganda, ..." ($t = 15.6^{***}$) and "fake, false, ..." ($t = 14.2^{***}$), referrals to the truth "fact, research, ..." ($t = 8.5^{***}$), and negative character portraits such as "liar,

[4]ComLex has some unnamed categories, in which case I use three words in that category as the category name.

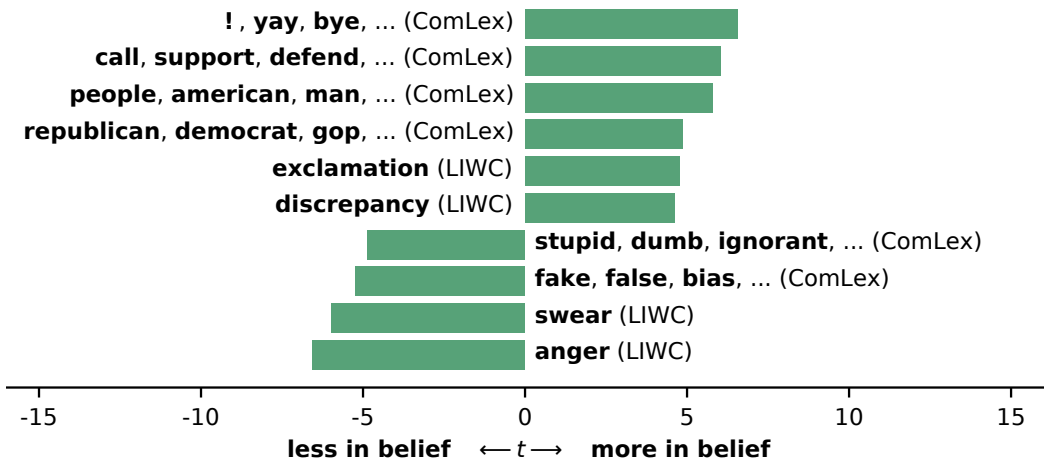

Figure 3.20: **Linguistic difference between tweets expressing belief and others.** Tweets expressing belief contains more exclamation (e.g., "!", "yay") and discrepancy, and less falsehood awareness signals (e.g., "lie", "fake", "stupid") and negative emotions. Significance of difference is obtained by t-tests with $p < 0.01$ after Bonferroni correction.

crook, ..." ($t = 10.3^{***}$) and "stupid, dumb, ..." ($t = 8.7^{***}$). These results are intuitive and provide face-validity to the existing linguistic study of misinformation responses, where similar signals were used to insinuate users' disbelief [94]. In addition, tweets expressing disbelief also contain more negative emotions ($t = 7.6^{***}$) and negation (e.g., "no, not", $t = 4.3^{***}$), less positive emotions ($t = -5.6^{***}$) and discrepancy (e.g., "should, would", $t = -6.1^{***}$).

Figure 3.20 shows that tweets expressing belief contain less falsehood awareness signals, including referrals to falsehood "fake, false, ..." ($t = -5.2^{***}$) and negative character portrait "stupid, dumb, ..." ($t = -4.8^{***}$). This is intuitively the opposite of disbelief. In addition, tweets expressing belief also contain more exclamation (for both LIWC exclamation marks, $t = 4.8^{***}$, and ComLex "!, yay, ..." category, $t = 6.6^{***}$) and discrepancy ($t = 4.6^{***}$), and less negative reactions such as swear (e.g., "damn, fuck", $t = -6.0^{***}$) and anger (e.g., "hate, kill", $t = -6.6^{***}$).

3.5.2 Experiments with Classification Models

Given these observed difference in language usage, my next question is *if such difference can be used to identify tweets that express (dis)belief?* To answer this question, I experiment with NLP models to build classifiers.

Chance. I first experiment with a chance classifier where I assign random probabilities for both disbelief and belief labels to demonstrate trivial performance baselines.

Lexicon-derived features with linear models. As a continuation of § 3.5.1, I run experiments using lexicon-derived features with linear models. For each tweet, I concatenate all mapped frequencies f_c across all categories c to a vector representation $\vec{f}$ (92 dimensions for LIWC and 300 for ComLex), and then feed these vector representations to a Logistic Regression (LR) layer for classification.

These models should perform better than trivial baselines, as they include the language signals I observed in § 3.5.1. However, their performance is still inherently limited, as such methods only capture the semantics of unigrams while ignoring the dependency between words (e.g., co-reference, phrases). Thus, these models are incapable of comprehending an entire tweet at the sequence level.

Neural transfer-learning models. To boost performance, I embed the entire sequence and leverage state-of-the-art neural transfer-learning [166] methods for the task. I experiment with three pre-trained models: BERT [42], XLNet [236], and RoBERTa [126].

This method follows a *pre-training-fine-tuning* paradigm. During the *pre-training* phase, transformer [220] or transformer-XL [39] based models are trained on large, unlabeled corpus with certain objectives, e.g., BERT and RoBERTa are trained to predict missing words in sentences, XLNet is trained to predict last tokens in factorization orders of sentences. During this process, a randomly initialized model is adjusted by back-propagation of loss, and its weights are progressively updated to embed knowledge of human language.

During the *fine-tuning* phase, models are initialized with pre-trained weights and then re-train on labeled data over specific tasks. This process tunes an already sophisticated model to perform specific downstream tasks, thus the model is expected to achieve high performance on a small labeled dataset.

To experiment with these neural models, I first preprocess tweets through the same pipeline designed in the pre-training phase, which includes tokenizing tweets at the sub-word level using specific tokenizer, and then padding or truncating the sequence to a specific length.[5] Next, these sequences are fed to an input layer which is connected to a pre-trained model. After all parameters flow through the model, I replace the last layer of the model with a double-label classification layer to predict (dis)belief. Finally, I compare the predictions and labels, calculate the cross entropy loss, and back-propagate errors. This training process is done iteratively for a certain number of epochs, as determined by cross validation on the training set.

[5] Although longer sequences are truncated to a maximum sequence length, information loss is expected to be rare, considering that commonsense writing styles usually put important (and thus identifiable) content in the beginning of comments [91].

As reported in the original papers, these models achieve state-of-the-art performance on a wide range of generic NLP tasks. Thus, I expect that they can increase performance for my task (versus the linear models) without designing domain-specific neural architectures.

Experimental setup. I randomly split the dataset into 80% (5,448) training set and 20% (1,361) testing set. My linear models were trained until convergence, which completed within one minute. I set up the neural models (BERT, XLNet, and RoBERTa) using the same neural architecture, hyperparameters, vocabularies, and tokenizers as the base models described in the original papers,[6] and I trained them for three epochs, which completed within two hours on a single Titan X Pascal GPU.

Evaluation metrics. All of the models I experiment with are probabilistic classifiers that assign a probability $\mathbb{P}$ to the positive label (i.e., disbelief or belief) and the remaining $1 - \mathbb{P}$ to the negative label (i.e., not disbelief or not belief). I then obtain the predicted label by setting a threshold $\tau \in [0, 1]$ to cut off the probability distribution so that inputs with $\mathbb{P} > \tau$ are assigned with positive labels and inputs with $\mathbb{P} < \tau$ are assigned with negative labels.

Before discussing my thresholding strategy (i.e., the choice of τ), I evaluate each classifier on the testing set using precision-recall curves that I obtained by varying τ between 0 and 1. After I choose the threshold τ, I evaluate each classifier on the testing set using unbiasedness (defined later in § 3.5.3), binary-, macro-, and micro-F_1 scores under τ.[7]

Results. The precision-recall curves of all classifiers are shown in Figure 3.21 and Figure 3.22. Linear classifiers with lexicon-derived features (LIWC+LR and ComLex+LR) outperform trivial baseline methods and achieve their best binary-F_1 scores near 0.6 for disbelief (Figure 3.21) and 0.5 for belief (Figure 3.22). Neural transfer-learning based classifiers (BERT, XLNet and RoBERTa) have the best performance, achieving their best binary-F_1 scores around 0.8 for disbelief (Figure 3.21) and 0.7 for belief (Figure 3.22). The performances of the three neural classifiers are similar, with RoBERTa being slightly better than BERT and XLNet, aligning with the results in [126] for generic NLP tasks.

3.5.3 Thresholding Scores for Measurement

In the real world, the thresholding strategy is linked to specific downstream tasks: some common strategies include applying the default $\tau = 0.5$, choosing τ that maximizes F_1/accuracy scores,

[6] Due to equipment constraints, I am unable to run large models released from these papers.

[7] For binary labels, micro-F_1 is equivalent to accuracy.

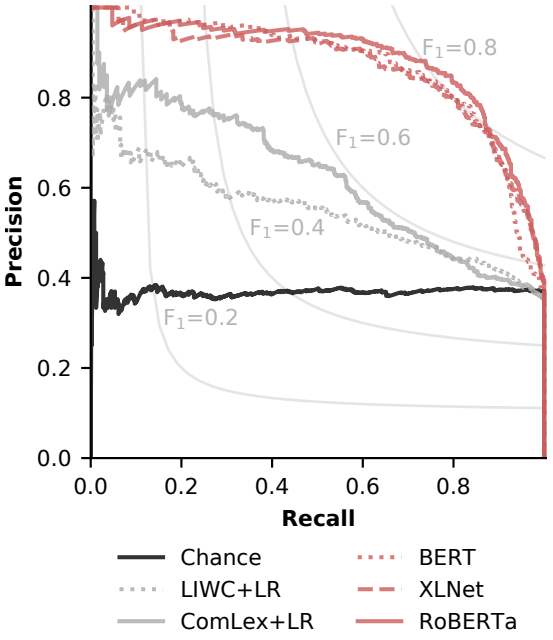

Figure 3.21: **Precision-recall curves for predicting disbelief.** Linear classifiers (LIWC+LR and ComLex+LR) achieve best binary-F_1 scores near 0.6, and neural-transfer classifiers (BERT, XLNet, and RoBERTa) achieve best binary-F_1 scores around 0.8. Isolines for binary-F_1 scores are shown.

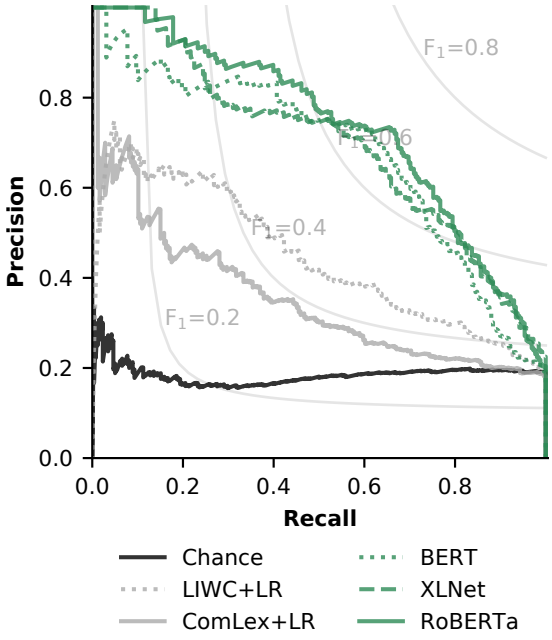

Figure 3.22: **Precision-recall curves for predicting belief.** Linear classifiers (LIWC+LR and ComLex+LR) achieve best binary-F_1 scores near 0.5, and neural-transfer classifiers (BERT, XLNet, and RoBERTa) achieve best binary-F_1 scores around 0.7. Isolines for binary-F_1 scores are shown.

choosing τ under certain precision/recall guarantees, etc.

In my case, however, the application is to use the learned classifier as a proxy for human experts, to measure (dis)belief at scale. Therefore the classifier is expected to make statistically *unbiased* estimations comparing to the underlying label distribution. This means that a desirable τ should equalize error rates between false positives and negatives, so that errors can be balanced out when the classifier is applied onto a large dataset.

Specifically, consider the following confusion matrix:

		Human experts: Positive	Human experts: Negative	
Predictions	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
		TP+FN	FP+TN	N

Table 3.2: **Evaluation results for disbelief prediction.** Chance and linear classifiers can achieve unbiasedness for the disbelief label but exhibit poor performance. All three neural classifiers can achieve unbiasedness for the disbelief label. RoBERTa also has the best F_1 scores. The chosen thresholds τ, unbiasedness, binary-, macro-, and micro-F_1 scores under τ for all experimented classifiers on the testing set are shown.

Classifier	**Disbelief**				
	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1
Chance	0.654	✓	0.354	0.494	0.533
LIWC+LR	0.415	✓	0.548	0.647	0.675
ComLex+LR	0.364	✓	0.586	0.683	0.712
BERT	0.374	✓	0.801	0.840	0.850
XLNet	0.514	✓	0.798	0.839	0.850
RoBERTa	0.436	✓	**0.817**	**0.855**	**0.864**

Consider a tweet expressing (dis)belief as label b, then the underlying prevalence $\mathbb{E}(b)$ in the sample is the number of positive labels (TP+FN) divided by the sample size (N). Using a trained classifier to predict b, the estimated prevalence $\mathbb{E}(\hat{b})$ is then the number of predicted positive labels (TP+FP) divided by the sample size (N). An unbiased classifier should make $\mathbb{E}(b) = \mathbb{E}(\hat{b})$, i.e.,

$$\mathbb{E}(b) = \frac{\text{TP}(\tau) + \text{FN}(\tau)}{\text{N}} = \frac{\text{TP}(\tau) + \text{FP}(\tau)}{\text{N}} = \mathbb{E}(\hat{b}), \tag{3.1}$$

and therefore,

$$\text{FP}(\tau) = \text{FN}(\tau). \tag{3.2}$$

To verify unbiasedness, I choose a threshold τ using Equation 3.2 for every classifier from the training set, and then apply the same threshold τ on the testing set and conduct hypothesis tests on Equation 3.2 again. If Equation 3.2, as the null hypothesis, is not rejected, the classifier under threshold τ is unbiased. I use the χ^2 test and set the significance level as $p < 0.01$ after Bonferroni correction.

The final evaluation results for all experimented classifiers are shown in Table 3.2 and Table 3.3. Chance and linear classifiers, with their simple structure, can easily achieve unbiasedness for both disbelief and belief labels. However, this unbiasedness is moot given their poor performance, as I hypothesize that prevalence will shift in the measurement dataset, i.e., if I apply the Chance classifier under the chosen threshold for measurement, the resulting distribution would be the same as my training data, whose distribution is not representative (as discussed in § 3.4.1). For the neural

Table 3.3: **Evaluation results for belief prediction.** Chance and linear classifiers can achieve unbiasedness for the belief label but exhibit poor performance. Only RoBERTa can achieve unbiasedness for the belief label. RoBERTa also has the best F_1 scores. The chosen thresholds τ, unbiasedness, binary-, macro-, and micro-F_1 scores under τ for all experimented classifiers on the testing set are shown.

Classifier	**Belief**				
	Threshold τ	Unbias?	Binary-F_1	Macro-F_1	Micro-F_1
Chance	0.814	✓	0.170	0.490	0.691
LIWC+LR	0.306	✓	0.450	0.666	0.806
ComLex+LR	0.279	✓	0.371	0.612	0.761
BERT	0.646	✗	0.620	0.773	0.877
XLNet	0.593	✗	0.646	0.785	0.877
RoBERTa	0.451	✓	**0.671**	**0.800**	**0.884**

classifiers, all three can achieve unbiasedness for the disbelief label but only RoBERTa can achieve unbiasedness for the belief label. In addition, RoBERTa has the best performance evaluated by F_1 scores, therefore I choose it as the classifier to measure (dis)belief at scale.

3.6 Measuring (Dis)belief via Applying Neural Models

As an application of my classifier, I leverage it to measure (dis)belief at scale and explore my proposed research questions. My measurement study leverages the unlabeled dataset collected in § 3.1 that contains 1,672,687 comments collected from Facebook, 113,687 from Twitter, and 828,000 from YouTube written in response to 5,303 fact-checked claims. These claims are drawn from the entire archive of Snopes and PolitiFact's articles between their founding and January 9, 2018.

The applicability of my trained classifier on this dataset is suggested by **(a)** the same data collection method, i.e., gathering all comments on social media made in response to seed claims identified from fact check articles; and **(b)** the consistent style of informal English language in social media comments. I preprocess the dataset the same way as my experiments, and then feed the dataset to the RoBERTa-based classifier using my chosen τ as the threshold to predict (dis)belief labels on each comment. This process runs within six hours on a single Titan X Pascal GPU.

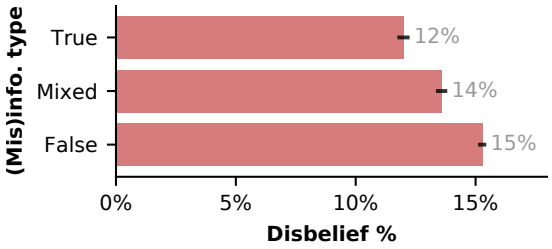

Figure 3.23: **Overall prevalence of expressed disbelief.** For true/mixed/false claims on social media, 12%/14%/15% of comments express disbelief. As the veracity of the claims decreases, the prevalence of expressed disbelief increases.

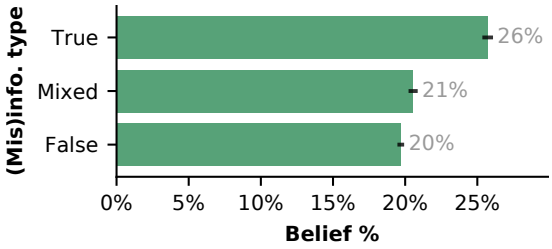

Figure 3.24: **Overall prevalence of expressed belief.** For true/mixed/false claims on social media, 26%/21%/20% of comments express belief. As the veracity of the claims decreases, the prevalence of expressed belief also decreases.

3.6.1 Measuring the Prevalence of (Dis)belief

First, I investigate **RQ1.3**, *do audiences believe in misinformation, and if so, to what extent?* i.e., an estimation of the prevalence of (dis)belief.

This prevalence intuitively varies by the types of (mis)information, therefore I aggregate the veracity of the original claims into three (mis)information types: **(a)** true, if the claims are rated as "true" by Snopes or PolitiFact — these claims contain no misinformation, and their responses were shown to follow distinctive patterns versus others [94]; **(b)** mixed, if the claims are rated as "mostly true", "half true", or "mixed" — these claims contain some misinformation but also some truth; and **(c)** false, if the claims are rated as "mostly false", "false", or "pants on fire!" — these claims contain mostly falsehood.

Next, I aim to estimate the prevalence of (dis)belief in comments in the dataset. However, some of these comments are impacted by a powerful confounding variable: the existence of a fact-check article. To mitigate this, I filter out comments that were posted *after* the corresponding fact-check article was published. Note that, even with this filtering, the remaining comments could still be biased in the claimants distribution.

Finally, I group the remaining comments by the (mis)information type, average their (dis)belief labels (1 if estimated to express (dis)belief and 0 otherwise), and show the results in Figure 3.23 and Figure 3.24.

I observe that as veracity of claims decrease, disbelief increases while belief decreases. As shown in Figure 3.23, I estimate that 12%, 14%, and 15% of comments express disbelief in response to true, mixed, and false claims, respectively; Figure 3.24 shows that 26%, 21%, and 20% of comments express belief in response to true, mixed, and false claims, respectively. These findings suggests that

at least some people commenting on misinformation have the ability to distinguish falsehood, which resonates with the results from existing studies on belief in misinformation [6, 143, 157].

However, the difference in the prevalence of (dis)belief across (mis)information types is relatively small, and for claims that were verified to be true, I estimate that only 26% of comments express belief while 12% express disbelief. One potential explanation for this observation is that the partisan environment drives the public to suspect any claims raised from the opposite ideological group regardless of veracity [74, 75, 84]. Another, though less likely, explanation is that media literacy education equips the public with curiosity to query and doubt all claims, even when the claim is consistent with existing facts [83, 178]. Both explanations are worthy of deeper investigation by future work.

3.6.2 Effects of Time and Fact-Checks on (Dis)belief

RQ1.4, *does the prevalence of expressed (dis)belief in misinformation vary over time?*, and **RQ1.5**, *does the prevalence of expressed (dis)belief in misinformation vary after fact-checking?*, ask for the effects of time and fact-checks. These two questions confound together along the temporal dimension, therefore I investigate them simultaneously. I focus on their effects on false claims, which restricts my analysis to 1,395,293 comments.

To investigate **RQ1.4** and **RQ1.5**, I formulate the following model: I denote a comment as m, its corresponding claim as C_m, its corresponding fact-check for the claim as F_m, and Δ_{e_1,e_2} as the time difference (unit: days) between event e_1 and event e_2 ($\Delta_{e_1,e_2} > 0$ if e_2 happens after e_1). Then, $\Delta_{C_m,m}$ represents the time delay between a comment and its claim, and $\Delta_{F_m,m}$ represents the time delay between a comment and the fact-check of its claim.

Under these notations, the following model captures the linear effects of time and fact-checks:

$$\hat{b} = \beta_0 + \underbrace{\beta_1 \cdot \Delta_{C_m,m}}_{\textbf{RQ1.4}} + \underbrace{\beta_2 \cdot \mathbb{I}_+(\Delta_{F_m,m})}_{\textbf{RQ1.5}} + \epsilon, \tag{3.3}$$

where $\hat{b}$ is the underlying prevalence of (dis)belief estimated by the classifier (defined in § 3.5.3), $\mathbb{I}_+$ is the identity function of positive numbers that returns 1 if the input is positive and 0 otherwise, $\epsilon \sim N(0, \sigma^2)$ is normally distributed noise centered at 0, and β_0, β_1, β_2 are the parameters to be estimated.

This model is similar to the traditional *difference-in-difference* model from causal estimation methods, where the (broadly defined) time variable Δ and the intervention variable $\mathbb{I}$ are regressed

Table 3.4: **Regression results for the effects of time and fact-checks.** OLS is used to estimate parameters for constant effect ($\hat{\beta}_0$), time effect ($\hat{\beta}_1$), and effect of fact-check ($\hat{\beta}_2$) on 1,395,293 comments in response to false information. There is an extremely slight time effect of falsehood awareness, where disbelief increases 0.001% and belief decreases 0.002% per day after the initial claim. Controlling the time effect, disbelief increases 5% and belief decreases 3.4% after a fact-check.

Parameters	**Disbelief**		**Belief**	
	Estimation	p-value	Estimation	p-value
$\hat{\beta}_0$	$+1.52 \times 10^{-1}$	***	$+1.98 \times 10^{-1}$	***
$\hat{\beta}_1$	$+9.96 \times 10^{-6}$	***	-2.19×10^{-5}	***
$\hat{\beta}_2$	$+5.00 \times 10^{-2}$	***	-3.41×10^{-2}	***
# of samples	$1,395,293$		$1,395,293$	

jointly to estimate their respected effects [116]. In my setting, Δ is defined as the time difference between a comment m and its corresponding claim C_m, and $\mathbb{I}$ is a binary variable identified by the time difference between a comment m and its corresponding fact-check F_m.

I use Ordinary Least Square (OLS) to estimate Equation 3.3 for $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$. Here, $\hat{\beta}_0$ represents the constant effects of the underlying initial (dis)belief; $\hat{\beta}_1$ represents the time effect **RQ1.4**, i.e., for every unit of $\Delta_{c_m,m}$, (dis)belief is changed by $\hat{\beta}_1$; $\hat{\beta}_2$ represents the effect of fact-checks **RQ1.5**, i.e., after fact-checks (the threshold of $\mathbb{I}_+$, $\Delta_{F_m,m} > 0$), (dis)belief is changed by $\hat{\beta}_2$.

As shown in Table 3.4, there is an extremely slight time effect, where disbelief increases 0.001% and belief decreases 0.002% per day after the initial false claims. This effect may be caused by social dynamics, where past comments embed the "wisdom of the crowd" at identifying misinformation, which then impacts future users who engage with the claims [103, 216]. Controlling for the time effect, disbelief increases 5% and belief decreases 3.4% after the publication of a fact-check article, which reinforces existing work on the positive effects of fact-checks [65, 78, 213]. Note that although the prevalence of (dis)belief is altered by fact-checks, the mechanism behind such positive effects is still unknown: does the fact-check correct the existing false belief of the same group of users, or does the publication of the fact-check attract a different group of users to comment on the claim with disbelief (therefore altering the overall prevalence)?

3.6.3 Difference of (Dis)belief across Platforms

Finally I look at the difference in (dis)belief across social media platforms. I process the dataset the same way as § 3.6.1, except that here I group data by social media platforms instead of misinformation types.

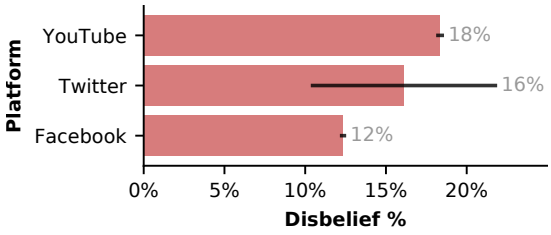

Figure 3.25: **Platforms difference of expressed disbelief.** Facebook comments express less disbelief than YouTube. However, the difference is not significant for Twitter.

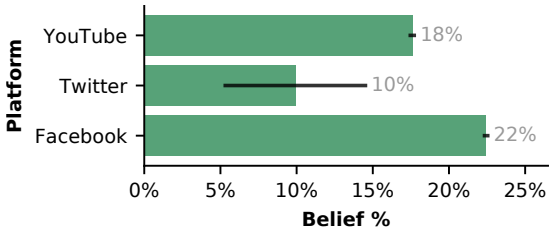

Figure 3.26: **Platforms difference of expressed belief.** Facebook comments express more belief than YouTube, and YouTube comments express more belief than Twitter.

As shown in Figure 3.25 and Figure 3.26, the prevalence of (dis)belief varies across social media platforms. Figure 3.25 shows that for disbelief, Facebook comments express less disbelief than YouTube, while the difference is not significant for Twitter. Figure 3.26 shows that for belief, Facebook comments express more belief than YouTube, whose comments express more belief than Twitter.

Note that this aggregation ignores other confounders, e.g., claim and audience distributions, therefore the result only suggest an overall difference in (dis)belief prevalence across platforms. This reinforces my position (articulated in § 3.4.1) that analyzing Twitter alone is insufficient to represent the misinformation ecosystem.

3.7 Summary of Audiences' Response

In this chapter, I measure audiences' response to misinformation and answered the following 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.

- **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.

There are several limitations of the study in this chapter.

Claimant and topical bias. First, fact-checked claims are, in general, made by high-profile claimants (e.g., political pundits or well-known organizations), therefore excluding claims from the common crowd. There is, to our knowledge, no existing work discussing the relative importance of claims erroneously made (or misinterpreted) by the common crowd in the misinformation ecosystem, therefore I am unable to estimate to what extend this exclusion affects our measurement. Second, most of the articles from Snopes and PolitiFact are focused on politics or political issues, therefore our measurement is also heavily focused on these topics. Other popular misinformation topics, such as health [17] or scientific [52] misinformation, could be less polarized and thus alter the underlying distributions of (dis)belief.

Proxy validity. The use of comments to understand social interaction is common in social media studies. However, a comment may not reflect the true underlying belief of a person. The Hawthorne effect [137] would suggest that social media users are aware of being observed by the public and thus change their behaviors. Social identity [206] and normative influence theory [104] would suggest that a comment could be posted just to cater to the preference of a person's ideological group, instead of capturing their true belief. Additionally, the (dis)belief of people who retweet the claim without commenting are not captured in our approach. Therefore, we emphasize that our study measures *expressed* (dis)belief in the misinformation ecosystem, and our results should be interpreted together with existing qualitative and experimental studies [6, 157].

Bots and likewise. Although comments from bot and bot-like (e.g., the Internet Research Agency (IRA)) users are not cleaned in the dataset, recent studies show that bots mostly spread repeated information rather than commenting [196], and the IRA had very limited commenting activity comparing to the entire Twitter population [88, 238]. We compared our training dataset verses an IRA account dataset released by Twitter and found no overlap [64]. Therefore, the existence of bots should have minimal effects on our results. Note that the limited commenting activity of IRA does not imply limited *impact*, as a comment can influence subsequent comments. That said, comments under such influence, as long as they are from real users, are intended to be captured in our measurement.

This chapter deliver some optimistic results, e.g., increased disbelief and decreased belief as information veracity decrease, (albeit slightly) increased disbelief and decreased belief for false claims over time, a positive effect of fact-checks. However, these results do not undermine the fundamentally concerning consequences of misinformation, especially since we also found some pessimistic results, e.g., considerable suspicion of truthful claims. Despite several notable limitations mentioned above, I hope this work will be a helpful addition to the literature that complements existing qualitative and experimental studies of (mis)information.

Chapter 4

Platforms

In this chapter, I investigate platforms' moderation practice and the following RQs:

- **RQ2.1**, *does the political leaning of a video affect the moderation decision of its comments?*
- **RQ2.2**, *does the extremeness of a video affect the moderation decision of its comments?*
- **RQ2.3**, *does the veracity of content in a video affect the moderation decision of its comments?*
- **RQ2.4**, *does the fact-check of a video affect the moderation decision of its comments?*

To conveniently conduct hypothesis testing, I first reframe each RQ to a null hypothesis that the variable "does not effect" the outcome. Namely, **RQ2.1** - **RQ2.4** becomes:

- **H1a**$_0$, *the political leaning of a video does not affect the moderation decision of its comments.*
- **H1b**$_0$, *the extremeness of a video does not affect the moderation decision of its comments.*
- **H2a**$_0$, *the veracity of content in a video does not affect the moderation decision of its comments.*
- **H2b**$_0$, *the fact-check of a video does not affect the moderation decision of its comments.*

The conceptual framework of these hypotheses is shown in Figure 4.1.

4.1 Platforms' Moderation on Misinformation - an YouTube Dataset

To test hypotheses **H1a**$_0$-**H2b**$_0$, I filter the dataset collected in § 3.1 to 84,068 comments posted on 258 YouTube videos, and link them with labels including *outcome* (was a comment moderated),

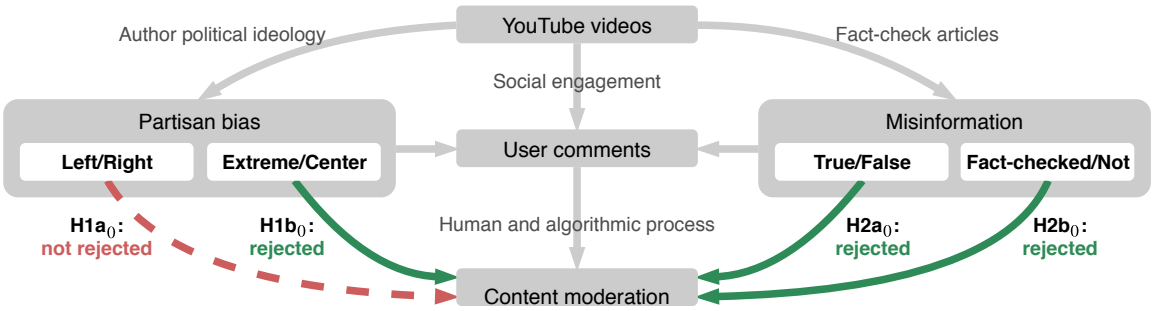

Figure 4.1: **Conceptual framework of four hypotheses.** I investigate the effect of partisanship (i.e., left/right, extreme/center) and misinformation (i.e., true/false, fact-checked/not) on comment moderation. Potential confounders include social engagement on YouTube videos (e.g., views and likes) and linguistics in comments (e.g., hate speech).

treatments (corresponding to four hypothesized variables), and *controls* for confounding variables (i.e., social engagement and the linguistic features of comments). In this section, I describe the data collection and labeling methods with an illustrative example in Figure 4.2.

4.1.1 Moderation Decision - the Outcome Variable

In § 3.1, I crawled Snopes and PolitiFact in January 2018, identified all fact-check articles that linked to posts on social media, including videos on YouTube, and then crawled all the comments attached to these posts. This dataset contains over 2K YouTube videos with 828K comments. Figure 4.2 shows an example article from PolitiFact [204] that fact-checked a YouTube video from Red State Media [185].

To determine whether each comment in the dataset was moderated (1) or not (0), I recrawled all of the YouTube videos in June 2018. I label comments that appeared in the first crawl but not the second as *moderated*. There are two limitations of this labeling method: a) I do not know why or who moderated each comment, and I discuss this limitation more deeply in later sections; and b) my dataset only contains comments that were moderated after January and before June 2018. Figure 4.2 shows four example comments from my dataset, two of which were moderated.

4.1.2 Political Leaning and Extremeness - Treatment Variables

I use two measures for partisanship: its direction (i.e., left (0) or right (1)) for $H1a_0$ and magnitude (i.e., extreme (1) or center (0)) for $H1b_0$ of each video in my dataset. This information is not contained in the original dataset [94]. To gather this information, I leverage partisan scores from previous

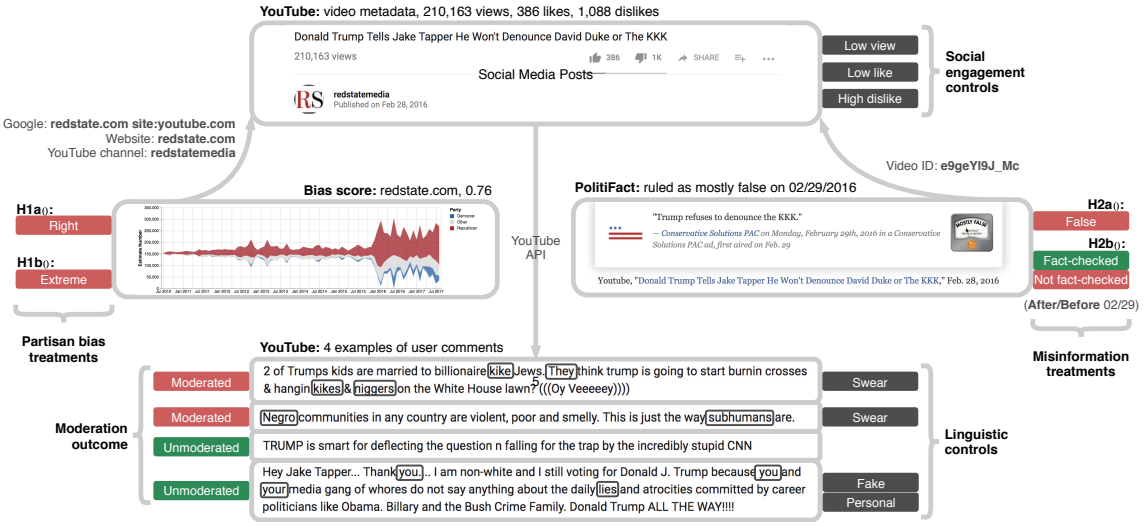

Figure 4.2: **Data collection process and an illustrative example.** Starting from a fact-check article on PolitiFact, I collect the misinformation treatment and a YouTube video ID. Another starting point is the partisan score for the website "redstate.com", where I collect the partisanship treatment and then use Google to get the corresponding channel name. I then use YouTube API to collect the video metadata and link previous data by video ID and channel name respectively. I also collect user comments and labeled their linguistic treatments using *ComLex*. Finally, I compare two crawls to identify moderated comments.

research [191]. In brief, these scores were constructed using a virtual panel of registered US voters. Voters were linked to their Twitter accounts, and then the partisan score of a website was measured by the relative proportion of how it was shared by Democrats and Republicans. This dataset contains scores for 19K websites and the scores range from -1 (shared entirely by Democrats) to 1 (shared entirely by Republicans).

Since the basic unit of my analysis is YouTube videos, not websites, I used Google Search as an intermediary to link a YouTube channel to its website. I entered all 19K website domains as queries into Google Search and added a filter to only return results from the YouTube domain. For each query, I located the first search result containing a link to a YouTube channel (if one existed on the first page of search results), and compared the ID of that channel to the IDs of all channels in my dataset. If I found a match, I associated the partisan score of that website to videos in my dataset from that channel.

Using this process, I were able to associate partisanship labels to 258 YouTube videos from my dataset, originating from 91 unique channels. Example channels include "MacIverInstitute", "John McCain", "BarackObamadotcom", etc. The remaining videos were posted by users and channels that

had little-to-no presence off of YouTube. For direction of partisanship, I label each video as *left* or *right* depending on whether its partisanship score is < 0 or > 0, respectively. Further, for magnitude of partisanship, I labels each video as *extreme* or *center* depending on whether the absolute value of its partisanship score is > 0.5 or < 0.5, respectively.[1]

For example, as shown in Figure 4.2, the partisan score for "redstate.com" is 0.76. I use Google to search the query "redstate.com site:youtube.com" and follow the first link that contains a YouTube channel ID, which leads me to the Red State Media YouTube channel [186]. This enables me to label all Red State Media videos in my dataset as *right* and *extreme*.

4.1.3 Misinformation and Fact-Checks - Treatment Variables

I use two measures for misinformation: the veracity of each video (i.e., true (1) or false (0)) for $H2a_0$ and whether each comments was posted before (0) or after (1) the video was fact-checked for $H2b_0$. The dataset from Jiang and Wilson already contains articles from Snopes and PolitiFact with veracity rulings and timestamp.

I label a video as *true* if the corresponding fact-check article determined that it was true, otherwise I label the video as *false*.[2] For *before/after* labels, I compare the timestamp of each comment to the timestamp of the corresponding fact-check article. The example in Figure 4.2 shows that PolitiFact judged this video to be false on February 29, 2016.

4.1.4 Social Engagement - Control Variables

I also collected social engagement information (i.e., views, likes, and dislikes) as potential controls, e.g., a video with many dislikes could attract more flaggers and therefore cause more moderation. I bin the number of views to an integer in the range 0 (low, $< 25\%$ quantile) to 3 (high, $> 75\%$ quantile) based on quantiles of the view distribution. Similarly, I process likes/dislikes by normalizing them with the number of views to get like/dislike rates per video, then bin them in the same manner as views.[3]

[1] I discuss results using alternative thresholds in later sections.

[2] Thus, my binary veracity label encodes the presence or absence of misinformation in a video, regardless of magnitude. I use a binary encoding for veracity because Jiang and Wilson found that users exhibit significantly different linguistic patterns in comments depending on whether misinformation is present.

[3] This step improves the model performance in later sections. Continuous data are vulnerable to outliers, and number of likes/dislikes without normalization shows high multicollinearity with number of views, i.e., highly viewed videos have more likes and dislikes. (Original data: Spearman $\rho = 0.949^{***}$ for views/likes, and $\rho = 0.887^{***}$ for views/dislikes. After normalization and binning: $\rho = 0.249^{***}$ for views/likes, and $\rho = -0.625^{***}$ for views/dislikes.)

The example video in Figure 4.2 has 210,163 views, 386 likes (0.184% like rate) and 1,088 dislikes (0.518% dislike rate), which I label as *low view* (25% quantile), *low like* (25% quantile), and *high dislike* (75% quantile).

4.1.5 Linguistic Signals - Control Variables

I use a lexicon-based approach to control for the linguistics of each comment, as linguistics are the primary moderation criteria in YouTube's community guidelines [237] and have been found to affect moderation in practice [29, 200].

For this task, I use an existing lexicon called *Comlex* [94] that contains 28 categories (56 subcategories) of human evaluated words extracted from user comments on social media, i.e., the same context as my study. Prior work has found that using contextually appropriate lexicons yields better results than generic ones [120].[4] I apply standard text pre-processing techniques to the comments in my dataset using NLTK [127] (e.g., tokenization, case-folding, and lemmatization) before mapping them into ComLex.

I select eight word categories that significantly ($p < 0.001$) affect moderation likelihood for comments, determined by a preliminary linear regression model:[5] *swear* (including hate speech, e.g., "fuck", "bitch", "nigger"), *laugh* (e.g., "lol", "lmao", "hahaha"), *emoji* (e.g., "😂", "😊", "💩"), *fake* (fake awareness, e.g., "lie", "propaganda", "bias"), *administration* (e.g., "mayor", "minister", "attorney"), *American* (cities and states, e.g., "nyc", "texas", "tx"), *nation* (other nations, e.g., "canada", "mexico", "uk"), and *personal* (e.g., "your", "my", "people's"). I construct eight binary variables for each comment in my dataset; each variable is 1 if the given comment includes a word from that category.

Figure 4.2 shows four examples of user comments under the video. The first comment contains the hate lemmas "kike" and "nigger", therefore it is labeled as *swear*. Similarly, the second contains "negro" and "subhuman" so it is also labeled as *swear*. The last comment contains the lemma "lie" which is a word from the *fake* awareness category, and the lemma "your" and "you" which are from the *personal* category, therefore these variables are 1. All other linguistic variables that contains no words are labeled as 0.

Table 4.1: **Statistics of the YouTube comment dataset.** Mean with 95% confidence intervals after labeling are shown for each measured variable, including the outcome variable, treatment and control variables.

Type	Variable	Value	Mean ± 95% CI
Outcome	Moderated/Not	1/0	0.032 ±0.001
Misinformation	True/False	1/0	0.132 ±0.002
	After/Before Fact-check		0.332 ±0.003
Partisan bias	Right/Left	1/0	0.472 ±0.003
	Extreme/Center		0.716 ±0.003
Engagement	Views	0-3	1.407 ±0.008
	Likes		1.438 ±0.007
	Dislikes		1.411 ±0.008
Linguistic	Swear	1/0	0.102 ±0.002
	Laugh		0.052 ±0.002
	Emoji		0.024 ±0.001
	Fake		0.086 ±0.002
	Administration		0.041 ±0.001
	American		0.022 ±0.001
	Nation		0.016 ±0.001
	Personal		0.239 ±0.003

4.1.6 Overview of Data

Summary statistics are shown in Table 4.1.

4.2 Criteria to Measure Effects

Recent advances in fairness research provide many criteria to measure effects (or bias), each aiming to formalize different desiderata [15]. Most of these criteria characterize the joint or conditional probability between involved variables (e.g., decision, sensitive features), and can be approximately classified to two categories: *independence* and *separation* [87]. In this section, I use $\mathbf{H1a}_0$ (political leaning) as an example to introduce these two criteria, and they apply to $\mathbf{H1b}_0$, $\mathbf{H2a}_0$, and $\mathbf{H2b}_0$ in the same way.

[4] I also applied generic lexicons such as LIWC. I discuss these results in later sections.

[5] This step is designed to select relevant categories. Including all categories would harm the results of my causal model due to overfitting in the logistic regressions to calculate propensity scores.

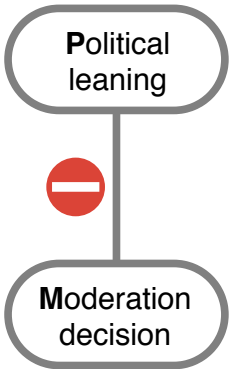

Figure 4.3: **Graph models of the independence criterion.** Null hypothesis $\mathbf{H}_0^{ind}$: M $\perp\!\!\!\perp$ P.

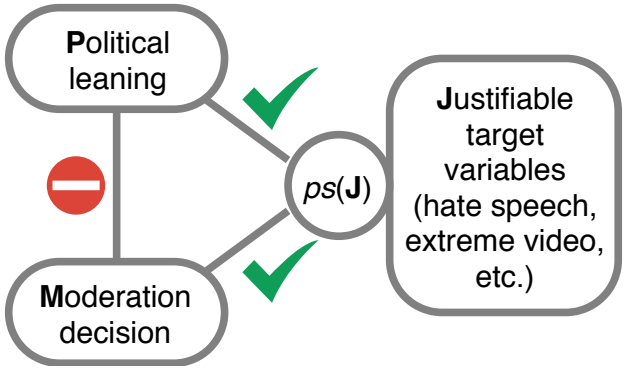

Figure 4.4: **Graph models of the separation criterion.** Propensity scoring function $ps(\mathrm{J})$ is used to summarize J to a scala, hence 2nd null hypothesis $\mathbf{H}_0^{sep}$: M $\perp\!\!\!\perp$ P $\mid ps(\mathrm{J})$.

4.2.1 Independence - a Correlational Criterion

Independence, also referred to as *demographic parity*, is a fairness criterion that requires the decision variable and the sensitive feature to be statistically independent. In the context of political bias and content moderation, an item on social media (e.g., post, comment) can be associated with its political leaning $\mathrm{P} = \{\text{left}, \text{right}\}$ and moderation decision $\mathrm{M} = \{\text{moderated}, \text{alive}\}$. This criterion requires these two variables to satisfy M $\perp\!\!\!\perp$ P, which, given that P is a binary variable, is equivalent to:

$$\mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{left}\} = \mathbb{P}\{\mathrm{M} \mid \mathrm{P} = \text{right}\}. \tag{4.1}$$

Since independence simply describes a correlation between the outcome and the sensitive feature (treatment variable), I refer to it as a *correlational perspective* in the following sections. The graphic model of independence criterion is shown in Figure 4.3. To allege political bias under this criterion, then, requires empirical evidence to reject (4.1) as the null hypothesis $\mathbf{H}_0^{ind}$ with statistical confidence.

Although this criterion is intuitive and has been applied in many studies [**?**, 85], its desirability is context-dependent: e.g., moderation decisions are intended to be made based on the toxicity of content, and if toxicity is unevenly distributed across the political spectrum, the pursuit for independence may be unachievable and even undesirable.

4.2.2 Separation - a Causal Criterion

Separation, also referred to as *equalized odds*, is a type of conditional independence that allows dependence between the decision variable and the sensitive feature, but only to the extend that can be

justified by target variables. For content moderation, such target variables can include hate speech, extreme videos, etc. Denoting a universe of justifiable target variables as J, this criterion requires $\mathrm{M} \perp\!\!\!\perp \mathrm{P} \mid \mathrm{J}$, which, given that P is a binary variable, is equivalent to $\forall$ J:

$$\mathbb{P}\left\{\mathrm{M} \mid \mathrm{P} = \text{left}, \mathrm{J}\right\} = \mathbb{P}\left\{\mathrm{M} \mid \mathrm{P} = \text{right}, \mathrm{J}\right\}. \tag{4.2}$$

This criterion is also widely adopted in previous studies, especially when the correlation between sensitive features and target variables is inherent [10, 111, 215].

A practical limitation of this criterion is that stable estimators of (4.2) requires matched observational pairs conditional on J. Therefore, as J contains more variables, matching becomes more difficult. An alternative method is to summarize all of the target variables into one scalar, i.e., $f : \mathbb{R}^{|\mathrm{J}|} \to \mathbb{R}$. A particular example of f is *propensity scoring* defined as: $ps(\mathrm{J}) := \mathbb{P}\left\{\mathrm{P} = \text{left (or right)} \mid \mathrm{J}\right\}$ [194]. It is proven that if (4.2) holds and $\mathbb{P}\left\{\mathrm{P} \mid \mathrm{J}\right\} \in (0, 1)$, then $\forall\, ps(\mathrm{J})$, $\mathbb{P}\left\{\mathrm{P} \mid ps(\mathrm{J})\right\} \in (0, 1)$ and:

$$\mathbb{P}\left\{\mathrm{M} \mid \mathrm{P} = \text{left}, ps(\mathrm{J})\right\} = \mathbb{P}\left\{\mathrm{M} \mid \mathrm{P} = \text{right}, ps(\mathrm{J})\right\}. \tag{4.3}$$

Separation describes a conditional dependence between the outcome and the treatment variable, and the estimation method, propensity scoring, is referred from causal inference models, therefore I refer to it as a *causal perspective* in the following sections.[6] The graphic model of propensity scored separation criterion is shown in Figure 4.4. To allege political bias under this criterion, then, requires empirical evidence to reject (4.3) as the null hypothesis $\mathbf{H}_0^{sep}$ with statistical confidence.

4.3 Hypothesis Testing on Comment Moderation

In this section, I conduct correlational analysis of my data to investigate the perception of partisan bias in content moderation, and argue that such bias is misperceived.

4.3.1 Independence and Correlational Perception of Effects

I frame the correlational perception of bias as the raw difference in moderation likelihood under each hypothesized variable, i.e., if moderation likelihood under one label (e.g., *right*) is significantly different from its dual (*left*), the corresponding null hypothesis is rejected (correlationally) by my dataset. The moderation likelihood under each hypothesized variable with 95% confidence interval

[6] The word "causal" refer to the causal inference method. The discussion of what constitutes a true causal effect is a philosophical question beyond the scope of this thesis.

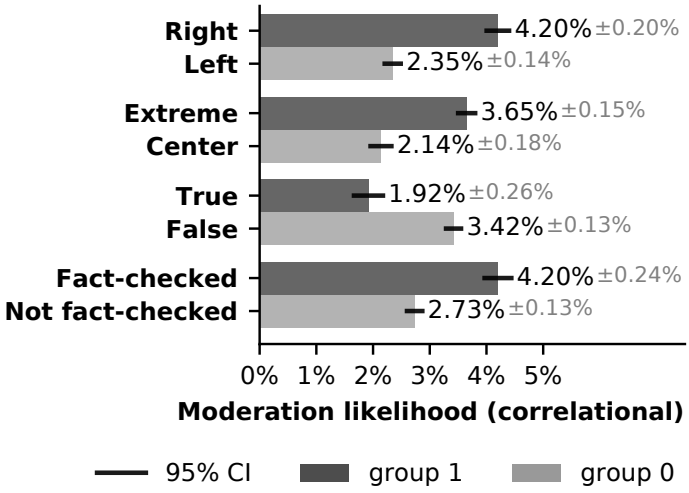

Figure 4.5: **Correlational difference in moderation likelihood.** Moderation likelihood for each group with 95% CI is shown. All four null hypotheses are rejected.

(CI) is shown in Figure 4.5. I perform a χ^2 test on the significance of difference in likelihood between each pair. Under this intuitive, but naïve, perception of bias, all null hypotheses are rejected.

For $H1a_0$, I see that there is a 79% increase in the moderation likelihood on comments from *right*-leaning videos versus *left*-leaning videos, and that the difference is significant ($\chi^2 = 231.0^{***}$). This finding seems to support, at least on the surface level, the claim that content moderation is biased against conservatives [99, 219]. For $H1b_0$, I observe a 71% increase in moderation likelihood from *center* to *extreme* channels, which is also significant ($\chi^2 = 125.2^{***}$). This observation could be caused by YouTube's efforts to monitor extremely partisan channels to prevent hateful content [30, 149, 152].

For $H2a_0$, I find that there is a 44% decrease in the likelihood that comments will be moderated when moving from *false* to *true* videos, and that this difference is significant ($\chi^2 = 69.6^{***}$). Similarly, for $H2b_0$, I observe a 54% increase in moderation likelihood for comments posted after a fact-check on the associated video is available, which is also significant ($\chi^2 = 129.1^{***}$). These findings may be related to YouTube's purported efforts to fight misinformation on their platform [3, 165, 177] by actively partnering with fact-checking organizations [71, 138].

Of course, the correlations I report in Figure 4.5 are potentially specious, since I do not control for correlations between these treatments or with other confounding variables. Therefore, **we do not endorse the findings presented in Figure 4.5.** Rather, I present these results merely to highlight why a person might erroneously believe that comment moderation on YouTube exhibits partisan bias.

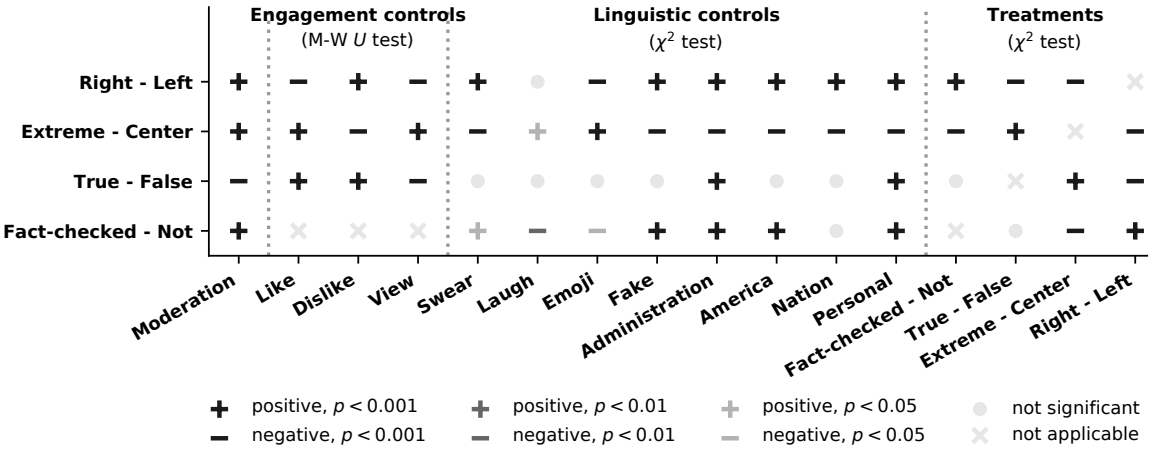

Figure 4.6: **Correlational difference for confounding variables.** The 1^{st} column repeats the observations I made for moderation likelihood. The 2^{nd} to 4^{th} columns show how social engagement correlates with hypothesized variables, the 5^{th} to 12^{th} columns show linguistic features, and 13^{th} to 16^{th} columns show how hypothesized variables correlate with each other. Each "**+**" represents a positive difference in mean and "**–**" a negative one. Significance, as suggested by χ^2 or Mann-Whitney (M-W) U test, is encoded with transparency.

4.3.2 The Problem of Confounding Variables

Comment moderation on YouTube is complicated. As shown in Figure 4.6, there are a set of potential confounding variables that correlate with my hypothesized variables. The 1^{st} column repeats my observations from Figure 4.5. The 2^{nd} to 4^{th} columns show how social engagement on videos correlates with the hypothesized variables, while the 5^{th} to 12^{th} columns show correlations with linguistic features. Finally, the 13^{th} to 16^{th} columns examine correlations between the hypothesized variables themselves. Each "**+**" represents a positive difference in mean and "**–**" a negative one. Significance, calculated using the χ^2 or Mann-Whitney (M-W) U test, is encoded with transparency.[7]

Take H1a_0 as an example. With respect to video-level confounders, *right*-leaning videos have significantly less views ($U = 0.310 \cdot 10^{9***}$) and likes ($U = 0.333 \cdot 10^{9***}$), but significantly more dislikes ($U = 0.408 \cdot 10^{9***}$) than *left*-leaning videos. This provides an alternative explanation for the seeming partisan bias of moderation: the higher dislike rate may result in more flagged comments, thus increasing the likelihood of moderation.

With respect to comment-level linguistics, *right*-leaning videos contain significantly more swear words ($\chi^2 = 671.2^{***}$), fake awareness signals ($\chi^2 = 1013.6^{***}$), discussion on administrative

[7] Since I present 57 independent χ^2 and M-W U tests, I use Bonferroni correction to counteract the problem of multiple hypothesis testing.

matters ($\chi^2 = 778.5^{***}$), references to city/states in America ($\chi^2 = 686.6^{***}$) and other nations ($\chi^2 = 117.1^{***}$), and personal pronouns ($\chi^2 = 423.7^{***}$), but less usage of emojis ($\chi^2 = 524.9^{***}$). This also provides alternative explanations for the seeming partisan bias of moderation: perhaps comments on *right*-leaning videos are more heavily moderated because they include more hate speech.

I also observe that *right*-leaning videos are significantly more likely to be fact-checked ($\chi^2 = 4738.9^{***}$) and *false* ($\chi^2 = 221.8^{***}$) than *left*-leaning videos. This reveals another complication: my hypothesized variables are correlated with each other. This suggests another alternative explanation for H1a_0: that misinformation is the driving force behind moderation, not partisanship.

Some of the correlations in Figure 4.6 are supported by findings from existing research. For example, I find no significant difference in fake awareness signals between *true* and *false* videos ($\chi^2 = 8.4, p = 0.004$), which agrees with previous work on people's inability to identify misinformation [156, 193, 227]. Additionally, I observe that comments posted after fact-checking contain more fake awareness signals ($\chi^2 = 149.7^{***}$), which suggests positive effects of fact-checking on people's expression of political beliefs [62, 176]. However, I also observe more swear word usage ($\chi^2 = 12.8^{*}$) which could be linked to "backfire" effects, where attempts to correct false beliefs makes things worse [159, 231].

To disentangle the effects of my hypothesized variables, I apply a causal model that controls for identified confounding variables. A causal effect is framed as the difference between "what happened" and "what would have happened" [170], e.g., H1a_0 is framed as "what would happen if a left-leaning video changed to right-leaning (while its partisanship magnitude, misinformation level, social engagement, etc. remained the same)". One way to estimate causal effects from observational data is called *matching*. The idea is to find *quasi-experiments* where subjects have similar controls but different treatments, and then compare their outcomes.

Several different matching methods have been proposed for causal inference, such as exact matching, Mahalanobis distance, and propensity scoring [209]. The latter two have been used within the Computer Science community [28,31,61,163]. One shortcoming of exact matching and Mahalanobis distance is that the matching is based on each confounding variable, meaning that the number of matches typically decreases as the number of confounders increases. Therefore, I use a propensity scoring method [194] that has been used extensively in the social [215], psychological [111], and biological [10] literatures.

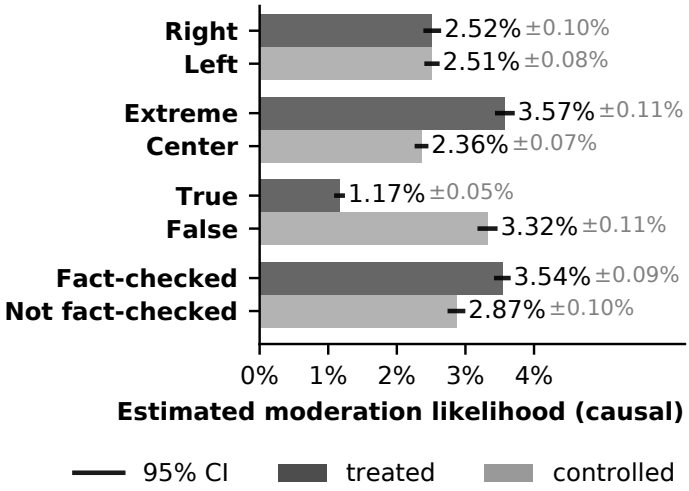

Figure 4.7: **Causal difference in moderation likelihood.** Moderation likelihood for controlled and treated groups with 95% CI is shown. $H1a_0$ is no longer rejected. Differences in the other 3 hypothesized variables are also changed.

4.3.3 Separation and Causal Perception of Effects

As introduced in § 4.2.2, the propensity score is the *probability of getting the treatment label*. It summarizes all of the confounding variables into one scalar. It has been proven that propensity scores are balancing scores, i.e., given a particular propensity score, the distribution of confounders that yield such a score is the same in the treated and controlled groups. Therefore, matching individuals with similar propensity scores mimics a quasi-experiment, at least for measured confounding variables. Additionally, if such an experiment is randomized given a measured set of confounders, then the treatment assignment is also randomized given the propensity scores, which justifies matching based on the propensity score rather than on the full spectrum of confounders (i.e., exact matching and Mahalanobis distance) [194].

For each of my hypotheses, I compute propensity scores using measured confounding variables and the other three hypothesized variables. I then match each treated/controlled sample with its *2-nearest neighbors* based on propensity scores.

Finally, I estimate causal effects, denoted as the Average Treatment Effect (ATE), by averaging the difference in mean for each treated/controlled pair and bootstrap CIs and p-values.

The estimated mean of each hypothesized variable with 95% CI is shown in Figure 4.7, where light (dark) bars represent the controlled (treatment) group. The causal effect estimation with 95%

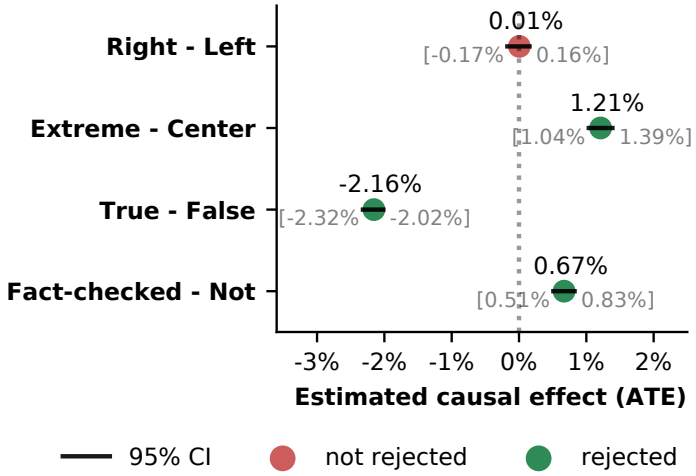

Figure 4.8: **Estimation of causal effect.** Average treatment effect (ATE) with 95% CI is shown. Significance level for null hypothesis is encoded with color. CIs using bootstrap are considered as conservative estimates.

bootstrapped CI[8] is shown in Figure 4.8. I depict $H1a_0$ in red since it is no longer rejected, while I depict the three hypotheses that are still rejected in green.[9]

$H1a_0$ is no longer rejected. In the controlled setting, the estimated moderation likelihood for comments under *left*-leaning videos is $2.51\% \pm 0.08\%$ and under *right*-leaning video is $2.52\% \pm 0.10\%$, which represents an estimated causal effect of 0.01% (95% CI: $[-0.17\%, 0.16\%]$). This difference is not significant ($p = 0.926$). This contradicts the correlational finding from the previous section, and shows that I have no evidence to reject the null hypothesis that comment moderation is not politically biased on average. Instead, this provides empirical evidence that conservative YouTube users and politicians have erroneously assumed that YouTube's moderation practices are biased against them. Rather, rightward political-lean is a proxy for other confounding variables.

$H1b_0$ is still rejected. The estimated moderation likelihood for comments under videos with *center* channels is $2.36\% \pm 0.07\%$ and with *extreme* channels is $3.57\% \pm 0.11\%$, which represents an estimated causal effect of $1.21\%^{***}$ (95% CI: $[1.04\%, 1.39\%]$). This corresponds to a 51% increase, which is smaller than the 71% increase from center to extreme channels I observed in the correlational tests. Regardless, I still find evidence that the magnitude of video partisanship impacts the likelihood of comment moderation. This finding may also partially explain accusations of biased content moderation, since I observe that there are a greater number of ideologically extreme right-leaning

[8] A recent study showed that such CIs are conservative estimates [11].

[9] Note that because I run four hypotheses simultaneously, I use Bonferroni correction to counteract the problem of multiple comparisons, i.e., 95% CIs are actually 98.75% CIs.

channels than similarly extreme left-leaning channels on YouTube.

H2a$_0$ is still rejected. The estimated moderation likelihood for comments under *false* videos is $3.32\% \pm 0.11\%$ and under *true* videos is $1.17\% \pm 0.05\%$, which represents an estimated causal effect of $-2.16\%^{***}$ (95% CI: $[-2.32\%, -2.02\%]$). This corresponds to a 65% decrease, which is larger than the 44% decrease from *false* to *true* videos I observed in the correlational tests, mainly because the estimated moderation likelihood for comments on true videos decreases. In sum, I find evidence that the veracity of videos affects the likelihood of moderation.

H2b$_0$ is still rejected. The estimated moderation likelihood for comments posted *before* fact-checking is $2.87 \pm 0.10\%$ and *after* fact-checking is $3.54\% \pm 0.09\%$, which represents an estimated causal effect of $0.67\%^{***}$ (95% CI: $[0.51\%, 0.83\%]$). This corresponds to a 23% increase, which is smaller than the 54% increase after fact-checking I observed in the correlational tests. This suggests that although confounding variables subsume a large part of the observed correlational difference, I still find evidence that comments are more likely to be moderated after the associated video is fact-checked.

4.4 Alternative Explanations and Robustness Check

Although I analyze my hypotheses within a relatively controlled setting, my analysis is still limited by available datasets and model specifications. In this section, I discuss the limitations and alternative explanations for my results.

4.4.1 Signals and Sources of Moderation

One limitation of my study is my inability to determine who moderated a given comment: the video uploader, a human moderator at YouTube, an algorithm, or the commenter themselves. To address this, I use simulations to investigate how my analysis would change under varying assumptions about the fraction of comments that are removed by commenters themselves. I assume a self-moderation rate r, i.e., the remaining $1 - r$ removed comments were moderated by YouTube's systems. I randomly sample $1 - r$ of the moderated comments in my dataset while keeping the unmoderated comments the same. As shown in Figure 4.9, self-moderation does not change my conclusion for H1a$_0$ for a spectrum of r from 0% to 50%. Although the effect size for H2a$_0$ and H2b$_0$ fluctuate as r increases, the direction of their effects are robust. The only exception is H1b$_0$: the direction of its causal effect does not hold when $r > 20\%$.

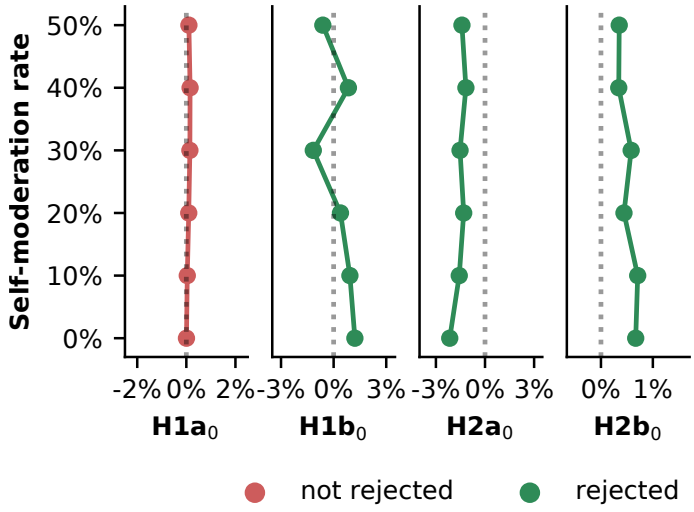

Figure 4.9: **Simulation of user moderation.** The effect of self moderation is minimal for $H1a_0$, $H2a_0$, and $H2b_0$, but $H1b_0$ does not hold under high rates ($r > 20\%$).

Note that this robustness check assumes a constant user moderation rate over all moderated comments, which oversimplifies reality. The moderation behavior of video uploaders and commenters are likely correlated with unmeasured variables, e.g., video uploaders may be more likely to moderate comments that disagree with their own position, either due to direction or extremity of partisanship. Investigating when and why self-moderation happens is beyond my current capabilities, therefore I leave it for future work.

4.4.2 Credibility of Fact-Checkers

My credibility labels are drawn from Snopes and PolitiFact, which are both confirmed by the International Fact-Checking Network to be non-partisan, fair, and transparent [180]. However, there are still accusations that their ratings are biased against political conservatives [153, 188, 197]. Although I do observe that right-leaning videos are more likely to be rated as false ($\chi^2 = 221.8^{***}$), I do not know if the political leaning actually causes this difference.

Exploring the bias of ratings from fact-checkers themselves is beyond the scope of this paper, but still, I investigate the hypothetical case where fact-checkers are systematically biased. I assume a bias b, where $b = \lambda$L represents a systematic bias against liberals and $b = \lambda$R represents bias against conservatives, and λ represents the magnitude of bias (0, non-existing; +1, slight; +2, high). I recalibrate all the veracity scores in my dataset given a value for b. For example, $b = +1$R

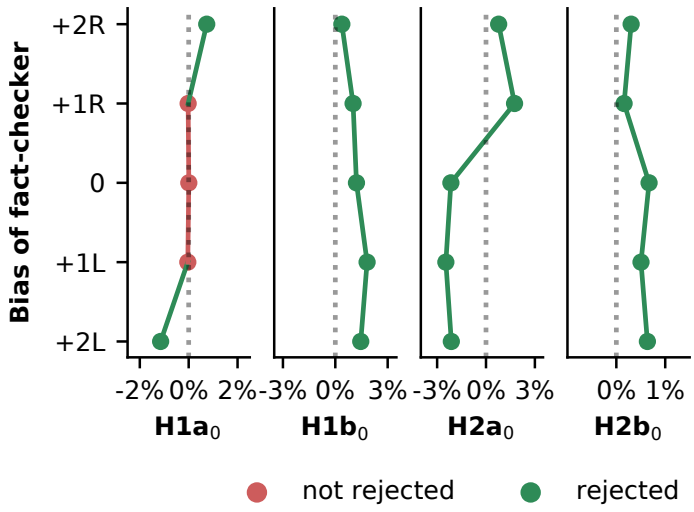

Figure 4.10: **Simulation of biased fact-checkers.** The effect of fact-checker bias is minimal for $H1b_0$ and $H2b_0$, and minimal for $H1a_0$ when bias is low ($\lambda \leq +1$).

represents a slight bias against conservatives, which I consider as a form of underrating right-leaning videos. Therefore, all conservative videos labeled as "mostly true" by the fact-checker will instead be considered true. Similarly, if $b = +2$R, then all conservative videos labeled as "half true" or "mostly true" by the fact-checker will instead be considered true.

The results of my causal models under various values of b are shown in Figure 4.10. $H2a_0$ is impacted the most, since it directly concerns video veracity. In contrast, the effect sizes of $H1b_0$ and $H2b_0$ fluctuate, but the direction of their effects are robust. For $H1a_0$, the result does not change with slight bias ($\lambda \leq +1$), but does change when fact-checkers are highly biased. Consider $b = +2$R, which means fact-checkers are highly biased against right-leaning videos: in the calibrated case, content moderation is also biased against right-leaning videos. *Vice versa* for $b = +2$L. Similarly, the results of $H2a_0$ also change in the same direction.

Note that **we do not support claims of bias against fact-checkers in any way**. I investigate this hypothetical scenario simply for the sake of thoroughness, i.e., to show that even if fact-checkers were slightly biased, it would not explain why comments on right-leaning videos are moderated more heavily than comments on left-leaning videos.

4.4.3 Alternative Thresholds and Control Variables

I now explore model dynamics under alternative thresholds and controls for my labels.

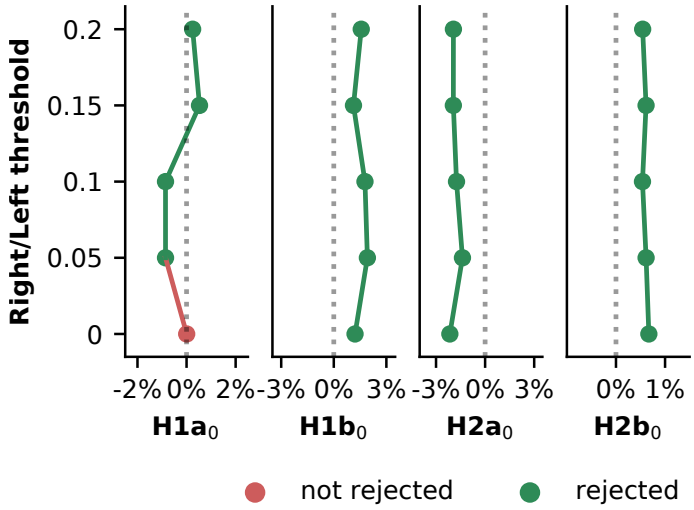

Figure 4.11: **Alternative $H1a_0$ (left/right) thresholds.** The effect of left/right thresholds is minimal for $H1b_0$, $H2a_0$ and $H2b_0$, but results for $H1a_0$ do not hold.

First, my label for right- and left-leaning video channels is based on the sign of partisanship score. However, it is conceivable that scores near zero may not indicate perceptible partisanship [191]. Therefore, I set a minimum threshold for partisanship scores, i.e., only absolute scores greater than the threshold are labeled right/left, others are considered neutral and not used for analysis. As shown in Figure 4.11, such thresholding has minimal impact on $H1b_0$, $H2a_0$, and $H2b_0$, but does impact $H1a_0$.[10] However, since the effect fluctuates between leftward and rightward bias, the claim for "conservative bias" is still not supported overall.

Next, I investigate how alternative thresholds for extreme/center labels affects my results by replacing my original threshold 0.5 with a spectrum from 0.3 to 0.7. As shown in Figure 4.12, this change has minimal impact on all hypotheses with two exceptions. a) I observe leftward bias for $H1a_0$ under threshold 0.3; although this bias is statistically significant, the difference is only 0.37% which yields minimal practical impact. b) The bias flips for $H1b_0$ under threshold 0.7, but this is caused by poor model performance since such extremely partisan video channels are rare in my dataset (leading to a sample of < 1000 moderated comments).

Third, I examine an alternative set of linguistic controls using LIWC [171, 214]. Although the ComLex lexicon is context-specific, it has not been as extensively used as LIWC. I derived five categories from LIWC: *swear, money, work, biological process,* and *punctuation*,[11], use them in

[10] This is partially due to the partisan bias scores of comments in my dataset not being balanced between left and right.
[11] Determined by a preliminary linear regression for $p < 0.001$.

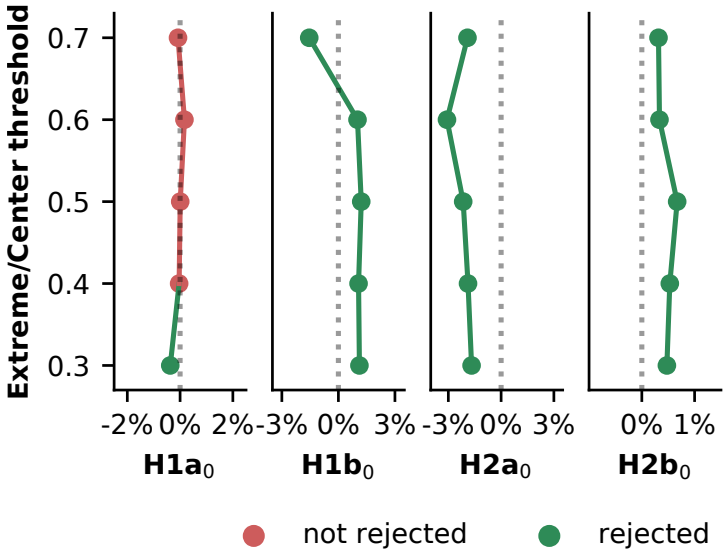

Figure 4.12: **Alternative $H1b_0$ (extreme/center) thresholds.** The effect of extreme/center thresholds is minimal for most hypotheses, except for $H1a_0$ and $H1b_0$.

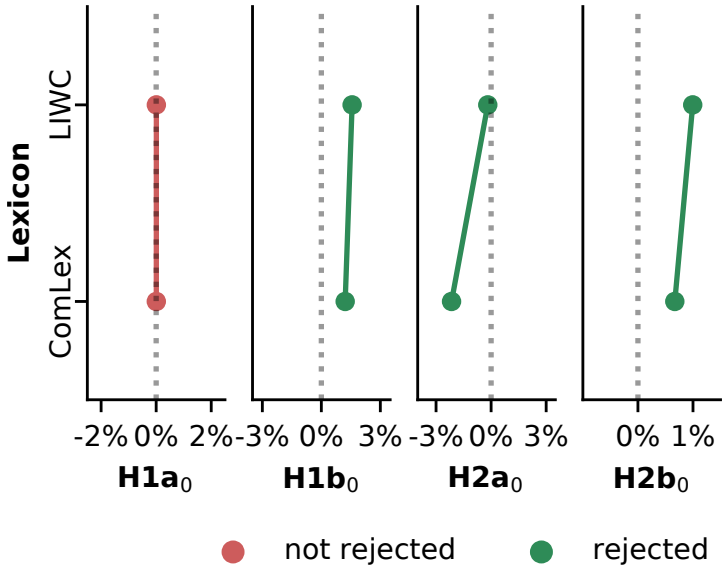

Figure 4.13: **Alternative linguistic controls.** The effect of alternative linguistic controls using lexicon LIWC instead of ComLex is minimal for all hypotheses.

place of the linguistic controls from ComLex, and rerun my model. As shown in Figure 4.13, the difference between using ComLex and LIWC is minimal for all hypotheses.

4.5 Summary of Platforms Moderation

In this chapter, I investigate platforms' moderation practice using YouTube as a lens and answered 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.

There are several limitations of the study in this chapter, besides the ones mentioned in § 4.4.

Concerns regarding causal models. There are two main concerns when using causal models. The first is *reverse causality* [135], which refers to the case where the direction of a causal effect may be the opposite of what is assumed, or the causal effect is a two-way relationship. Reverse causality does not apply to my study, since in my dataset the outcome variable (comment moderation) comes strictly after a video is posted, when all my hypothesized variables are already determined. Another concern is *unmeasured confounding variables* [192], which refers to factors that might affect the outcome and correlate with treatments but are not controlled in the model. My controlled confounders include social engagement with YouTube videos and linguistics in user comments, which are intuitive and highly relevant given YouTube's community guidelines [237] and prior studies [29, 94, 200]. However, this set is admittedly incomplete; unmeasured factors such as user characteristics, comment volume, the presence of "bots," etc., could still skew the results of propensity scoring models [105]. Nevertheless, the results from propensity scoring show significant improvement comparing to correlational analysis [40]. Again, although causal models analyze relationships between treatments (i.e., hypotheses) and outcome (i.e., moderation), they do not explain intermediate factors. For example, it could be that extreme partisanship and high-level misinformation directly affect the

attention and decision-making of algorithmic or human moderators [3, 30, 149, 152, 165, 177]. Or it could be that fact-check messages draw more efforts from concerned users to flag content for moderation [71, 138].

Representation and generalization. The YouTube videos in my dataset are covered by the datasets from [94] and [191], which means they were published by identifiable entities that have web presences off YouTube, and were influential enough to draw the attention of fact-checkers. In other words, the videos in my study are higher-profile than average on YouTube. Measured by number of views, my sample of YouTube videos has a mean of 4,311,320 $\pm$ 38,942 views, which is significantly higher than the average views measured by previous studies [33, 59, 142]. Thus, my findings may not be representative across all videos on YouTube. That said, the vast majority of videos on YouTube receive very few views and comments, meaning they are not viable or interesting candidates for study. Instead, by focusing on high-profile videos, I present results that I believe are more relevant to the YouTube community and policymakers. I use YouTube as a lens to investigate comment moderation as I believe that this is a vitally important endeavor at this moment in time, given the prevailing political climate. That said, I caution that my findings may not generalize beyond YouTube. Further, platform moderation policies are notoriously fickle, meaning that my findings may not generalize over time.

My study advances the call for researchers to engage with issues of societal and political importance, especially as they pertain to a healthy web and concerns of partisan bias and free speech [26, 28, 45, 46, 113]. The major design implication stemming from my findings concerns the non-transparent deletion of comments on YouTube. Opaque moderation practices, regardless of whether they are fully or semi-automated, are a breeding ground for theories like the one we've refuted here – anti-conservative bias in moderation practices. Indeed, this is both a motivation of my study and one of the limitations of my dataset: there is no record of when, why, or by who a comment was deleted. Although moderation is absolutely a critical component of healthy social media systems [23], platform providers should consider designs that are more constructive and transparent.

Towards this goal, I recommend that deleted comments be preserved and protected. That is, comments are still moderated under existing policies, but the original comment is hidden behind a notification that it has been moderated. Then, if a user or researcher is interested in what was moderated and why, they can click on the notification to view the original comment alongside the specific policy violations that caused it to be moderated. Additional meta-information could also be

provided about who moderated the comment – the platform or the channel owner – and whether the comment was flagged by automated systems. This design serves two purposes. First, it would give the commenter an explanation for why their comment was deleted and provide them with feedback on how to improve their discourse. Second, because the comment and its policy violations are preserved, it provides transparency and feedback to the community at large. This transparency, in turn, may discourage public figures from making false claims about why comments were moderated, since external researchers will have the ability to fact check such claims and mitigate the damage done to the platform in terms of user trust [232].

The second benefit (transparency) could negate the first (feedback), however, if the user who posted the deleted comment is exposed to the community: the user may be shamed into no longer participating or worse [106, 107]. Instead of learning how to be civil, they may simply go elsewhere. For example, researchers found that Reddit's ban of two hate speech subreddits was effective in reducing overall hate speech usage on the site, but noted that this ban had simply "made these users (from banned subreddits) *someone else's problem*" and "likely did not make the internet safer or less hateful" [28]. To avoid this outcome, the comment should be preserved, but the offending user should be anonymized. The goal of this design is to educate, to give a human being the opportunity to learn, not to exclude. Further research is needed to investigate how this may play out in practice.

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] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 579–586. Association for Computational Linguistics, 2005.

[3] E. Alvarez. Youtube ceo talks misinformation, creators and comments at sxsw. Engadget, 3 2018.

[4] M. A. Amazeen. Revisiting the epistemology of fact-checking. *Critical Review*, 27(1):1–22, 2015.

[5] M. A. Amazeen. Checking the fact-checkers in 2008: Predicting political ad scrutiny and assessing consistency. *Journal of Political Marketing*, 15(4):433–464, 2016.

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

[7] 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.

[8] A. Arif, J. J. Robinson, S. A. Stanek, E. S. Fichet, P. Townsend, Z. Worku, and K. Starbird. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 155–168. ACM, 2017.

[9] S. E. Asch and H. Guetzkow. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, pages 222–236, 1951.

[10] P. C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12):2037–2049, 2008.

[11] P. C. Austin and D. S. Small. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*, 33(24):4306–4319, 2014.

[12] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2010)*, volume 10, pages 2200–2204, 2010.

[13] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[14] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *IJCAI*, volume 7, pages 2670–2676, 2007.

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

[16] L. Becker, G. Erhart, D. Skiba, and V. Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 333–340, 2013.

[17] A. J. Berinsky. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, 47(2), 2017.

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

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

[20] 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.

[21] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.

[22] S. Brown. Likert scale examples for surveys, 2010.

[23] C. Buni and S. Chemaly. The secret rules of the internet. The Verge, Apr. 2016.

[24] C. Burfoot and T. Baldwin. Automatic satire detection: Are you having a laugh? In *Proc. of ACL*, 2009.

[25] C. J. Calhoun. *Social Theory and the Politics of Identity*. 1994.

[26] S. Chancellor and S. Counts. Measuring employment demand using internet search data. In *Proc. of CHI*, page 122, 2018.

[27] S. Chancellor, J. A. Pater, T. A. Clear, E. Gilbert, and M. De Choudhury. #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proc. of CSCW*, 2016.

[28] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *PACM on HCI*, 1(CSCW):1–22, Dec. 2017.

[29] 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.

[30] R. Chatterjee and P. Dave. Youtube set to hire more staff to review extremist video content. Independent, 12 2017.

[31] L. Chen, R. Ma, A. Hannák, and C. Wilson. Investigating the impact of gender on rank in resume search engines. In *Proc. of CHI*, page 651, 2018.

[32] J. Cheng, M. Bernstein, C. Danescu-Niculescu-Mizil, and J. Leskovec. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1217–1230. ACM, 2017.

[33] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Proc. of IWQoS*, 2008.

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

[35] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193, 2015.

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

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

[38] 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.

[39] Z. Dai, Z. Yang, Y. Yang, W. W. Cohen, J. Carbonell, Q. V. Le, and R. Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proc. of ACL*, 2019.

[40] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proc. of CHI*, pages 2098–2110, 2016.

[41] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi. Echo chambers in the age of misinformation. *arXiv preprint arXiv:1509.00189*, 2015.

[42] 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.

[43] N. Diakopoulos and M. Naaman. Towards quality discourse in online news comments. In *Proc. of CSCW*, 2011.

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

[45] 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.

[46] R. Epstein, R. E. Robertson, D. Lazer, and C. Wilson. Suppressing the search engine manipulation effect (SEME). *PACM on HCI*, 1(CSCW):1–22, Dec. 2017.

[47] A. Esuli and F. Sebastiani. Sentiwordnet: a publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, pages 417–422, 2006.

[48] Facebook. Community standards, 2018.

[49] Facebook. Investments to fight polarization, 2020.

[50] D. I. H. Farías, V. Patti, and P. Rosso. Irony detection in twitter: The role of affective content. *ACM ToIT*, 16(3), 2016.

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

[52] J. Farrell, K. McConnell, and R. Brulle. Evidence-based strategies to combat scientific misinformation. *Nature Climate Change*, 2019.

[53] E. Fast, B. Chen, and M. S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4647–4657. ACM, 2016.

[54] E. Fast, T. Vachovsky, and M. S. Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proc. of ICWSM*, pages 112–120, 2016.

[55] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to pretrain any-domain models for detecting emotion, sentiment and sarcasm. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'17)*, Copenhagen, Denmark, 9 2017.

[56] 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.

[57] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016.

[58] S. Fiegerman. Facebook, google, twitter to fight fake news with 'trust indicators', 2017.

[59] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto. On the dynamics of social media popularity: A youtube case study. *ACM ToIT*, 14(4):24, 2014.

[60] 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.

[61] E. Foong, N. Vincent, B. Hecht, and E. M. Gerber. Women (still) ask for less: Gender differences in hourly rate in an online labor marketplace. *PACM on HCI*, 2(CSCW):53, 2018.

[62] 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.

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

[64] V. Gadde and Y. Roth. Enabling further research of information operations on twitter. *Twitter Blog*, 17, 2018.

[65] R. K. Garrett, E. C. Nisbet, and E. K. Lynch. Undermining the corrective effects of media-based political fact checking? the role of contextual cues and naïve theory. *Journal of Communication*, 63(4):617–637, 2013.

[66] M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57(3), 2019.

[67] 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.

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

[69] E. Gilbert, C. Lampe, A. Leavitt, K. Lo, and L. Yarosh. Conceptualizing, creating, & controlling constructive and controversial comments: A cscw research-athon. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 425–430. ACM, 2017.

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

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

[72] R. González-Ibánez, S. Muresan, and N. Wacholder. Identifying sarcasm in twitter: a closer look. In *Proc. of HLT*, 2011.

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

[74] 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.

[75] 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.

[76] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, pages 153–164. SIAM, 2012.

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

[78] A. Hannak, D. Margolin, B. Keegan, and I. Weber. Get back! you don't know me like that: The social mediation of fact checking interventions in twitter conversations. In *ICWSM*, 2014.

[79] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proc. of CSCW*, pages 1914–1933, 2017.

[80] K. S. Hasan and V. Ng. Stance classification of ideological debates: Data, models, features, and constraints. In *Proc. of IJCNLP*, 2013.

[81] N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu. The quest to automate fact-checking. *world*, 2015.

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

[83] R. Hobbs and A. Jensen. The past, present, and future of media literacy education. *Journal of media literacy education*, 1(1), 2009.

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

[85] D. Hu, S. Jiang, R. E. Robertson, and C. Wilson. Auditing the partisanship of google search snippets. In *Proc. of WWW*, 2019.

[86] Y. L. Huang, K. Starbird, M. Orand, S. A. Stanek, and H. T. Pedersen. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 969–980. ACM, 2015.

[87] B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proc. of FAT**, 2019.

[88] J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, and E. Gilbert. Still out there: Modeling and identifying russian troll accounts on twitter. *arXiv*, 2019.

[89] B. Jackson. Factcheck, 2018.

[90] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. Online harassment and content moderation: The case of blocklists. *ACM ToCHI*, 25(2):1–33, Mar. 2018.

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

[92] S. Jiang, L. Chen, A. Mislove, and C. Wilson. On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi. In *Proc. of WWW*, 2018.

[93] 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.

[94] 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.

[95] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *Data Mining (ICDM), 2014 IEEE International Conference on*, pages 230–239. IEEE, 2014.

[96] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *AAAI*, pages 2972–2978, 2016.

[97] K. Joseph, L. Friedland, W. Hobbs, D. Lazer, and O. Tsur. Constance: Modeling annotation contexts to improve stance classification. In *Proc. of EMNLP*, 2017.

[98] D. Jurafsky and J. H. Martin. *Speech and language processing*, volume 3. Pearson London:, 2014.

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

[100] 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.

[101] L. K. Kaye, S. A. Malone, and H. J. Wall. Emojis: Insights, affordances, and possibilities for psychological science. *Trends in cognitive sciences*, 21(2):66–68, 2017.

[102] M. W. Kearney. Trusting news project report. *Reynolds Journalism Institute*, 7 2017.

[103] 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. In *Proc. of WSDM*, 2018.

[104] D. L. Kincaid. From innovation to social norm: Bounded normative influence. *Journal of health communication*, 9(S1), 2004.

[105] G. King and R. Nielsen. Why propensity scores should not be used for matching. 2016.

[106] K. Klonick. Re-shaming the debate: Social norms, shame, and regulation in an internet age. *SSRN Electronic Journal*, 2015.

[107] K. Klonick. The new governors: The people, rules, and processes governing online speech. (ID 2937985), 2017.

[108] T. Kriplean, C. Bonnar, A. Borning, B. Kinney, and B. Gill. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1188–1199. ACM, 2014.

[109] 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.

[110] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proc. of AAAI*, 2015.

[111] S. T. Lanza, J. E. Moore, and N. M. Butera. Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American journal of community psychology*, 52(3-4):380–392, 2013.

[112] B. Laslett. Unfeeling knowledge: Emotion and objectivity in the history of sociology. In *Sociological Forum*, volume 5, pages 413–433. Springer, 1990.

[113] J. Lazar, J. Abascal, S. Barbosa, J. Barksdale, B. Friedman, J. Grossklags, J. Gulliksen, J. Johnson, T. McEwan, L. Martínez-Normand, et al. Human–computer interaction and international public policymaking: a framework for understanding and taking future actions. *Foundations and Trends® in Human–Computer Interaction*, 9(2):69–149, 2016.

[114] 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.

[115] 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.

[116] M. Lechner et al. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3), 2011.

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

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

[119] 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.

[120] M. Li, Q. Lu, and Y. Long. Are manually prepared affective lexicons really useful for sentiment analysis. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 146–150, 2017.

[121] Q. V. Liao and W.-T. Fu. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proc. of CHI*, pages 2359–2368, 2013.

[122] Q. V. Liao and W.-T. Fu. Can you hear me now?: mitigating the echo chamber effect by source position indicators. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 184–196. ACM, 2014.

[123] Q. V. Liao and W.-T. Fu. Expert voices in echo chambers: effects of source expertise indicators on exposure to diverse opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2745–2754. ACM, 2014.

[124] K. W. Lim and W. Buntine. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1319–1328. ACM, 2014.

[125] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin': Evolution of twitter users and behavior. In *Proc. of ICWSM*, 2014.

[126] 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.

[127] E. Loper and S. Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, volume 1, pages 63–70, 2002.

[128] X. Lu, W. Ai, X. Liu, Q. Li, N. Wang, G. Huang, and Q. Mei. Learning from the ubiquitous language: an empirical analysis of emoji usage of smartphone users. In *Proceedings of the*

2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pages 770–780. ACM, 2016.

[129] 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.

[130] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

[131] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11):2579–2605, 2008.

[132] A. Magdy and N. Wanas. Web-based statistical fact checking of textual documents. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 103–110. ACM, 2010.

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

[134] D. B. Margolin, A. Hannak, and I. Weber. Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, pages 1–24, 2017.

[135] G. S. Marquis, J.-P. Habicht, C. F. Lanata, R. E. Black, and K. M. Rasmussen. Association of breastfeeding and stunting in peruvian toddlers: an example of reverse causality. *International journal of epidemiology*, 26(2):349–356, 1997.

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

[137] R. McCarney, J. Warner, S. Iliffe, R. Van Haselen, M. Griffin, and P. Fisher. The hawthorne effect: a randomised, controlled trial. *BMC Medical Research Methodology*, 7(1), 2007.

[138] H. McCracken. Youtube will use wikipedia to fact-check internet hoaxes. Fast Company, 3 2018.

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

[140] D. Mikkelson. Snopes, 2018.

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

[142] J. M. Miotto and E. G. Altmann. Predictability of extreme events in social media. *PLoS One*, 9(11):e111506, 2014.

[143] A. Mitchell, J. Gottfried, J. Kiley, and K. E. Matsa. Political polarization & media habits. *Pew Research Center*, 21, 2014.

[144] S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[145] S. M. Mohammad and P. D. Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[146] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[147] 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.

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

[149] D. Z. Morris. Hate speech: Youtube restricts extremist videos. Fortune, 8 2017.

[150] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Proc. of ICWSM*, 2013.

[151] S. Murray and C. Lima. Trump accuses social media giants of 'silencing millions of people'. Politico, 2018.

[152] S. News. New youtube recruits to monitor online extremist propaganda 'wrong approach'. Sputnik International, 6 2017.

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

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

[155] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

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

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

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

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

[160] 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.

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

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

[163] A. Olteanu, O. Varol, and E. Kiciman. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. of CSCW*, 2017.

[164] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 309–319. Association for Computational Linguistics, 2011.

[165] V. Palladino. Youtube to fight fake news with links to real news and context. Ars Technica, 7 2018.

[166] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE TKDE*, 22(10), 2009.

[167] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135, 2008.

[168] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, volume 10, pages 79–86. Association for Computational Linguistics, 2002.

[169] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, pages 369–374. ACM, 2016.

[170] J. Pearl. *Causality*. Cambridge university press, 2009.

[171] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

[172] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[173] G. Pennycook and D. G. Rand. Assessing the effect of "disputed" warnings and source salience on perceptions of fake news accuracy. 2017.

[174] R. Plutchik. Emotions: A general psychoevolutionary theory. *Approches to emotion*, pages 197–219, 1984.

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

[176] 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.

[177] N. Y. Post. Youtube committing $25m to fight fake news. New York Post, 7 2018.

[178] W. J. Potter. *Media literacy*. Sage Publications, 2018.

[179] 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.

[180] Poynter. Verified signatories of the ifcn code of principles. Poynter, 2018.

[181] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 729–740, 2017.

[182] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599. Association for Computational Linguistics, 2011.

[183] L. Qiu, H. Lin, J. Ramsay, and F. Yang. You are what you tweet: Personality expression and perception on twitter. *Journal of Research in Personality*, 46(6):710–718, 2012.

[184] 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.

[185] RedStateMedia. Donald trump tells jake tapper he won't denounce david duke or the kkk. YouTube, 2 2016.

[186] RedStateMedia. Youtube homepage for redstatemedia. YouTube, 2018.

[187] R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

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

[189] M. A. Riordan. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology*, 36(5):549–567, 2017.

[190] R. E. Robertson, S. Jiang, K. Joseph, L. Friedland, D. Lazer, and C. Wilson. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM: Human-Computer Interaction*, 2(CSCW), 11 2018.

[191] 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.

[192] J. M. Robins, A. Rotnitzky, and D. O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. 2000.

[193] 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.

[194] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[195] 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.

[196] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. The spread of misinformation by social bots. *arXiv*, 2017.

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

[198] A. Sharockman. Politifact, 2018.

[199] 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.

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

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

[202] B. Shi and T. Weninger. Fact checking in heterogeneous information networks. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 101–102. International World Wide Web Conferences Steering Committee, 2016.

[203] 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.

[204] R. Snyder. Pro-rubio super pac ad tying trump to kkk misses the mark. PolitiFact, 2 2016.

[205] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove. Potential for discrimination in online targeted advertising. In *Proc. of FAT**, pages 5–19, 2018.

[206] J. E. Stets and P. J. Burke. Identity theory and social identity theory. *Social psychology quarterly*, 2000.

[207] P. J. Stone, D. C. Dunphy, and M. S. Smith. The general inquirer: A computer approach to content analysis. 1966.

[208] C. T. Street and K. W. Ward. Improving validity and reliability in longitudinal case study timelines. *European Journal of Information Systems*, 21(2), 2012.

[209] E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1), 2010.

[210] R. Suarez and K. Flynn. Facebook, twitter issue policy changes to manage fake news and hate speech, 2017.

[211] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro. Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*, 2017.

[212] H. Tafjel and J. C. Turner. The social identity theory of intergroup behavior. *Psychology of intergroup relations*, pages 7–24, 1986.

[213] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proc. of WWW*, 2015.

[214] Y. R. Tausczik and J. W. Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54, 2010.

[215] F. J. Thoemmes and E. S. Kim. A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, 46(1):90–118, 2011.

[216] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause. Fake news detection in social networks via crowd signals. In *Companion Proc. of WWW*, 2018.

[217] Twitter. Rules and policies, 2018.

[218] Twitter. About different types of tweets, 2020.

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

[220] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017.

[221] G. Veletsianos, R. Kimmons, R. Larsen, T. A. Dousay, and P. R. Lowenthal. Public comment sentiment on educational videos: Understanding the effects of presenter gender, video format, threading, and moderation on YouTube TED talk comments. *PLoS One*, 13(6), June 2018.

[222] 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.

[223] 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.

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

[225] 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.

[226] X. Wang, C. Yu, S. Baumgartner, and F. Korn. Relevant document discovery for fact-checking articles. In *The 2018 Web Conference Companion (WWW 2018 Companion)*, Lyon, France, 4 2018.

[227] 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.

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

[229] 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.

[230] 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.

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

[232] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proc. of CHI*, page 656, 2018.

[233] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7):589–600, 2014.

[234] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1), 2010.

[235] D. Yang, A. Lavie, C. Dyer, and E. Hovy. Humor recognition and humor anchor extraction. In *Proc. of EMNLP*, 2015.

[236] 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.

[237] YouTube. Community guidelines, 2018.

[238] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proc. of WebSci*, 2019.

[239] A. X. Zhang, M. Igo, M. Facciotti, and D. Karger. Using student annotated hashtags and emojis to collect nuanced affective states. In *Proceedings of the Fourth (2017) ACM Conference on Learning@Scale*, pages 319–322. ACM, 2017.

[240] X. Zhou, X. Wan, and J. Xiao. Attention-based lstm network for cross-lingual sentiment classification. In *Proc. of EMNLP*, 2016.