

**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**

June 2021

To my family.

Contents

List of Figures	v
List of Tables	ix
Acknowledgments	x
Abstract of the Dissertation	xi
1 Introduction	1
1.1 Audiences' Response	2
1.2 Platforms' Moderation	5
1.3 Storytellers' Strategies	7
1.4 Outline	9
2 Background	10
2.1 Misinformation and Its Consequences	10
2.1.1 Foundations of Misinformation	10
2.1.2 Fact-Checking as an Intervention	12
2.1.3 Belief and Disbelief in Misinformation	13
2.2 Content Moderation and Its Controversy	14
2.2.1 Platforms and Community Guidelines	15
2.2.2 Effects of Content Moderation	15
2.2.3 Bias of Human and Algorithms	16
2.3 Manipulation Strategies and Its Complication	17
2.3.1 Theorized Misinformation Types	17
2.3.2 Evolution of Misinformation Stories	17
2.4 Natural Language Processing for Social Science	18
2.4.1 Bag-of-Words and Lexicons	18
2.4.2 Sequence Classification and Neural Models	19
2.4.3 Interpretability, Explainability and Rationalization	19
3 Audiences	21
3.1 Audiences' Comments to Misinformation — an Unlabeled Dataset	21
3.1.1 Data Collection from Fact-Checks and Social Media	22

3.1.2	Overview of Fact-Checked Claims and Comments	23
3.2	Lexicon Construction for Linguistic Signals	24
3.2.1	Building ComLex via Clustering Word Embeddings	25
3.2.2	Human Evaluation of ComLex	25
3.2.3	Comparing ComLex with LIWC and Empath	27
3.2.4	Application of ComLex on Related Tasks	28
3.3	Unsupervised Exploration of Linguistic Signals	29
3.3.1	Effect of Misinformation on Audiences' Response	29
3.3.2	Linguistic Signals after Fact-Checking	34
3.4	Audiences' (Dis)belief to Misinformation - a Labeled Dataset	37
3.4.1	Another Data Collection from Fact-Checks and Social Media	38
3.4.2	Annotation of (Dis)belief Labels	38
3.4.3	Overview of (Mis)information and (Dis)belief	40
3.5	Modeling (Dis)belief with Supervised Learning	41
3.5.1	Exploratory Analysis of Linguistic Signals	41
3.5.2	Experiments with Classification Models	43
3.5.3	Thresholding Scores for Measurement	46
3.6	Measuring (Dis)belief via Applying Neural Models	48
3.6.1	Measuring the Prevalence of (Dis)belief	48
3.6.2	Effects of Time and Fact-Checks on (Dis)belief	49
3.6.3	Difference of (Dis)belief across Platforms	51
3.7	Summary of Audiences' Response	52
3.7.1	Research Questions and Answers	52
3.7.2	Limitations	53
3.7.3	Concluding Thoughts	54
4	Platforms	55
4.1	Platforms' Moderation on Misinformation - an YouTube Dataset	56
4.1.1	Moderation Decision - the Outcome Variable	56
4.1.2	Political Leaning and Extremeness - Treatment Variables	57
4.1.3	Misinformation and Fact-Checks - Treatment Variables	58
4.1.4	Social Engagement - Control Variables	58
4.1.5	Linguistic Signals - Control Variables	59
4.1.6	Overview of YouTube Videos and Comments	60
4.2	Criteria to Measure Effects	60
4.2.1	Independence - a Correlational Criterion	61
4.2.2	Separation - a Causal Criterion	61
4.3	Hypothesis Testing on Comment Moderation	62
4.3.1	Independence and Correlational Perception of Effects	62
4.3.2	The Problem of Confounding Variables	64
4.3.3	Separation and Causal Perception of Effects	66
4.4	Alternative Explanations and Robustness Check	68
4.4.1	Signals and Sources of Moderation	68
4.4.2	Credibility of Fact-Checkers	69
4.4.3	Alternative Thresholds and Control Variables	71

4.5	Summary of Platforms' Moderation	73
4.5.1	Research Questions and Answers	73
4.5.2	Limitations	73
4.5.3	Concluding Thoughts	74
5	Storytellers	76
5.1	Rationalized Neural Models	76
5.1.1	Problem Formulation and Rationalization Methods	77
5.1.2	Domain Knowledge as Weak Supervision	80
5.2	Rationalizing Public Datasets	81
5.2.1	Datasets, Domain Knowledge, and Experimental Setup	81
5.2.2	Comparing Rationalization Design Choices	83
5.3	Rationalizing Fact-Checks	84
5.3.1	Fact-Check Data and Domain Knowledge	85
5.3.2	Experiments on Fact-Checks	86
5.3.3	Structure of Misinformation Stories	89
5.4	Evolution of Misinformation	90
5.4.1	Evolution Over the Last Ten Years	90
5.4.2	2016 vs. 2020 US Presidential Elections	91
5.4.3	H1N1 vs. COVID-19 Pandemics	92
5.5	Summary of Storytellers' Strategies	93
5.5.1	Research Questions and Answers	93
5.5.2	Limitations	94
5.5.3	Concluding Thoughts	94
6	Conclusion	95
6.1	Summary of Contributions	95
6.2	Overview of Limitations	95
6.3	Concluding Remarks	95
	Bibliography	96

List of Figures

3.1	Interaction between social media and fact-checking websites. Following the publication of a post on Twitter, Facebook, YouTube, etc., Snopes.com and PolitiFact.com fact-check it and rate its veracity. Meanwhile, users comment on the post and sometimes refer to fact-check articles once they are released.	22
3.2	Distribution of veracity for posts from PolitiFact.com and Snopes.com. I map textual descriptions of veracity to ordinal values. I ignore descriptions that cannot be categorized such as <i>full flop</i> , <i>half flip</i> , <i>no flip</i> from PolitiFact.com and <i>legend</i> , <i>outdated</i> , <i>unproven</i> , <i>undetermined</i> , <i>research in progress</i> , <i>miscaptioned</i> , <i>misattributed</i> , <i>correct attribution</i> , <i>not applicable</i> , etc. from Snopes.com.	23
3.3	Veracity of posts fact-checked by both PolitiFact.com and Snopes.com. The veracity rulings are strongly correlated ($\rho = 0.671^{***}$).	23
3.4	Distribution of veracity for deleted posts. The likelihood of post deletion is negatively correlated with the veracity of posts ($r_{pb} = -0.052^{***}$).	23
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” ($\bar{\mu} = 4.506$) with moderate inter-rater agreement ($\bar{r} = 0.531$).	26
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” ($\bar{\mu} = 4.359$) with strong inter-rater agreement ($\bar{r} = 0.675$).	26
3.7	Comparing ComLex with LIWC. Each scatter plot shows the correlation of ComLex and LIWC for a similar word cluster. Selected clusters including <i>family</i> ($r = 0.883^{***}$), <i>pronoun</i> ($r = 0.877^{***}$) and <i>preposition</i> ($r = 0.833^{***}$) show very strong correlation.	27
3.8	Comparing ComLex with Empath. Each scatter plot shows the correlation of ComLex and Empath for a similar word cluster. Selected clusters including <i>monster</i> ($r = 0.949^{***}$), <i>timidity</i> ($r = 0.904^{***}$) and <i>ugliness</i> ($r = 0.908^{***}$) show very strong correlation.	28
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.	30

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] (<i>three example words</i>).	31
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.	34
3.12	Percentage of reference for PolitiFact.com and Snopes.com. Each pie chart shows the percentage of posts that contains <i>politifactref</i> or <i>snopesref</i> over all posts checked by the website.	35
3.13	Semantics of reference for PolitiFact.com and Snopes.com. The learned embedding, which encodes the semantics of <i>politifactref</i> or <i>snopesref</i> , is plotted along with other words in <i>fact</i> and three <i>fake</i> clusters. Dimensions are reduced from 100 to 2 using t-SNE. References to PolitiFact.com and Snopes.com carry similar semantics as other words expressing <i>fact</i> in the right part of the figure, as oppose to words expressing <i>fake</i> in the left part of the figure.	35
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.	36
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.com, PolitiFact.com) are identified as reference to fact-checking websites, while words in red blocks (i.e., fuck, damn) are mapped in the <i>swear</i> word cluster.	37
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.	39
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.	40
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.	40
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.	42
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.	42

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.	45
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.	45
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.	49
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.	49
3.25	Platforms difference of expressed disbelief. Facebook comments express less disbelief than YouTube. However, the difference is not significant for Twitter. . . .	51
3.26	Platforms difference of expressed belief. Facebook comments express more belief than YouTube, and YouTube comments express more belief than Twitter.	51
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).	56
4.2	Data collection process and an illustrative example. Starting from a fact-check article on PolitiFact.com, 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 <i>ComLex</i> . Finally, I compare two crawls to identify moderated comments.	57
4.3	Graph models of the independence criterion. Null hypothesis H_0^{ind} : $M \perp\!\!\!\perp P$. . .	61
4.4	Graph models of the separation criterion. Propensity scoring function $ps(J)$ is used to summarize J to a scala, hence 2nd null hypothesis H_0^{sep} : $M \perp\!\!\!\perp P \mid ps(J)$. . .	61
4.5	Correlational difference in moderation likelihood. Moderation likelihood for each group with 95% CI is shown. All four null hypotheses are rejected.	63
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.	64

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.	66
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.	67
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 ($r > 20\%$).	69
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 ($\lambda \leq +1$).	70
4.11	Alternative H1a₀ (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.	71
4.12	Alternative H1b₀ (extreme/center) thresholds. The effect of extreme/center thresholds is minimal for most hypotheses, except for H1a ₀ and H1b ₀	72
4.13	Alternative linguistic controls. The effect of alternative linguistic controls using lexicon LIWC instead of ComLex is minimal for all hypotheses.	72
5.1	A snippet of the misinformation structure. Each line is a snippet from a fact-check. Key phrases identifying the misinformation types are highlighted. Phrases with similar semantics are clustered in colored boxes. This structure is a sample of my final results.	77
5.2	Hard and soft rationalization methods. Hard rationalization is an end-to-end model that first uses input x to generate rationales z , and then uses unmasked tokens to predict y . Soft rationalization is a three-phased model that first uses input x to predict y and outputs importance scores s , then binarizes s to rationales z , and finally uses unmasked tokens to predict y again as evaluation for faithfulness. . . .	78
5.3	Structure of misinformation types. The ten identified clusters (colored) offer empirical confirmation of theorized misinformation types, contain novel fine-grained clusters, and reorganize the structure of misinformation stories.	87
5.4	Evolution of misinformation over the last ten years. Conspiracy theories, fabricated content, and digital manipulation have increased in prevalence. The prevalence of (arguably) less politicized stories (e.g., legends and tales, pranks and jokes, mistakes and errors) has decreased. (95% confidence intervals.)	90
5.5	Misinformation between notable events. The most prevalent misinformation type for both US presidential elections is fabricated content, while the 2016 election has more hoaxes and satires. The H1N1 pandemic in 2009 has more legends and tales, while the COVID-19 pandemic attracts more conspiracy theories. (95% confidence intervals.)	92

List of Tables

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.	28
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.	47
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.	47
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.	50
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.	60
5.1	Evaluation results for hard and soft rationalization methods. My experiments show that: (a) hard rationalization requires a sensitive hyperparameter λ_z to regularize rationales (h_2 to h_0); (b) soft rationalization achieves the best $F_1(y)$ overall, but $\Pr(z)$ depends on the rationale extraction approach (s_2/s_3 to s_0); (c) domain knowledge as weak supervision improves $\Pr(z)$ for both hard (h_1 to h_0) and soft (s_1 to s_0) rationalization while maintaining similar $\%(z)$ and $F_1(y)$; (d) soft rationalization achieves better $\Pr(z)$ in a fair comparison (s_1 to h_1).	83

Acknowledgments

When I look back on the last five years of my Ph.D. journey, I feel encouraged and supported by my advisor, collaborators, colleagues, friends, and family.

I owe a debt of gratitude to Christo Wilson, who is the best advisor I could ever hope for as a student: he has given me the freedom to explore research areas of my interests and offered immeasurable knowledge and wisdom for my thesis topic. Yet Christo is much more than a good advisor; he is, above all things, a kind person: he tolerates and forgives my lapses and always stands by me even when I made mistakes, e.g., when I misused proprietary datasets and antagonized collaborators, when I forgot to shut down servers and wasted thousands of dollars, and when I could not finish writing papers and missed promised deadlines. Not only did Christo offer me countless help and support for my research, but he has also taught me, through his actions, how to treat others and myself. For this, I am eternally grateful.

During my Ph.D., I had the privilege to collaborate with many of the leading experts in my field: Alan Mislove, Abigail Evans, and David Lazer from Northeastern University; Miriam Metzger and Andrew Flanagin from UC Santa Barbara; Kenneth Joseph from SUNY Buffalo; Nitin Joglekar from Boston University; among others. Their guidance, advice, support, and knowledge have profoundly impacted my research projects and monumentally helped this thesis. I am incredibly grateful for their presence, and I am honored to have three of them on my thesis committee.

I also had the pleasure to intern at several industrial labs and work with many talented colleagues: Shiou Tian Hsu from Facebook; Cong Yu, Simon Baumgartner, and Abe Ittycheriah from Google Research; William Groves, Sam Anzaroot, and Alejandro Jaimes from Dataminr; among others. My research with them has offered instrumental inspirations for my thesis.

Of course, I am grateful to all my friends and labmates: Muhammad Ali, Muhammad Abu Bakar Aziz, Le Chen, Avijit Ghosh, Dan Guo, Yao Hao, Desheng Hu, Ting Huang, Fangfan Li, Yanjun Liu, John Martin, Ming Min, Ronald Edward Robertson, Zhichuang Sun, Jieyi Tang, Jiewen Zhu, among others. Their companionship has made my Ph.D. journey as enjoyable as it is rewarding.

Finally, and most importantly, I give all my thanks to my parents, Yonghong Jiang and Lingfang Zhou, for their unconditional love and caring.

Abstract of the Dissertation

Measuring the Misinformation Ecosystem:

Audiences, Platforms, and Storytellers

by

Shan Jiang

Doctor of Philosophy in Computer Science

Northeastern University, June 2021

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 disciplines have investigated the misinformation problem from multiple perspectives, ranging from the socio-psychological foundations of audiences' susceptibility to algorithmic solutions aiding platforms' intervention against the spread of misinformation. Yet, large-scale empirical study is still needed 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 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 versus 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 extent audiences disbelieve or believe in these stories. Applying supervised classifiers trained to identify (dis)belief, I estimate 12%/15% of audiences express disbelief, and 26%/20% of them express belief in true/false information.

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

hybrid process of removing toxic content to maintain community standards. Using YouTube as a lens, this part investigates how misinformation and partisanship of videos interact with moderation practices on their corresponding comments. I observe that videos containing verifiably false content have heavier moderation of their comments, especially when the comments are posted after a fact-check article. Additionally, I find no evidence to support allegations of political bias in content moderation on YouTube, when justifiable factors (e.g., hate speech) are controlled.

Storytellers generate misinformation and then release them onto platforms. The third part of the thesis structurizes storytellers' behaviors and explores prevalent types of misinformation to date, by rationalizing fact-check articles. My intuition is that key phrases in a fact-check article that identify the misinformation type(s) (e.g., doctored images, urban legends) also act as rationales that determine the verdict of the fact-check (e.g., false). I experiment on rationalized models with domain knowledge as weak supervision to extract these phrases as rationales, and then cluster semantically similar rationales to summarize prevalent misinformation types. Using archived fact-check articles from Snopes.com, I identify ten types of misinformation stories. I discuss how these types have evolved over the last ten years and compare their prevalence between the 2016/2020 US presidential elections and the H1N1/COVID-19 pandemics.

Altogether, my work presents an overview of the misinformation ecosystem to date, as well as methodologies and tools for measuring it. The empirical findings in the thesis are derived from computational approaches based on observational data, and are reproducible from repositories that I have publicly released. 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 [271] to deliberate hoaxes [140, 141], propaganda [29, 157, 216, 263], disinformation [135, 263], and recently (and controversially) “fake news” [51, 271].

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 [92], and 44% of Americans age 18 or older visited at least one untrustworthy website [93]. Years later, 29% of US adults in a survey believed that the “exaggerated threat” of the COVID-19 pandemic purposefully damaged former US president Donald Trump [259], and 77% of Trump’s supporters believed “voter fraud” manipulated the 2020 US presidential election in spite of a complete lack of evidence [203].

To date, misinformation has been documented across the globe, e.g., in Africa [272], Asia [125], and Europe [77]. As one example 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 [207, 236]. Yet, misinformation continues to be posted on social media by politicians, partisan pundits, and even ordinary users [266].

The proliferation of misinformation has been raising societal concerns about its potential consequences. For example, in the political context, fabricated stories and partisan opinions may polarize the public [144], alter voters’ perceptions about candidates [4, 61], and erode trust in institutions [47], therefore posing a threat to the democracy [103, 178]; In health-related domains, the COVID-19 “infodemic” may substantially impact individuals’ intentions to vaccinate and attitudes about governmental enforcement of pandemic-related regulations [60].

CHAPTER 1. INTRODUCTION

Existing surveys and experiments across disciplines have investigated the misinformation problem from multiple perspectives, ranging from the socio-psychological foundations of audiences' susceptibility [84, 186, 225, 270] to algorithmic solutions aiding platforms' intervention on the spread of misinformation [73, 211, 216, 238, 263, 264, 267]. Yet, large-scale empirical study is still needed 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 three key players:

- *Audiences*, who receive and respond to misinformation [114, 117, 167].
- *Platforms*, through which misinformation reaches its audiences [115, 116].
- *Storytellers*, who generate misinformation [118].

I approach this study using computational methods on observational data. I have publicly released corresponding datasets and code repositories to make my 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 these falsehoods or inaccuracies. The first part of the thesis explores if and how this misinformation affects its audiences [114, 117, 167].

Although scholars are still debating whether misinformation impacted the outcome of the 2016 US presidential election [4, 93], exposure to misinformation may still harm audiences by promoting partisanship, reducing trust in civic institutions, and discouraging reasoned conversation [28, 80]. Research suggests that audiences are indeed vulnerable to misinformation because of psychological and sociological predispositions [84, 186, 225, 270]. Furthermore, misinformation often uses inflammatory and sensational language [216, 263, 264] that can alter audiences' emotions, which are a core component of how they perceive their political world [161], and can sometimes affect their perceived bias of information [273].

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. For example, a tweet posted by Donald

CHAPTER 1. INTRODUCTION

Trump claiming that Barack Obama was born in Kenya was later fact-checked by both Snopes.com and PolitiFact.com and found to be false [68, 168]. 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 [49, 90]. 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, to 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 [79, 95, 208, 280], while others found that it has minimal impact [146, 189] and sometimes even “backfires” on its audience [189–191]. In fact, the work of Snopes and PolitiFact.com has itself become politicized by those who view their work as biased, and this has led to attempts to discredit fact-checks [183, 220, 232].

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.com and PolitiFact.com to obtain veracity rulings (i.e., from true to false). Then, I build an emotional and topical lexicon, named *ComLex*, using a hybrid method of natural language processing (NLP) techniques and human validation. I use this lexicon later 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*.

CHAPTER 1. INTRODUCTION

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 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 [9, 187]. 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 [55], XLNet [286], and RoBERTa [155]). 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

CHAPTER 1. INTRODUCTION

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 [115, 116]. Besides the misinformation problem [50], social media platforms have also been subject to heightened levels of controversy and scrutiny for other issues, e.g., violent hate speech [192] and partisanship [4].

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 [145], investing in more artificial intelligence to filter content [85], and partnering with fact-checking organizations to identify misinformation [88]. In terms of policy, the platforms are updating their community guidelines with expanded definitions of what they believe constitutes hate speech, harassment, misinformation, etc [65, 257, 287].

Platforms' use of content moderation to police politically-sensitive topics has led to backlash from ideological conservatives, who claim that social media platforms are biased against them and are censoring their views [124, 260]. 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)" [30, 58]. 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" [101]. These concerns are driven by multiple factors, including anecdotal reports that: Facebook's Trending News team did not promote stories from conservative media outlets [188], Twitter "shadow banned" conservative users [184], fact-checking organizations are biased [220], and selective reporting by partisan news agencies [10].

However, there is no scientific evidence that social media platforms' content moderation practices

CHAPTER 1. INTRODUCTION

exhibit systematic partisan bias [115, 234, 235]. On the contrary, there are many cases where ideologically liberal users were moderated, although these cases have received less attention in the media [164]. 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 [87]. Further, there is evidence suggesting that users tend to overestimate bias in moderation decisions [235].

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 H_0 that the variable has no effect on comment moderation, and then use two formal criteria (i.e., *independence* and separation [22]) 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 [223] and misinformation labels from Snopes.com or PolitiFact.com [117]. 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) [177] and the linguistics in comments (e.g., hate speech) [41, 235]. Therefore, I re-analyze my 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 ~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.

CHAPTER 1. INTRODUCTION

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

I approach these hypotheses using an empirical method for auditing black-box decision-making processes [229] 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 and then release them onto platforms. The third part of the thesis structurizes storytellers’ strategies and explores prevalent types of misinformation to date [118].

From storytellers’ perspectives, misinformation can be generated in numerous ways, e.g., fabricating or manipulating content, making false context or connection. However, existing studies mostly adopted the term “misinformation” as a coarse concept for any false or inaccurate information, which incorporates a broad spectrum of misinformation stories. Although misinformation *types* have been theorized and categorized by practitioners [271], there is, to my knowledge, no empirical research that has systematically measured these prevalent types of misinformation stories. Therefore, this part aims to unpack the coarse concept of misinformation and structurize it to fine-grained story types.

I, again, leverage fact-checks as a corpus in this part. As a critical component of fact-checks’ semi-structured journalistic style, fact-checks often embed the (mis)information type(s) within their steps of reasoning [112]. For example, consider the following quote from a Snopes.com fact-check with a verdict of **false** [64]:

“...For instance, some started sharing a **doctored photograph** of Thunberg with alt-right boogeyman George Soros (the original photograph featured former Vice President Al Gore).”

The key phrase **doctored photograph** in the quote identifies the misinformation type of the fact-checked story. With a large corpus of fact-checks, these phrases would accumulate and reveal prevalent types of misinformation stories.

CHAPTER 1. INTRODUCTION

Extracting these phrases is a computational task. My intuition is that such phrases in a fact-check also act as *rationales* that determine the verdict of the fact-check. In the previous example, the verdict is **false** in part *because* the story contains a **doctored photograph**. Therefore, a neural model that predicts the verdict of a fact-check would also use the misinformation types as rationales.

To realize this intuition, I experiment on existing rationalized neural models to extract these phrases [110, 143], and, to target specific kinds of rationales, I additionally propose to include domain knowledge as weak supervision in the rationalizing process. Using public datasets as validation [37, 289], I evaluate the performance variation of different rationalized models, and show that including domain knowledge consistently improves the quality of extracted rationales.

After selecting the most appropriate method, I conduct an empirical investigation of prevalent misinformation types. Using archived fact-checks from Snopes.com, spanning from its founding in 1994 to 2021, I extract rationales by applying the selected model with theorized misinformation types for weak supervision [271], and then cluster rationales based on their semantic similarity to summarize prevalent misinformation types.

Using my derived lexicon of these clustered misinformation stories, I then explore the evolution of misinformation types over the last ten years. Overall, this part investigates the following RQs:

- **RQ3.1**, *what are the prevalent types of misinformation stories in the US over the last ten years?* I identify ten types of misinformation stories, including urban legends and tales, altered or doctored images, hoaxes and pranks, bogus scams, mistakes and errors, fabricated content, baseless conspiracies, satires and parodies, fictitious content, and sensational clickbait.
- **RQ3.2**, *how has the prevalence of misinformation types evolved over the last ten years?* Heavily politicized misinformation types, such as fabricated and misleading content and conspiracy theories have nearly doubled over the last ten years, while the prevalence of arguably less politicized stories, such as legends and tales, hoaxes and pranks, have decreased.
- **RQ3.3**, *how has the prevalence of misinformation types evolved between the 2016 and the 2020 US presidential elections?* The prevalence of many misinformation types are similar between the two elections, while the 2016 election has more hoaxes and satires. The most prevalent type during both elections is fabricated content and conspiracy theories.
- **RQ3.4**, *how has the prevalence of misinformation types evolved between the H1N1 and the COVID-19 pandemics?* The prevalence of certain misinformation types are significantly

CHAPTER 1. INTRODUCTION

different between two pandemics. Notably, the H1N1 pandemic has many more legends and tales, while COVID-19 has more conspiracy theories.

1.4 Outline

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

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**, § 4 investigates platforms' moderation practice on misinformation and answers **RQ2.1-RQ2.4**, § 5 structurizes storytellers' strategies to generate misinformation and answers **RQ3.1-RQ3.4**, and finally, § 6 discusses limitations of the thesis and concludes.

Chapter 2

Background

This chapter introduces the background of the misinformation ecosystem and positions my work around related areas.

From audiences’ perspective, § 2.1 introduces the concept of misinformation and its potential consequences. From platforms’ perspective, § 2.2 introduces content moderation policies and practices and the related controversy around partisan bias. From a storytellers’ perspective, § 2.3 introduces theorized misinformation types and the complexity of this topic. Finally, § 2.4 reviews methodology and introduces how natural language processing (NLP) can be helpful for social science research.

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 [29, 157, 216, 263], 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* [238]. Some scholars prefer to use “misinformation” to broadly refer to all false and inaccurate information regardless

CHAPTER 2. BACKGROUND

of intent [54, 93, 105, 146, 273], while others prefer the more modern (but polarizing) term “fake news” [4, 140, 141, 238]. Other scholars restrict “misinformation” to unintentional inaccuracies, and use “disinformation” for deliberate deception [135, 271]. “Propaganda” typically refers to intentional and strictly political information [29, 157], 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 [82, 225, 270]. Another theory called *confirmation bias* shows that people prefer to accept information that confirms their existing beliefs [186]. Sociological theories including *social identity theory* [36, 252] and *normative influence theory* [12] 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 [84].

Audiences’ vulnerability to misinformation affects their behavior and communication. For example, in-lab experiments have shown that exposure to biased information online [222] may significantly impact voting behavior [61, 62], while naïve information sharing may promote homophilous “echo chambers” of information [54, 149, 150].

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 [127, 172].

Another line of work measured the spread and impact of misinformation, finding that “fake news” spread faster than truthful news [266], and that a large fraction of “fake news” are spread by “bots” [74, 231]. Misinformation is especially (and alarmingly) easy to be spread during crises, because people attempt to complete partial information using their natural sense-making processes [105], although such misinformation can sometimes be self-corrected by the crowd [11].

Early computational work is focused on the algorithmic model and detection of misinforma-

CHAPTER 2. BACKGROUND

tion [238]. 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 [21, 160], or by attempting to infer a claim from existing knowledge graphs [48, 100, 237, 268, 282]. 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 [73, 211, 216, 263, 264, 267].

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 [72, 126, 156, 221, 293].

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 [120, 214, 251], as well as different propagation network structures between fake and truthful news [94, 119]. Recent work has also proposed tools that actively solicit and analyze “flags” on misinformation from users [256]. 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 has 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 [44], harassment [199], or hate speech [177]. Understanding and detecting the linguistic variants present in these comment threads may help when implementing intervention and moderation systems [86, 111].

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

CHAPTER 2. BACKGROUND

people the flu significantly reduced people’s belief in the rumor, but also reduced some people’s willingness to vaccinate because of side effects [190, 191]. However, later research failed to duplicate the results [95]. This phenomenon is called the “backfire” effect, where attempting to intervene against misinformation only entrenches the original, false belief further [189].

Even without the backfire effect, there are several experiments that found that fact-checking has limited corrective effects [146, 189]. However, others found that people are willing to accept fact-checking even when the information challenges their ideological commitments [79, 208, 280]. 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.

Examples of major fact-checking organizations include Snopes.com [169], Politifact.com [233], and FactCheck.org [108]. 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 [7, 8]. 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 [96, 162], but that including biographical information about the author of the fact-check in the article itself reduces the effectiveness [82]. On online platforms, alert messages and tags that warn users to the presence of misinformation can help reduce the influence of this content [62, 204].

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 [134] or bias warnings [62] 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 socio-technical systems [134], e.g., Google search [112], YouTube [90], and Facebook [49].

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 [270] and confirmation bias theory [186] from psychology suggested that peo-

CHAPTER 2. BACKGROUND

ple tend to believe in information that resonates with their pre-existing (yet potentially false) beliefs. Social identity [245] and normative influence theory [129] 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 [172]. More recent reports from the Reuters Institute [187] and Pew Research Center [9] 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 [165].

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. Specifically, I focus on raising research questions for YouTube as a case study, as this is the platform I study in § 4.

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 [65, 257, 287].

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 [287]. 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 [145]. Additionally, a comment may be removed by the corresponding video uploader, or by the commenter themselves [287]. Besides these human efforts, YouTube also uses algorithms that automatically flag and moderate inappropriate content [85]. 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 [40]. A study that interviewed users of Twitter’s “blocklist” feature discussed how it can be used to prevent harassment [111].

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 [57] and automated moderation algorithms are prone to false positives and negatives [262]. These moderation strategies are also brittle: a study on Instagram found that users in pro-eating disorder

CHAPTER 2. BACKGROUND

communities invented variations of banned tags (e.g., “anorexie” instead of “anorexia”) to circumvent lexicon-based moderation [39].

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 [41]. 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 [235]. 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

Researchers have used algorithm auditing techniques [229] to investigate bias in black-box systems. Studies have found gender and racial bias on hiring sites [43], freelance markets [97], ridesharing platforms [113], and online writing communities [71]. 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 [243], while search engines may create “filter bubbles” that isolate users from ideologically opposing information [104, 148].

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 [20]. 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 [222].

As for content moderation, there have been several allegations that social media platforms are censoring or biased against political conservatives [124, 260]. 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 [181]. Therefore, in § 4, I investigate the veracity of these allegations as they pertain to YouTube 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).

There are at least two reason why misinformation may attract moderators attention on social

CHAPTER 2. BACKGROUND

media. First, the platforms are updating their policies to specifically target misinformation. For example, Facebook updated their policies “to remove misinformation that has the potential to contribute to imminent violence, physical harm, and voter suppression.” [66] Second, as I will show in § 3.3, misinformation does alter audiences’ comments and increases their usage of swear words and other phrases that might violate non-misinformation-related community guidelines. Therefore, misinformation content could draw more attention from moderators. For these reasons, 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 to the effect of fact-checking, **RQ2.4**, *does a fact-check of a video affect the moderation decision of its comments?*

2.3 Manipulation Strategies and Its Complication

Existing studies mostly adopted the term “misinformation” as a coarse concept, yet, storytellers have complicated manipulation strategies to generate misinformation. In this section, I briefly review theorized types of misinformation.

2.3.1 Theorized Misinformation Types

A study from First Draft theorized seven types of potential misinformation and disinformation types [271], including satire and parody, misleading content, imposter content, fabricated content, false connection, false context and manipulated content. The study then explain why each type is created along eight possible reasons, including poor journalism, to parody, to provoke or to “punk”, passion, partisanship, profit, political influence, and propaganda. However, to my knowledge, no empirical evidence has been connected to this typology, therefore, I aim to systematically study this topic and propose **RQ3.1**, *what are the prevalent types of misinformation stories in the US over the last ten years?*

2.3.2 Evolution of Misinformation Stories

Researchers have investigated the evolution of specific types of misinformation as case studies, e.g., state-sponsored disinformation [244, 279], fauxtography [269, 290], and conspiracy theories [205, 228]. Building on discovered structure of misinformation stories, I explore the evolution of misinformation over time and ask **RQ3.2**, *how has the prevalence of misinformation types evolved*

CHAPTER 2. BACKGROUND

over the last ten years? Additionally, I also explore the evolution of misinformation between major events and ask **RQ3.3**, *how has the prevalence of misinformation types evolved between the 2016 and the 2020 US presidential elections?* and **RQ3.4**, *how has the prevalence of misinformation types evolved between the H1N1 and the COVID-19 pandemics?*

2.4 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 and how I use these tools in this thesis.

2.4.1 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 [83, 104]. 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 [123, 197]. 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) [201, 254], which contains both emotional, topical, and syntactic categories. An alternative for LIWC is Empath [70], 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) [174, 175] is another human curated lexicon that is structured around Plutchik’s wheel of emotions [206]; 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) [246] which has more topics than LIWC but fewer emotions, and Affective Norms for English Words (ANEW) [32] and SentiWordNet [18, 63] which have more emotions than LIWC.

CHAPTER 2. BACKGROUND

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 [147]. In the case of social media, existing lexicons such as NRC Hashtag Emotion Lexicon (HashEmo) [173] and others [25, 151] 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 [158, 170, 202] and then use unsupervised learning to cluster words [70]. This method has been used by some studies in the misinformation domain to analyze stylistic features of articles [216, 264].

In this thesis, I use the bag-of-words and lexical approach in both § 3 for measuring linguistic signals and in § 4 for constructing linguistic controls. I note that 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 as supporting evidence to validate my findings. Additionally, I also present a performance evaluation between lexicons in terms of predictive ability in § 3.2.3.

2.4.2 Sequence Classification 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 [284]. Recent solutions for solving the sequence classification problem use neural architectures [136, 294] and pre-trained transfer-learning models [55, 155, 286].

Specific applications of the sequence classification problem are defined within domain-specific datasets. There are several existing NLP tasks that are related to the topics mentioned in this thesis. Stance detection, for example, aims to determine the for-or-against stance in comments for a two-sided argument (e.g., marijuana, gay marriage) [99, 122], and, in the political context, it often overlaps with ideology identification [213]. Classification tasks of other creative languages such as sarcasm [89], satire [35], irony [67], and humor [285] are also related to the misinformation in general.

CHAPTER 2. BACKGROUND

In this thesis, I use neural models in both § 3 for modeling (dis)belief and in § 5 for modeling fact-checks.

2.4.3 Interpretability, Explainability and Rationalization

Realizing my intuition for § 5 requires neural models to (at least shallowly) reason about predictions. To achieve this goal, there has been some recent studies on the interpretability and explainability of neural models [59, 153]. In this thesis, I focus on a specific problem formulation, *rationalization* [56, 143], which requires a model to output predictions, as well as the parts of inputs it used or focused on to make the decision.

Hard rationalization requires a model to output a binary mask of the input representing whether a token is selected as part of the rationales, and only the selected tokens can be used to make the final prediction. The initial proposed model was trained end-to-end with modularized architecture [143], and recent work improved the initial model with adversarial components [37, 288], reparameterization [23], etc.

Soft rationalization, in contrast, requires a model to output continuous importance scores of the input, and the final prediction is made based on importance-weighted input. *Attention* mechanism provides build-in access to such scores [19]. Although there have been debates on the properties of attention-based explanations [109, 230, 277], recent work showed that phrases extracted by intuitive rules on attention weights achieved reasonable performance comparing to human rationales [110].

In this thesis, I experiment with both rationalization methods in § 5. My goal is to understand how the hyperparameter selection affects the extracted rationales, and therefore choose the most appropriate one to investigate my research questions about misinformation types.

Chapter 3

Audiences

Audiences receive and respond to misinformation, and therefore their behaviors are potentially influenced by this falsehood or inaccuracies. In this chapter, I measure audiences’ response to misinformation, and explore **RQ1.1** to **RQ1.5**:

- **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?*

The rest of the chapter is organized as follows: § 3.1 introduces an unlabeled dataset for fact-checked claims and social media comments, § 3.2 builds a lexicon for analyzing the data, § 3.3 uses the lexicon to analyze the data in an unsupervised manner and explores **RQ1.1** and **RQ1.2**. Then, § 3.4 introduces another labeled misinformation dataset and my annotation schemes for (dis)belief, § 3.5 builds supervised NLP models to predict (dis)belief, § 3.6 applies the model onto my dataset and explores **RQ1.3** to **RQ1.5**. Finally, § 3.7 summarizes.

3.1 Audiences’ Comments to Misinformation — an Unlabeled Dataset

RQ1.1 and **RQ1.2** require an unsupervised exploration of audiences’ response to misinformation. Thus, 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.

CHAPTER 3. AUDIENCES

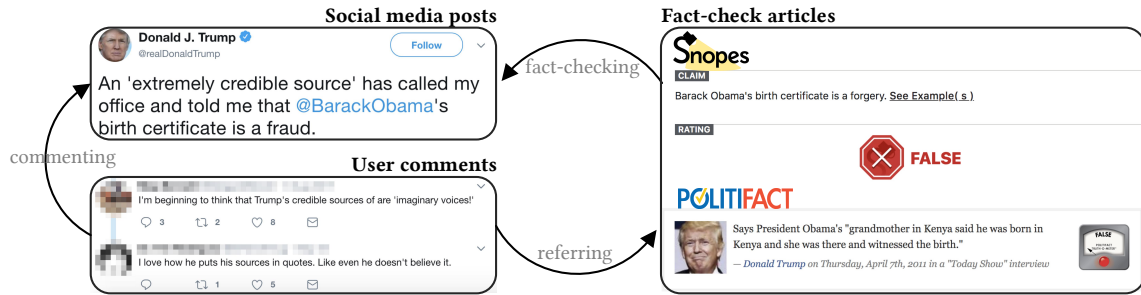


Figure 3.1: **Interaction between social media and fact-checking websites.** Following the publication of a post on Twitter, Facebook, YouTube, etc., Snopes.com and PolitiFact.com fact-check it and rate its veracity. Meanwhile, users comment on the post and sometimes refer to fact-check articles once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

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.com and PolitiFact.com, who then publish articles containing evidence for or against 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.com and Snopes.com as starting points. I choose PolitiFact.com and Snopes.com 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.com and Snopes.com, 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 ID¹ 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.

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

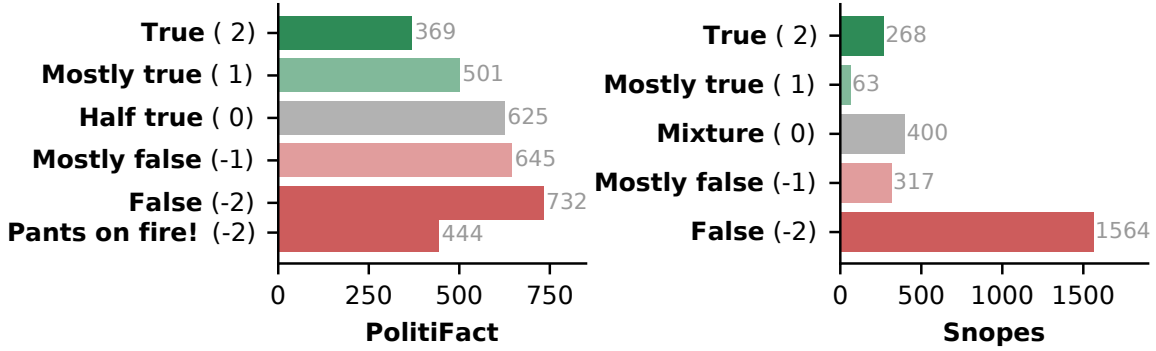


Figure 3.2: **Distribution of veracity for posts from PolitiFact.com and Snopes.com.** I map textual descriptions of veracity to ordinal values. I ignore descriptions that cannot be categorized such as *full flop*, *half flip*, *no flip* from PolitiFact.com and *legend*, *outdated*, *unproven*, *undetermined*,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

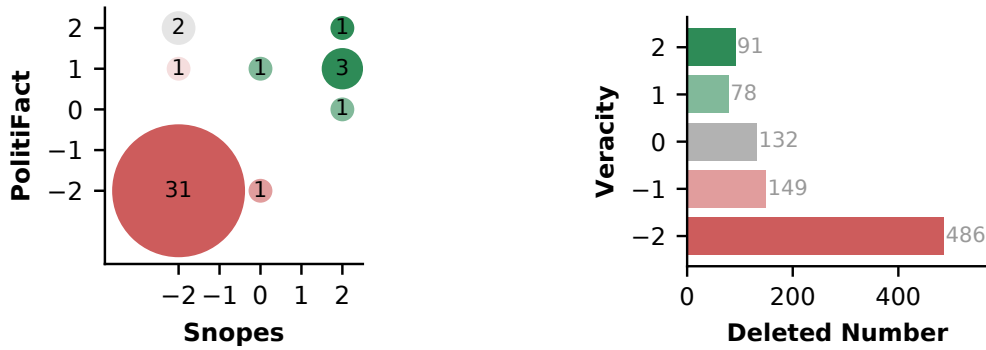


Figure 3.3: **Veracity of posts fact-checked by both PolitiFact.com and Snopes.com.** The veracity rulings are strongly correlated ($\rho = 0.671^{***}$).

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

3.1.2 Overview of Fact-Checked Claims and Comments

Overall, I collected 14,184 fact-check articles from Politifact.com and 11,345 from Snopes.com, 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.com and Snopes.com have different ruling criterion and therefore different textual descriptions for post veracity. To make them comparable, I translated their descriptions to a scale from -2 to 2 using the mapping shown in Figure 3.2. I view *pants on fire!* and *false* as -2 for PolitiFact.com, and ignore descriptions that cannot be categorized such as *full flop*, *half flip*,

CHAPTER 3. AUDIENCES

no flip from PolitiFact.com and *legend*, *outdated*, *unproven*, *undetermined*, *research in progress*, *miscaptioned*, *misattributed*, *correct attribution*, *not applicable*, etc. from Snopes.com. 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.com and Snopes.com, 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 [7, 8].

Finally, I collected user comments on the 5,303 fact-checked social media posts using their 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 [75, 249]. 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.

²* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

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* [31], 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 [5, 72]. In addition, I replaced all URLs that link to a Snopes.com or PolitiFact.com webpages with the special tokens *snopesref* or *politifactref*. This enables me to group all fact-checked posts from Snopes.com and PolitiFact.com 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* [219] to learn *Word2Vec* [170] 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 [185] 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*

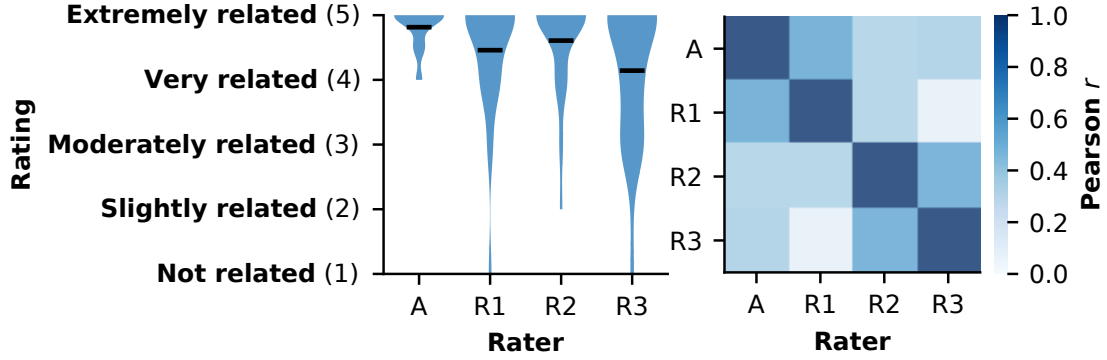


Figure 3.5: **Survey results for semantic closeness.** The left violin plot shows the rating distribution of four raters and the right heatmap shows the inter-rater correlations. Words in clusters are rated on average above “very related” ($\bar{\mu} = 4.506$) with moderate inter-rater agreement ($\bar{r}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

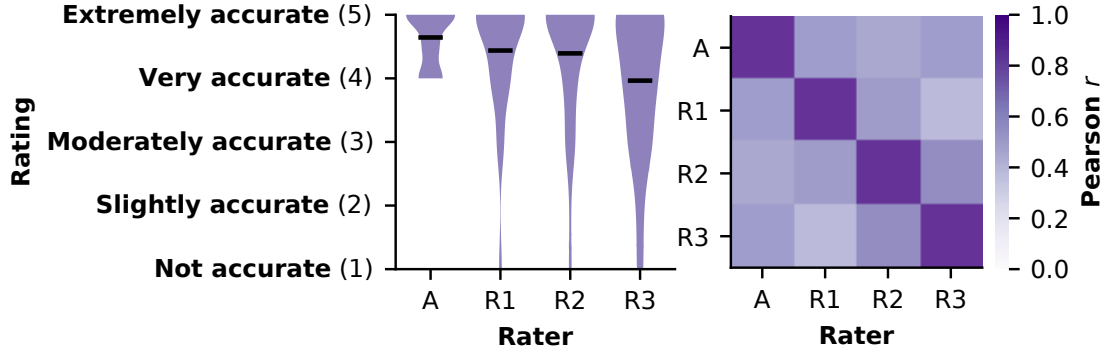


Figure 3.6: **Survey results for information accuracy.** The left violin plot shows the rating distribution of four raters and the right heatmap shows the inter-rater correlations. Cluster names and additional information are rated on average above “very accurate” ($\bar{\mu} = 4.359$) with strong inter-rater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

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 [33]. 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 ≥ 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 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

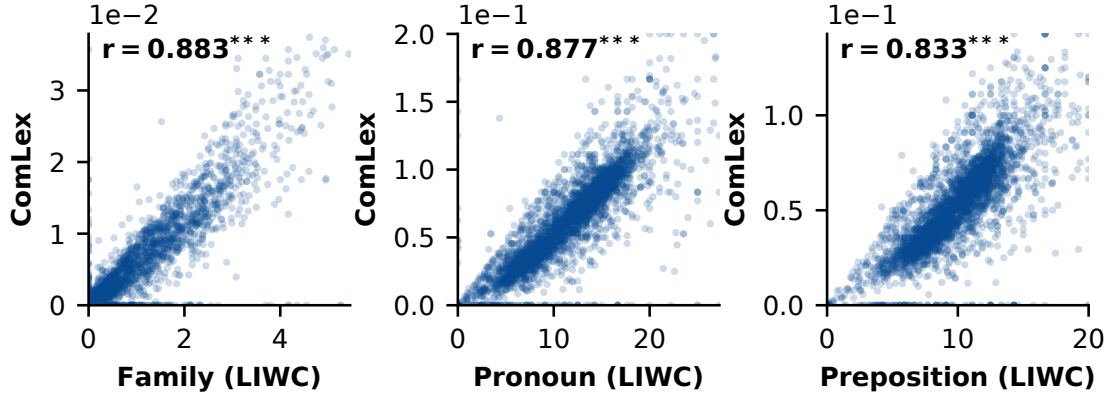


Figure 3.7: **Comparing ComLex with LIWC.** Each scatter plot shows the correlation of ComLex and LIWC for a similar word cluster. Selected clusters including *family* ($r = 0.883^{***}$), *pronoun* ($r = 0.877^{***}$) and *preposition* ($r = 0.833^{***}$) show very strong correlation.

“very related” on average (mean average $\bar{\mu} = 4.506$), 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 $\bar{r} = 0.531$).

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 ($\bar{\mu} = 4.359$), and the difference in μ among raters is significant ($H = 10.8^*$). As shown in the heatmap, the raters are strongly agreed with each other on average ($\bar{r} = 0.675$). 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 similar clusters such as *family* (Pearson $r = 0.883^{***}$), *pronoun* ($r = 0.877^{***}$) and *preposition* ($r = 0.833^{***}$).

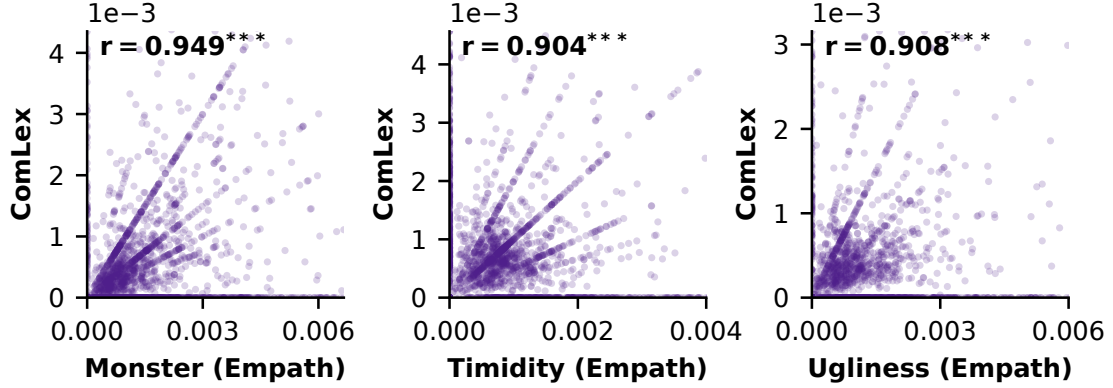


Figure 3.8: **Comparing ComLex with Empath.** Each scatter plot shows the correlation of ComLex and Empath for a similar word cluster. Selected clusters including *monster* ($r = 0.949^{***}$), *timidity* ($r = 0.904^{***}$) and *ugliness* ($r = 0.908^{***}$) show very strong correlation.

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 [194]	Human judges		56.9% - 61.9%
	GI	SVM	73.0%
	LIWC		76.8%
	ComLex		81.4%
	Learned unigrams		88.4%
Movie reviews [198]	Human judges		58.0% - 69.0%
	ComLex	SVM	72.3%
	Learned unigrams		72.8%

*All accuracy data are drawn from the original papers [194, 198] except for ComLex.

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

CHAPTER 3. AUDIENCES

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 [194], 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 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 [198], 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

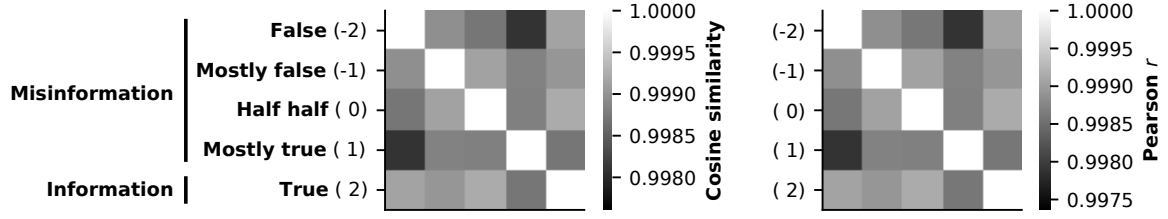


Figure 3.9: **Similarity matrix over veracity.** Heatmaps shows the similarity matrix over veracity using cosine similarity and Pearson correlation respectively. Using both measures, clear patterns of decreasing similarity are visible from -2 to 1, but the trend does not hold for 2.

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

CHAPTER 3. AUDIENCES

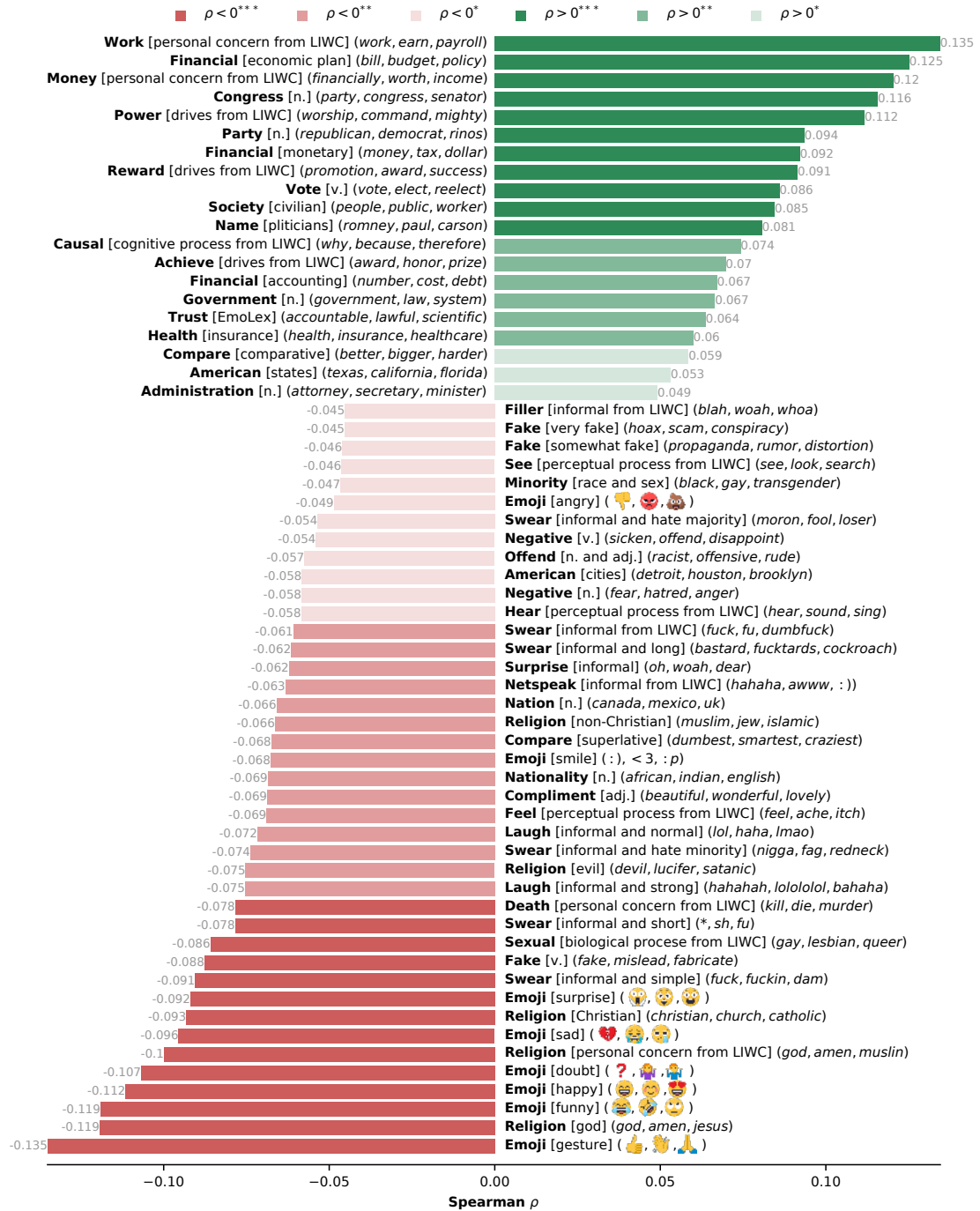


Figure 3.10: **Linguistic signals versus degree of misinformation.** Clusters with significance ρ are plotted, ranked by the sign and strength of correlation. A positive ρ indicates that the statistic increases with veracity, and vice versa. Clusters are labeled in the figure using the format: **name**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

CHAPTER 3. AUDIENCES

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* (👍, 🙌, 🙏, 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 [72, 156], I view them as important proxies for people’s actual emotional state [126, 221, 293] 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 [215]. In my data, I observe an increasing amount of people using negative or offensive words in comments as veracity decreases and swear words increase.

CHAPTER 3. AUDIENCES

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 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 [138].

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* ($t = 2.539^*$). In terms of effect size, combining these clusters together, their mean is 0.0074 for misinformation

CHAPTER 3. AUDIENCES

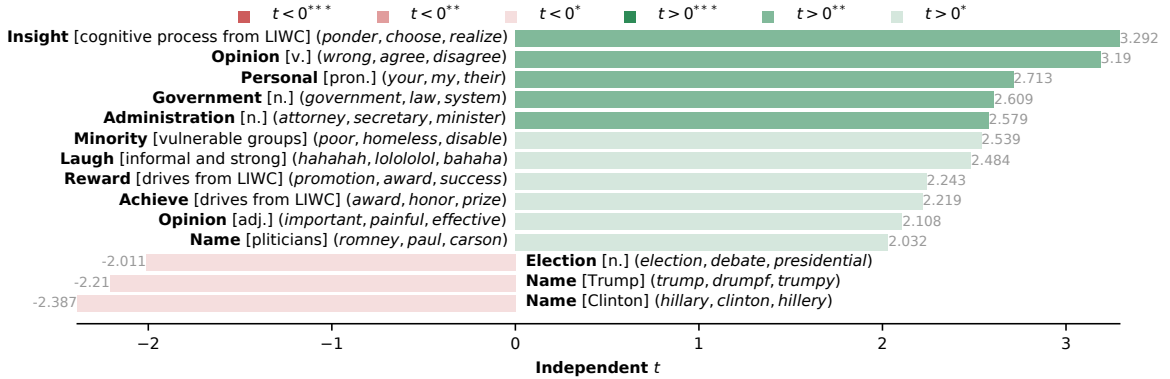


Figure 3.11: **Linguistic signals versus existence of misinformation.** Clusters with significance independent t are plotted, ranked by the sign and strength of difference. A positive t indicates that the mean of the statistic for accurate information is higher than misinformation, and vice versa.

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 [4, 93].

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.com and Snopes.com in user comments with special tokens *politifactref* and *snopesref*, respectively. I use these tokens for my analysis.

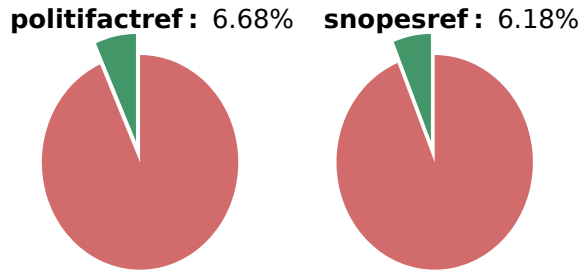


Figure 3.12: **Percentage of reference for PolitiFact.com and Snopes.com.** Each pie chart shows the percentage of posts that contains *politifactref* or *snopesref* over all posts checked by the website.

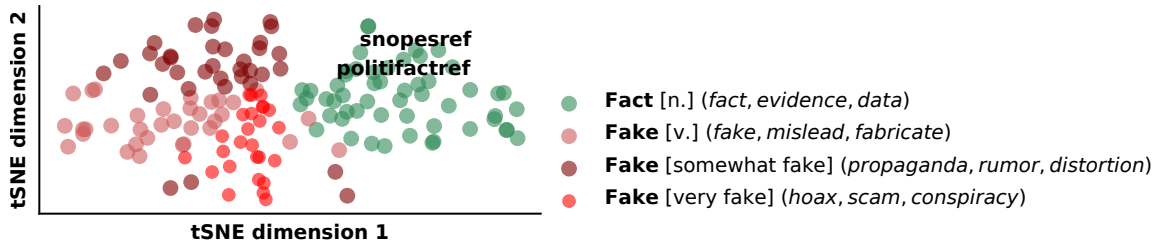


Figure 3.13: **Semantics of reference for PolitiFact.com and Snopes.com.** The learned embedding, which encodes the semantics of *politifactref* or *snopesref*, is plotted along with other words in *fact* and three *fake* clusters. Dimensions are reduced from 100 to 2 using t-SNE. References to PolitiFact.com and Snopes.com 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 3.12 shows the prevalence of *politifactref* and *snopesref*. For all posts that were fact-checked by PolitiFact.com, 6.68% of them have at least one comment that mentioned PolitiFact.com. The number for Snopes.com is similar at 6.18%. This gives me an overview of the prevalence of direct references to PolitiFact.com and Snopes.com 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) [159] to reduce the dimensionality of each vector to 2-dimensional space. As shown in the figure, references to Snopes.com and PolitiFact.com 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

CHAPTER 3. AUDIENCES

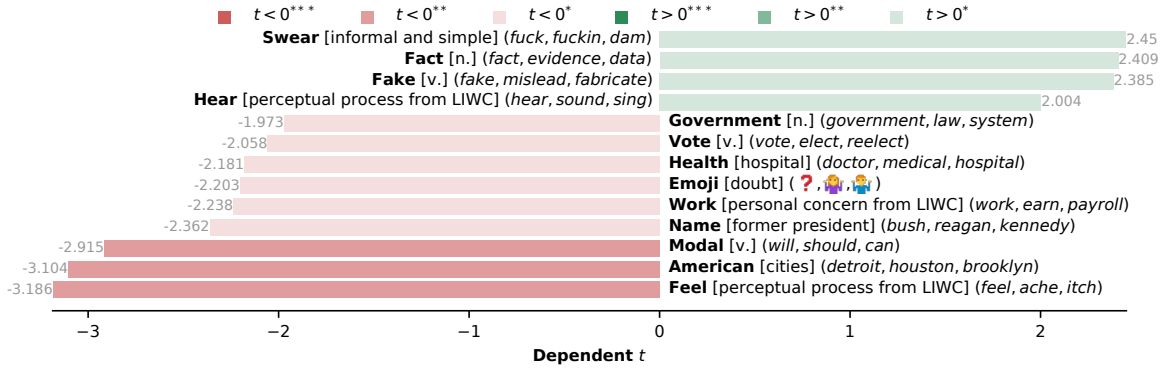


Figure 3.14: **Linguistic signals before and after fact-checking.** Clusters with significance dependent t are plotted, ranked by the sign and strength of difference. A positive t indicates that the mean of the statistic is higher after fact-checking than before, and vice versa.

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.com or Snopes.com, 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%

CHAPTER 3. AUDIENCES

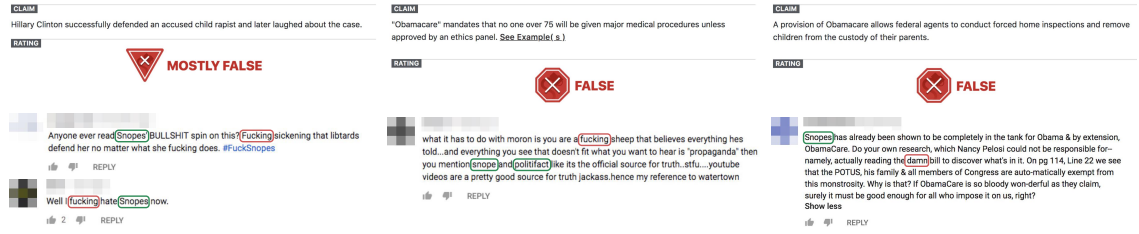


Figure 3.15: **Example comments of the backfire effect.** Three examples are given that include the post veracity from fact-check articles (top) and selected user comments indicating backfire effects (bottom). Words in green blocks (i.e., Snopes.com, PolitiFact.com) are identified as reference to fact-checking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

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 [273].

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 comments also tend to express doubt about the fact-checker themselves because the users perceive them to be biased and unreliable sources [183, 220, 232]. Note that these examples criticized Snopes or PolitiFact.com 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

CHAPTER 3. AUDIENCES

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.com’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 [154] 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 [258]. 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 [180], 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 [247], **(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

CHAPTER 3. AUDIENCES

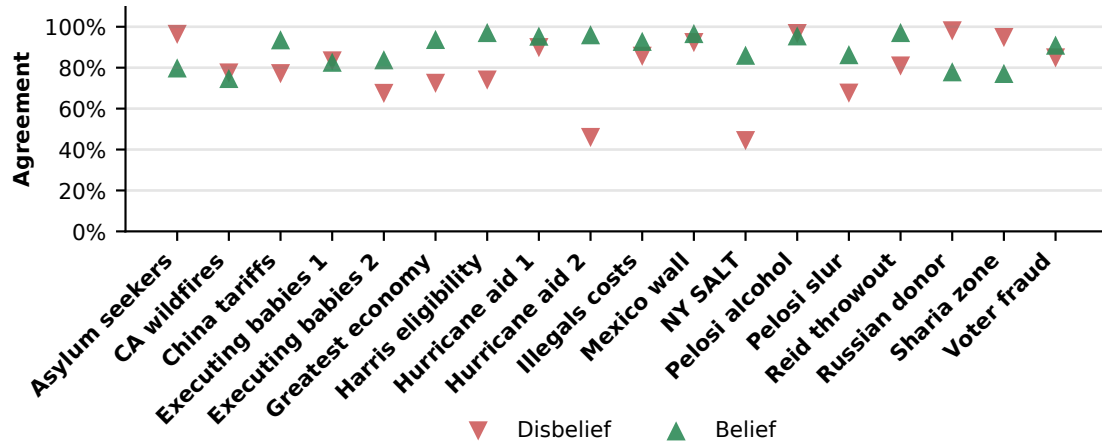


Figure 3.16: **Inter-annotator agreement per claim.** Out of 36 evaluated groups/labels, 66.7% are above 80% agreement and 88.9% are above 70% agreement.

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 agreement³ (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.

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

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

CHAPTER 3. AUDIENCES

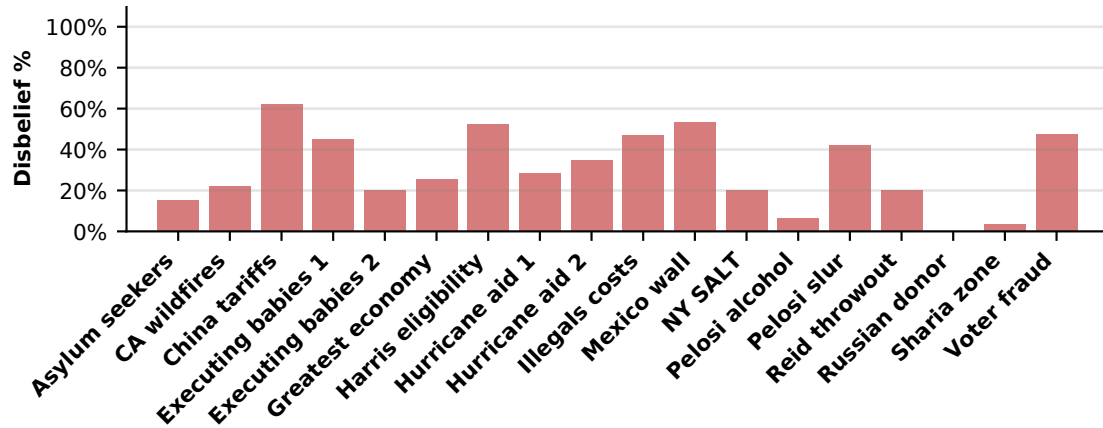


Figure 3.17: **Overview of the disbelief label per claim.** Disbelief distribution across 18 claims. The percentage of disbelief ranged from 0 to 62.4%, with a variance of 0.03.

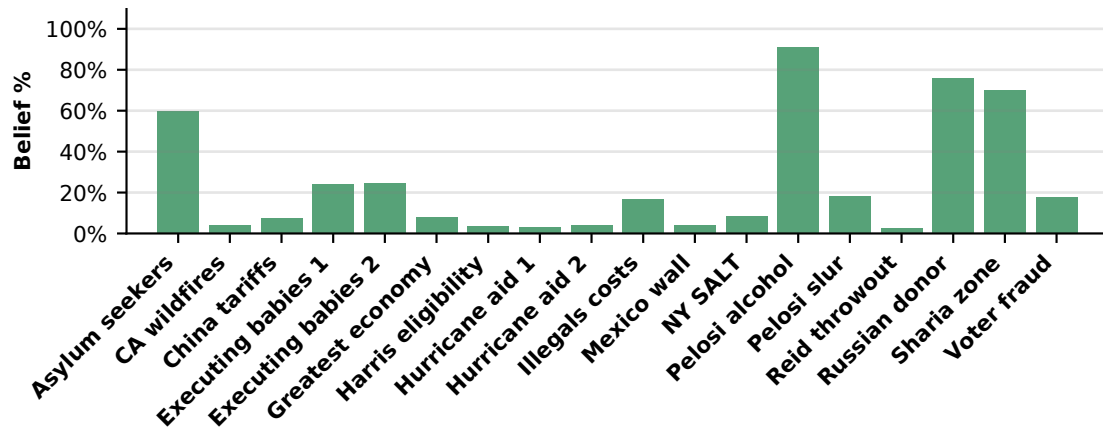


Figure 3.18: **Overview of the belief label per claim.** Belief distribution across 18 claims. The percentage of belief ranged from 2.8% to 91.1%, with a variance of 0.08.

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 (Mis)information and (Dis)belief

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 $r = -0.68^{***}$).

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?*

I adopt a lexicon-based method to explore this question, and choose two lexicons: **(a)** LIWC [254], 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 [117], 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 names⁴ 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, crook, ...” ($t = 10.3^{***}$) and “stupid, dumb, ...” ($t = 8.7^{***}$). These results are intuitive and provide

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

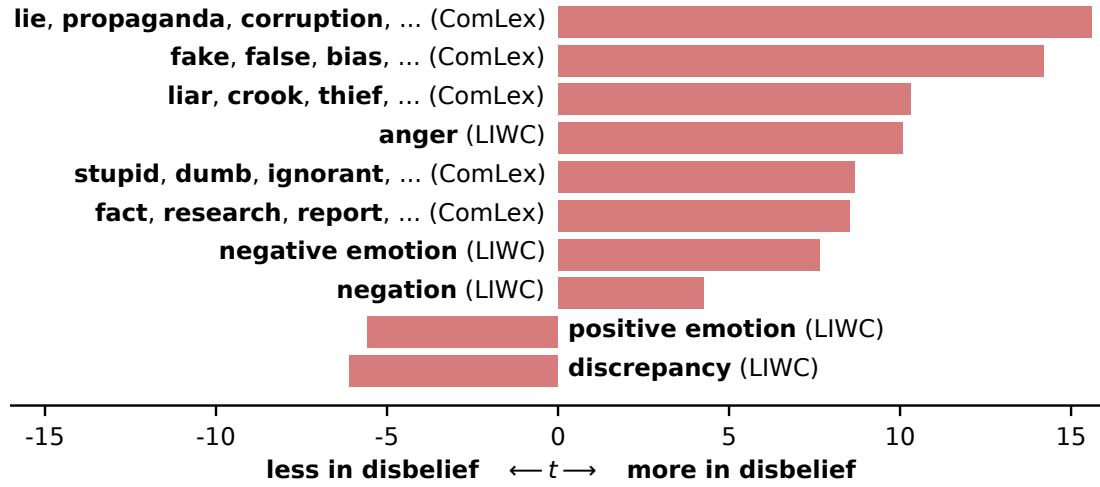


Figure 3.19: **Linguistic difference between tweets expressing disbelief and others.** Tweets expressing disbelief contains more falsehood awareness signals (e.g., “lie”, “fake”, “stupid”) and negative emotions, and less positive emotions and discrepancy. Significance of difference is obtained by t -tests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

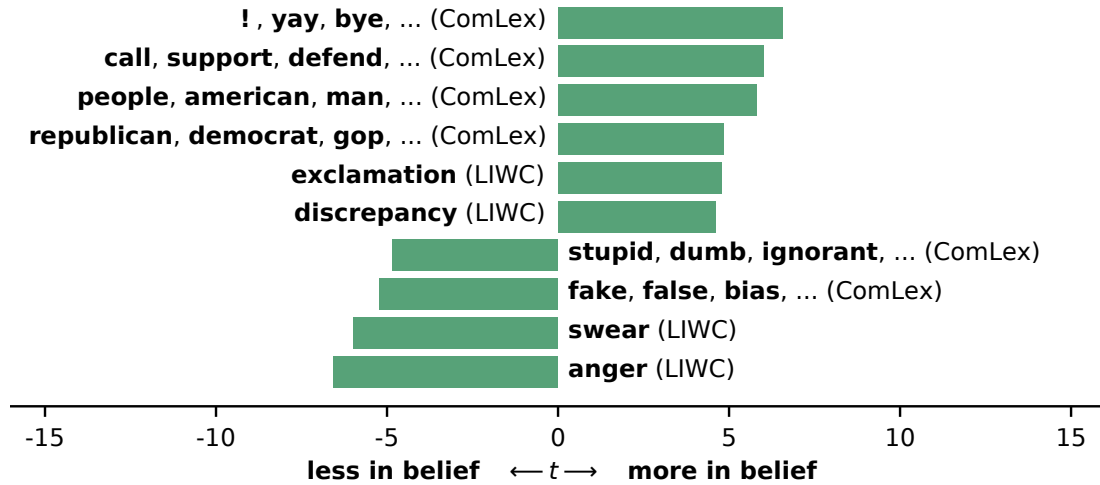


Figure 3.20: **Linguistic difference between tweets expressing belief and others.** Tweets expressing belief contains more exclamation (e.g., “!”, “yay”) and discrepancy, and less falsehood awareness signals (e.g., “lie”, “fake”, “stupid”) and negative emotions. Significance of difference is obtained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

face-validity to the existing linguistic study of misinformation responses, where similar signals were used to insinuate users’ disbelief [117]. 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 [196] methods for the task. I experiment with three pre-trained models: BERT [55], XLNet [286], and RoBERTa [155].

This method follows a *pre-training-fine-tuning* paradigm. During the *pre-training* phase, transformer [261] or transformer-XL [52] 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.

CHAPTER 3. AUDIENCES

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.

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

⁵Although 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 [112].

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

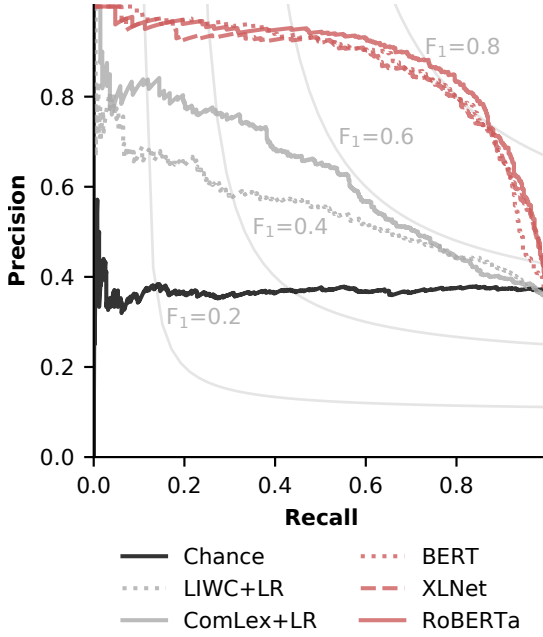


Figure 3.21: **Precision-recall curves for predicting disbelief.** Linear classifiers (LIWC+LR and ComLex+LR) achieve best binary- F_1 scores near 0.6, and neural-transfer classifiers (BERT, XLNet, and RoBERTa) achieve best binary- F_1 scores around 0.8. Isolines for binary- F_1 scores are shown.

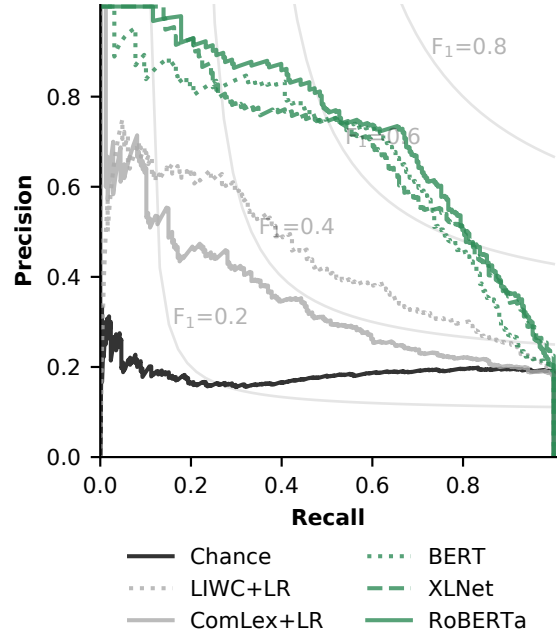


Figure 3.22: **Precision-recall curves for predicting belief.** Linear classifiers (LIWC+LR and ComLex+LR) achieve best binary- F_1 scores near 0.5, and neural-transfer classifiers (BERT, XLNet, and RoBERTa) achieve best binary- F_1 scores around 0.7. Isolines for binary- F_1 scores are shown.

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 [155] for generic NLP tasks.

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

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, 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	Negative	
Predictions	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
		TP+FN	FP+TN	N

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.,

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

and therefore,

$$FP(\tau) = FN(\tau). \quad (3.2)$$

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

CHAPTER 3. AUDIENCES

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

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

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 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.com and PolitiFact.com’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.

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.com or PolitiFact.com — these claims contain no misinformation, and their responses were shown to follow distinctive patterns versus others [117]; **(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.

CHAPTER 3. AUDIENCES

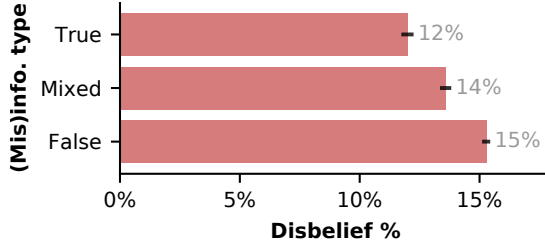


Figure 3.23: **Overall prevalence of expressed disbelief.** For true/mixed/false claims on social media, 12%/14%/15% of comments express disbelief. As the veracity of the claims decreases, the prevalence of expressed disbelief increases.

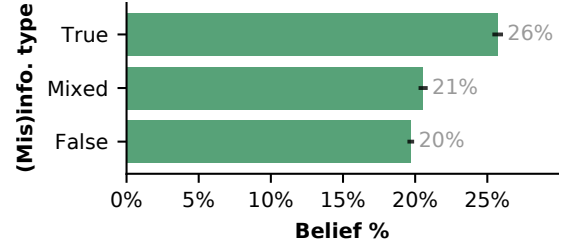


Figure 3.24: **Overall prevalence of expressed belief.** For true/mixed/false claims on social media, 26%/21%/20% of comments express belief. As the veracity of the claims decreases, the prevalence of expressed belief also decreases.

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 [9, 172, 187].

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 [92, 93, 103]. 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 [102, 210]. 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 ,

CHAPTER 3. AUDIENCES

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	

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:

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

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 $\beta_0, \beta_1, \beta_2$ 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 jointly to estimate their respected effects [142]. 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 $\hat{\beta}_2$.

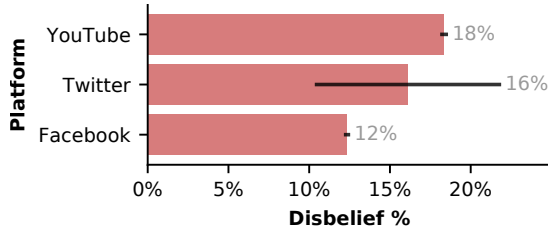


Figure 3.25: **Platforms difference of expressed disbelief.** Facebook comments express less disbelief than YouTube. However, the difference is not significant for Twitter.

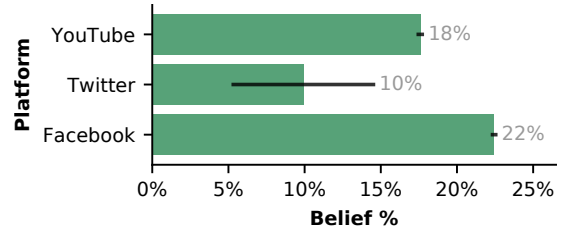


Figure 3.26: **Platforms difference of expressed belief.** Facebook comments express more belief than YouTube, and YouTube comments express more belief than Twitter.

As shown in 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 [128, 256]. 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 [82, 96, 253]. 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.

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

This section summarizes the chapter on audiences' response.

3.7.1 Research Questions and Answers

In this chapter, I investigated 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.

3.7.2 Limitations

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 my 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 extent this exclusion affects my measurement. Second, most of the articles from Snopes.com and PolitiFact.com are focused on politics or political issues, therefore my measurement is also heavily focused on these topics. Other popular misinformation topics, such as health [27] or scientific [69] 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 [165] would suggest that social media users are aware of being observed by the public and thus change their behaviors. Social identity [245] and normative influence theory [129] 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 my approach. Therefore, I emphasize that my study measures *expressed* (dis)belief in the misinformation ecosystem, and my results should be interpreted together with existing qualitative and experimental studies [9, 187].

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 [231], and the IRA had very limited commenting activity comparing to the entire Twitter population [107, 292]. I compared my training dataset verses an IRA account dataset released by Twitter and found no overlap [81]. Therefore, the existence of bots should have minimal effects on my 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 my measurement.

3.7.3 Concluding Thoughts

This chapter delivered 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, and a positive effect of fact-checks. However, these results do not undermine the fundamentally concerning consequences of misinformation, especially since I 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

Platforms play an essential role in how misinformation reaches its audience. In this chapter, I focus on platforms' moderation practices by investigating YouTube as a case study. I explore **RQ2.1** to **RQ2.4**:

- **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 considered variable *does not effect* the outcome, i.e., I rewrite **RQ2.1** to **RQ2.4** as:

- **H1a₀**, *the political leaning of a video does not affect the moderation decision of its comments.*
- **H1b₀**, *the extremeness of a video does not affect the moderation decision of its comments.*
- **H2a₀**, *the veracity of content in a video does not affect the moderation decision of its comments.*
- **H2b₀**, *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. The rest of the chapter is organized as follows: § 4.1 introduces how different control and treatment variables are connected in my dataset, § 4.2 introduces formal criteria to measure bias, § 4.3 uses these criteria as null hypothesis and answers **RQ2.1** to **RQ2.4**, § 4.4 conducts robustness checks and sensitivity analysis on my answers, and finally, § 4.5 summarizes.

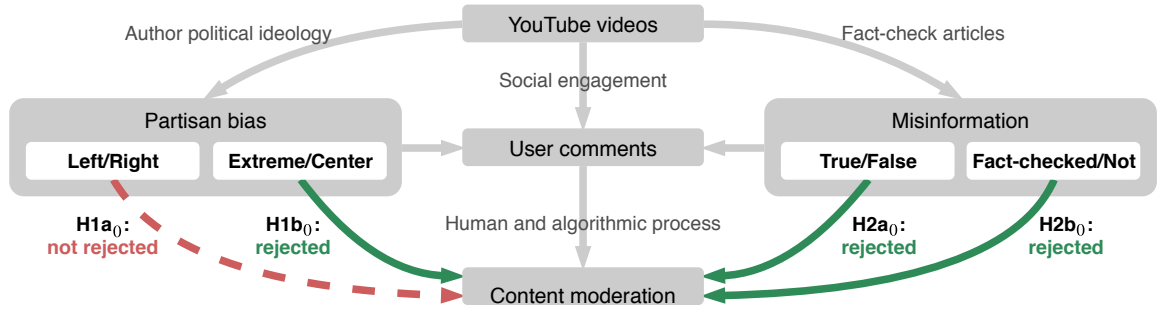


Figure 4.1: **Conceptual framework of four hypotheses.** I investigate the effect of partisanship (i.e., left/right, extreme/center) and misinformation (i.e., true/false, fact-checked/not) on comment moderation. Potential confounders include social engagement on YouTube videos (e.g., views and likes)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

4.1 Platforms' Moderation on Misinformation - an YouTube Dataset

To test hypotheses **H1a₀-H2b₀**, 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), *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.com and PolitiFact.com 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.com [242] that fact-checked a YouTube video from Red State Media [217].

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.

CHAPTER 4. PLATFORMS

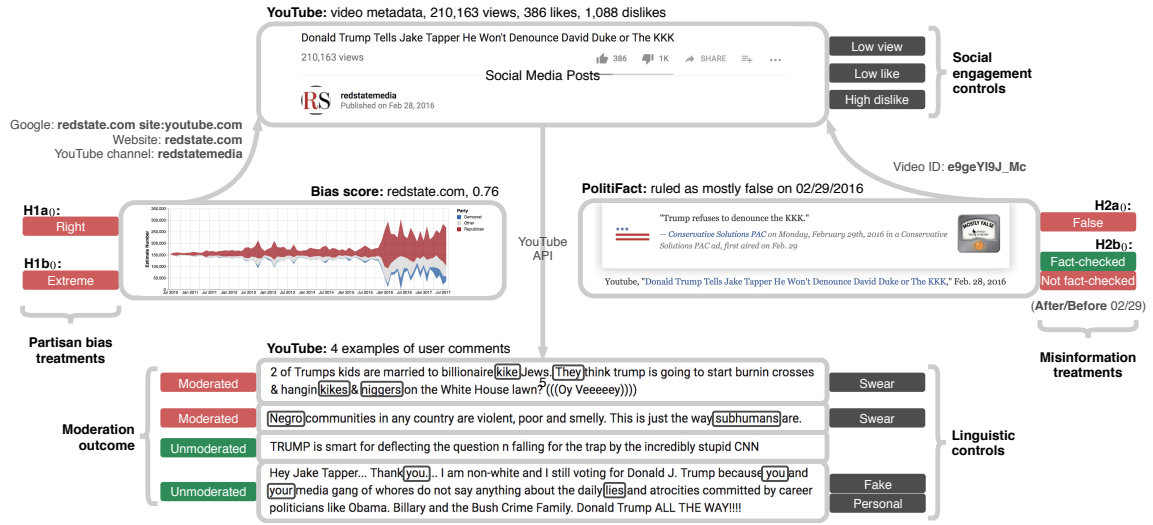


Figure 4.2: **Data collection process and an illustrative example.** Starting from a fact-check article on PolitiFact.com, I collect the misinformation treatment and a YouTube video ID. Another starting point is the partisan score for the website “redstate.com”, where I collect the partisanship treatment 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.

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 [117]. To gather this information, I leverage partisan scores from previous research [223]. 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.

CHAPTER 4. PLATFORMS

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 [218]. 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 H2a₀ and whether each comments was posted before (0) or after (1) the video was fact-checked for H2b₀. The dataset from Jiang and Wilson already contains articles from Snopes.com and PolitiFact.com 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

¹I discuss results using alternative thresholds in later sections.

²Thus, 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.

CHAPTER 4. PLATFORMS

normalizing them with the number of views to get like/dislike rates per video, then bin them in the same manner as views.³

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 [287] and have been found to affect moderation in practice [41, 235].

For this task, I use an existing lexicon called *Comlex* [117] 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 [147].⁴ I apply standard text pre-processing techniques to the comments in my dataset using NLTK [31] (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

³This 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.)

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

⁵This 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.

CHAPTER 4. PLATFORMS

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

the *personal* category, therefore these variables are 1. All other linguistic variables that contains no words are labeled as 0.

4.1.6 Overview of YouTube Videos and Comments

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 [22]. 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* [106]. In this section, I use **H1a₀** (political leaning) as an example to introduce these two criteria, and they apply to **H1b₀**, **H2a₀**, and **H2b₀** in the same way.

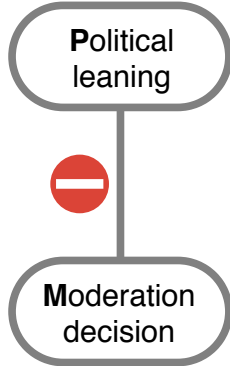


Figure 4.3: **Graph models of the independence criterion.** Null hypothesis \mathbf{H}_0^{ind} : $M \perp\!\!\!\perp P$.

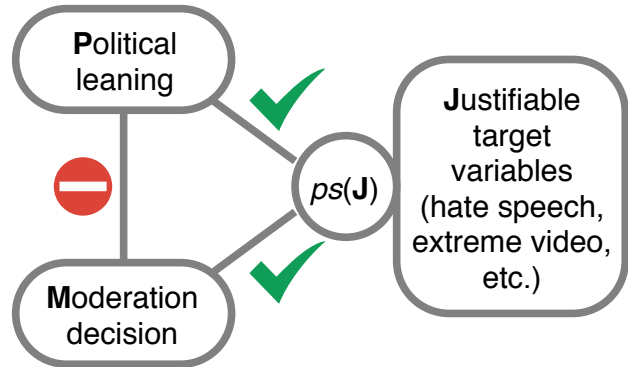


Figure 4.4: **Graph models of the separation criterion.** Propensity scoring function $ps(J)$ is used to summarize J to a scala, hence 2nd null hypothesis \mathbf{H}_0^{sep} : $M \perp\!\!\!\perp P \mid ps(J)$.

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 $P = \{\text{left}, \text{right}\}$ and moderation decision $M = \{\text{moderated}, \text{alive}\}$. This criterion requires these two variables to satisfy $M \perp\!\!\!\perp P$, which, given that P is a binary variable, is equivalent to:

$$\mathbb{P}\{M \mid P = \text{left}\} = \mathbb{P}\{M \mid P = \text{right}\} . \quad (4.1)$$

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 [104, 223], 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 $M \perp\!\!\!\perp P \mid J$, which, given that P is a binary variable, is equivalent to $\forall J$:

$$\mathbb{P}\{M \mid P = \text{left}, J\} = \mathbb{P}\{M \mid P = \text{right}, J\}. \quad (4.2)$$

This criterion is also widely adopted in previous studies, especially when the correlation between sensitive features and target variables is inherent [13, 137, 255].

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., $f : \mathbb{R}^{|J|} \rightarrow \mathbb{R}$. A particular example of f is *propensity scoring* defined as: $ps(J) := \mathbb{P}\{P = \text{left (or right)} \mid J\}$ [226]. It is proven that if (4.2) holds and $\mathbb{P}\{P \mid J\} \in (0, 1)$, then $\forall ps(J), \mathbb{P}\{P \mid ps(J)\} \in (0, 1)$ and:

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

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 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

⁶The 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.

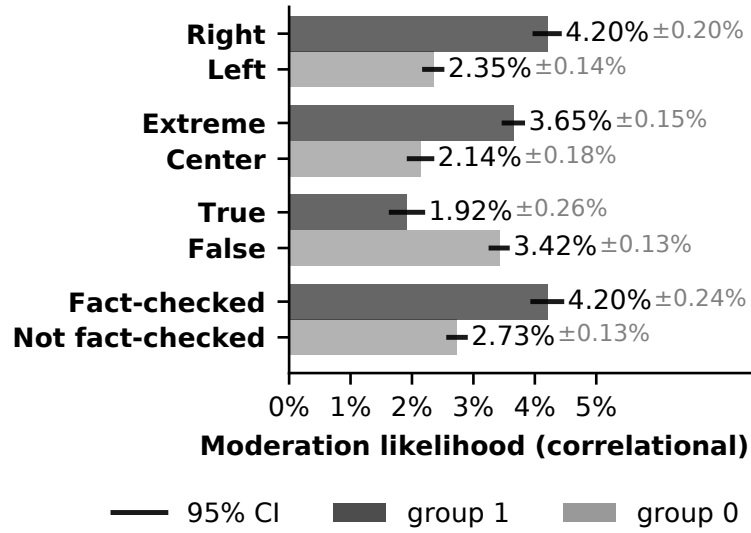


Figure 4.5: **Correlational difference in moderation likelihood.** Moderation likelihood for each group with 95% CI is shown. All four null hypotheses are rejected.

(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 H1a₀, 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 [124, 260]. For H1b₀, 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 [42, 179, 182].

For H2a₀, 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 H2b₀, 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 [6, 195, 209] by actively partnering with fact-checking organizations [88, 166].

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, **I do not endorse the findings presented in Figure 4.5.** Rather, I present these results merely to highlight

	Engagement controls (M-W U test)					Linguistic controls (χ^2 test)							Treatments (χ^2 test)			
Right - Left	+	-	+	-	+	●	-	+	+	+	+	+	+	-	-	×
Extreme - Center	+	+	-	+	-	+	+	-	-	-	-	-	-	+	×	-
True - False	-	+	+	-	●	●	●	●	+	●	●	+	●	×	+	-
Fact-checked - Not	+	×	×	×	+	-	-	+	+	+	●	+	×	●	-	+
	Moderation	Like	Dislike	View	Swear	Laugh	Emoji	Fake	Administration	America	Nation	Personal	Fact-checked - Not	True - False	Extreme - Center	Right - Left
<div><div><div>+</div><div>positive, $p < 0.001$</div></div><div><div>-</div><div>negative, $p < 0.001$</div></div></div> <div><div>+</div><div>positive, $p < 0.01$</div></div> <div><div>-</div><div>negative, $p < 0.01$</div></div> <div><div>+</div><div>positive, $p < 0.05$</div></div> <div><div>-</div><div>negative, $p < 0.05$</div></div> <div><div>●</div><div>not significant</div></div> <div><div>×</div><div>not applicable</div></div>																

Figure 4.6: **Correlational difference for confounding variables.** The 1st column repeats the observations I made for moderation likelihood. The 2nd to 4th columns show how social engagement correlates with hypothesized variables, the 5th to 12th columns show linguistic features, and 13th to 16th 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.

why a person might erroneously believe that comment moderation on YouTube exhibits partisan bias.

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 1st column repeats my observations from Figure 4.5. The 2nd to 4th columns show how social engagement on videos correlates with the hypothesized variables, while the 5th to 12th columns show correlations with linguistic features. Finally, the 13th to 16th 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 H1a₀ 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.

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

CHAPTER 4. PLATFORMS

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 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 H1a₀: 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 [186, 225, 270]. 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 [79, 208]. 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 [189, 280].

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” [200], e.g., H1a₀ 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 [248]. The latter two have been used within the Computer Science community [40, 43, 78, 193]. 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 [226] that has been used extensively in the social [255], psychological [137], and biological [13] literatures.

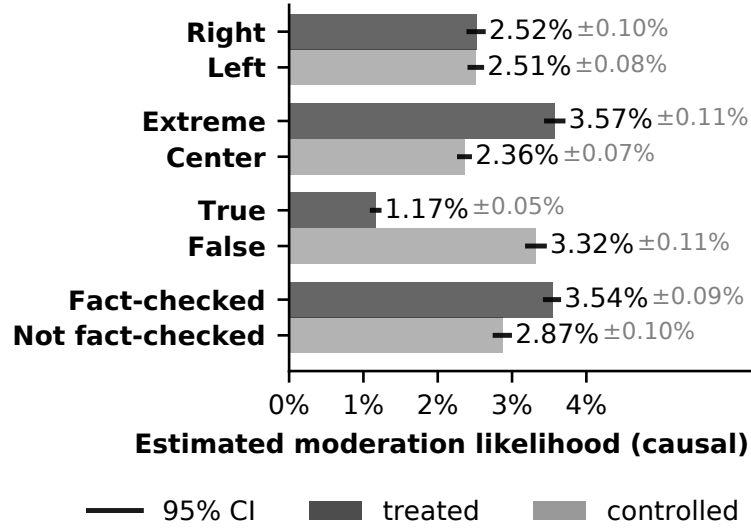


Figure 4.7: **Causal difference in moderation likelihood.** Moderation likelihood for controlled and treated groups with 95% CI is shown. H_{1a0} is no longer rejected. Differences in the other 3 hypothesized variables are also changed.

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) [226].

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%

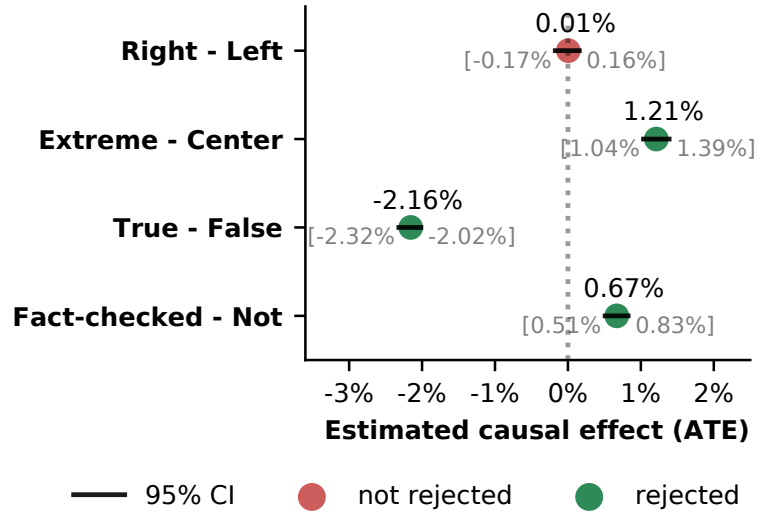


Figure 4.8: **Estimation of causal effect.** Average treatment effect (ATE) with 95% CI is shown. Significance level for null hypothesis is encoded with color. CIs using bootstrap are considered as conservative estimates.

bootstrapped CI⁸ 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

⁸A recent study showed that such CIs are conservative estimates [14].

⁹Note 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.

CHAPTER 4. PLATFORMS

moderation, since I observe that there are a greater number of ideologically extreme right-leaning channels than similarly extreme left-leaning channels on YouTube.

H2a₀ 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.

H2b₀ 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 H1a₀ for a spectrum of r from 0% to 50%. Although the effect size for H2a₀ and

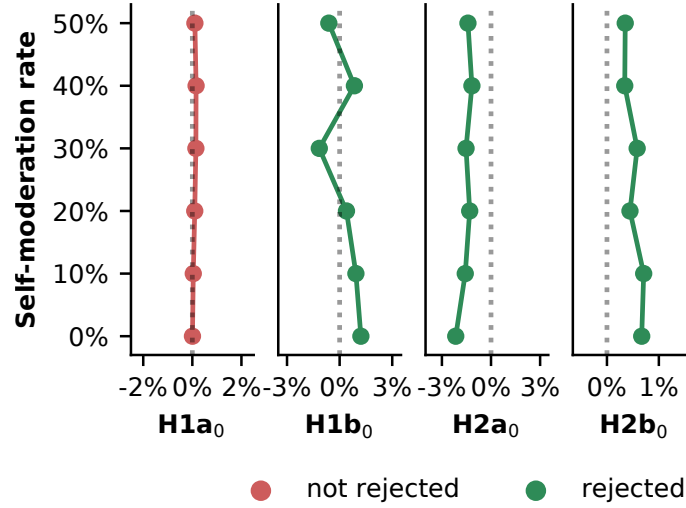


Figure 4.9: **Simulation of user moderation.** The effect of self moderation is minimal for H1a₀, H2a₀, and H2b₀, but H1b₀ does not hold under high rates ($r > 20\%$).

H2b₀ fluctuate as r increases, the direction of their effects are robust. The only exception is H1b₀: the direction of its causal effect does not hold when $r > 20\%$.

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.com and PolitiFact.com, which are both confirmed by the International Fact-Checking Network to be non-partisan, fair, and transparent [212]. However, there are still accusations that their ratings are biased against political conservatives [183, 220, 232]. 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 chapter, 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

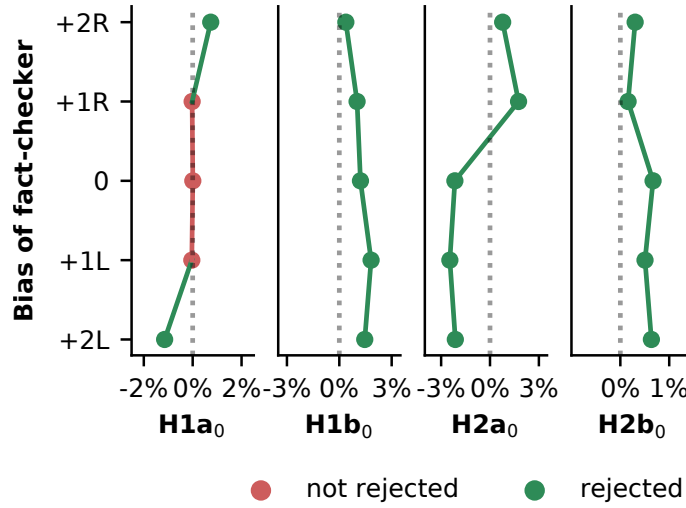


Figure 4.10: **Simulation of biased fact-checkers.** The effect of fact-checker bias is minimal for H1b₀ and H2b₀, and minimal for H1a₀ when bias is low ($\lambda \leq +1$).

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, $b = +1R$ 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 $b = +2R$, 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. H2a₀ is impacted the most, since it directly concerns video veracity. In contrast, the effect sizes of H1b₀ and H2b₀ fluctuate, but the direction of their effects are robust. For H1a₀, the result does not change with slight bias ($\lambda \leq +1$), but does change when fact-checkers are highly biased. Consider $b = +2R$, 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 $b = +2L$. Similarly, the results of H2a₀ also change in the same direction.

Note that **I 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.

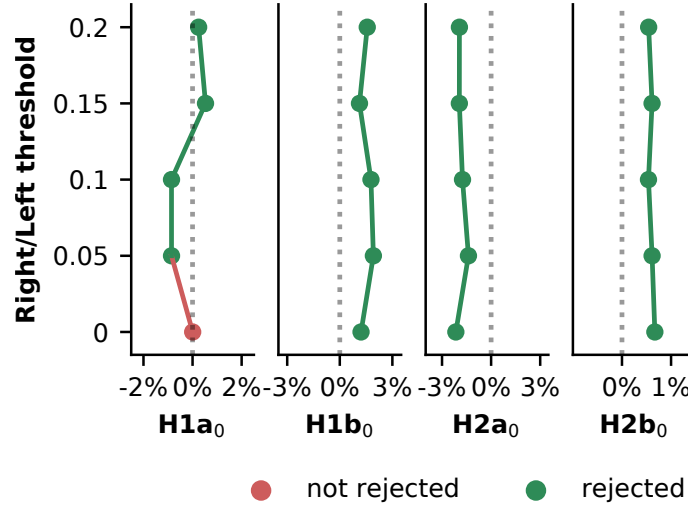


Figure 4.11: **Alternative H1a₀ (left/right) thresholds.** The effect of left/right thresholds is minimal for H1b₀, H2a₀ and H2b₀, but results for H1a₀ do not hold.

4.4.3 Alternative Thresholds and Control Variables

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

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 [223]. 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 H1b₀, H2a₀, and H2b₀, but does impact H1a₀.¹⁰ 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 H1a₀ 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 H1b₀ 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 [201, 254]. Although the

¹⁰This is partially due to the partisan bias scores of comments in my dataset not being balanced between left and right.

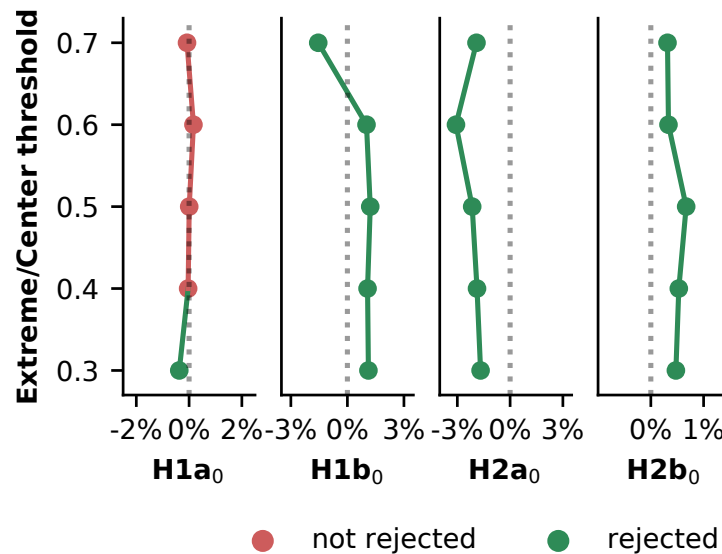


Figure 4.12: **Alternative H1b₀ (extreme/center) thresholds.** The effect of extreme/center thresholds is minimal for most hypotheses, except for H1a₀ and H1b₀.

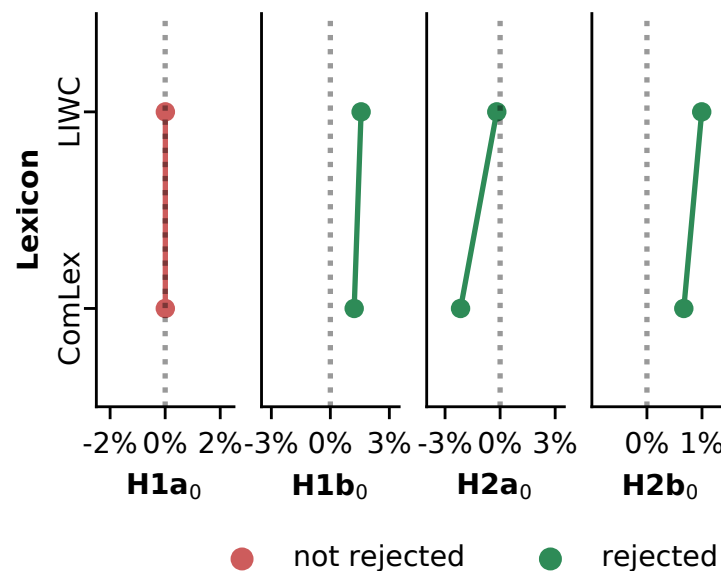


Figure 4.13: **Alternative linguistic controls.** The effect of alternative linguistic controls using lexicon LIWC instead of ComLex is minimal for all hypotheses.

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

¹¹Determined by a preliminary linear regression for $p < 0.001$.

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

This section summarizes the chapter on platforms' moderation practice using YouTube as a lens.

4.5.1 Research Questions and Answers

In this chapter, I investigated 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 ~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.

4.5.2 Limitations

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* [163], 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* [224], 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

CHAPTER 4. PLATFORMS

and highly relevant given YouTube’s community guidelines [287] and prior studies [41, 117, 235]. 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 [130]. Nevertheless, the results from propensity scoring show significant improvement comparing to correlational analysis [53]. 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 [6, 42, 179, 182, 195, 209]. Or it could be that fact-check messages draw more efforts from concerned users to flag content for moderation [88, 166].

Representation and generalization. The YouTube videos in my dataset are covered by the datasets from [117] and [223], 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 [45, 76, 171]. 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.

4.5.3 Concluding Thoughts

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 [38, 40, 61, 62, 139]. 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 have refuted here — anti-conservative bias in moderation practices. Indeed, this is both a motivation of

CHAPTER 4. PLATFORMS

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 [34], 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 [281].

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 [131, 132]. 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” [40]. 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.

Chapter 5

Storytellers

Storytellers generate misinformation and then release them onto platforms. In this part, I structurize storytellers' strategies and explore prevalent types of misinformation to date. I formulate **RQ3.1** to **RQ3.4**:

- **RQ3.1**, *what are the prevalent types of misinformation stories in the US over the last ten years?*
- **RQ3.2**, *how has the prevalence of misinformation types evolved over the last ten years?*
- **RQ3.3**, *how has the prevalence of misinformation types evolved between the 2016 and the 2020 US presidential elections?*
- **RQ3.4**, *how has the prevalence of misinformation types evolved between the H1N1 and the COVID-19 pandemics?*

The snippets of the discovered misinformation structure is shown in Figure 5.1. The rest of the chapter is organized as follows: § 5.1 introduces rationalized neural models, § 5.2 conducts experiments on public datasets to evaluate the performance of rationalization methods, § 5.3 conducts experiments on fact-checks and structurize misinformation types, 5.4 applies the structure and explores the evolution of misinformation stories, and finally, § 5.5 summarizes.

5.1 Rationalized Neural Models

Realizing my intuition (as described in § 1.3) requires neural models to (at least shallowly) reason about predictions. In this section, I introduce existing rationalized neural models and propose to

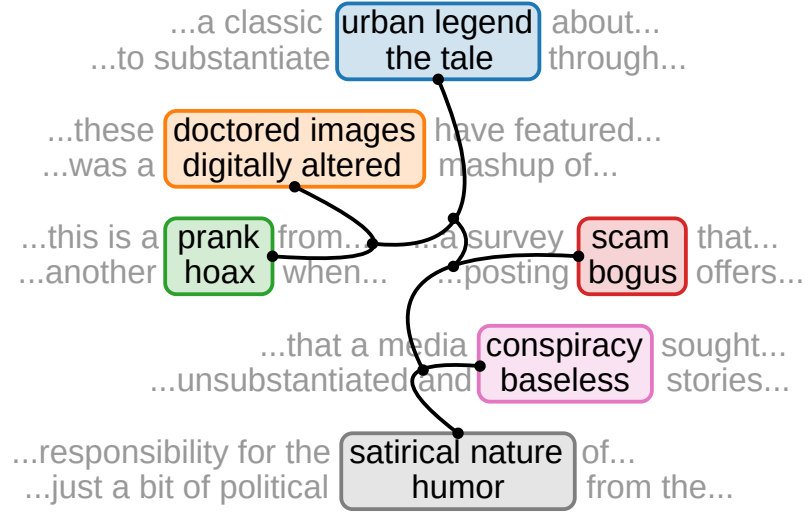


Figure 5.1: **A snippet of the misinformation structure.** Each line is a snippet from a fact-check. Key phrases identifying the misinformation types are highlighted. Phrases with similar semantics are clustered in colored boxes. This structure is a sample of my final results.

include domain knowledge as weak supervision in the rationalizing process.

5.1.1 Problem Formulation and Rationalization Methods

In a standard text classification problem, each instance is in a form of (\mathbf{x}, \mathbf{y}) . $\mathbf{x} = [\mathbf{x}^i] \in V_x^l$ is the input token sequence of length l , where V_x is the vocabulary of the input and i is the index of each token \mathbf{x}^i . $\mathbf{y} \in \{0, 1\}^m$ is the binary label of length m . Rationalization requires a model to output the prediction $\hat{\mathbf{y}}$ together with a binary mask $\mathbf{z} = [\mathbf{z}^i] \in \{0, 1\}^l$ of input length l , indicating which tokens are used (i.e., $\mathbf{z}^i = 1$) to make the decision. These tokens are called *rationales*.

Hard rationalization requires a model to directly output \mathbf{z} . Initially proposed by [143], the model first passes the input \mathbf{x} to a tagger¹ module and samples a binary mask \mathbf{z} from a Bernoulli distribution, i.e., $\mathbf{z} \sim \text{Tagger}(\mathbf{x})$, and then uses only unmasked tokens to make a prediction of \mathbf{y} , i.e., $\hat{\mathbf{y}} = \text{Predictor}(\mathbf{z}, \mathbf{x})$.²

The loss function of this method contains two parts. The first part is a standard loss for the prediction $L_y(\hat{\mathbf{y}}, \mathbf{y})$, which can be realized using common classification loss, e.g., cross entropy. The second part is a loss $L_z(\mathbf{z})$ aiming to regularize \mathbf{z} and encourage conciseness and contiguity of rationale selection, formulated by [143]. Recent work proposed to improve the initial model with an

¹This module was named *generator* by [143]. I name it *tagger* to distinguish it from the NLG problem.

²This module was named *encoder* by [143]. I name it *predictor*, consistent with [288], to distinguish it from the encoder-decoder framework.

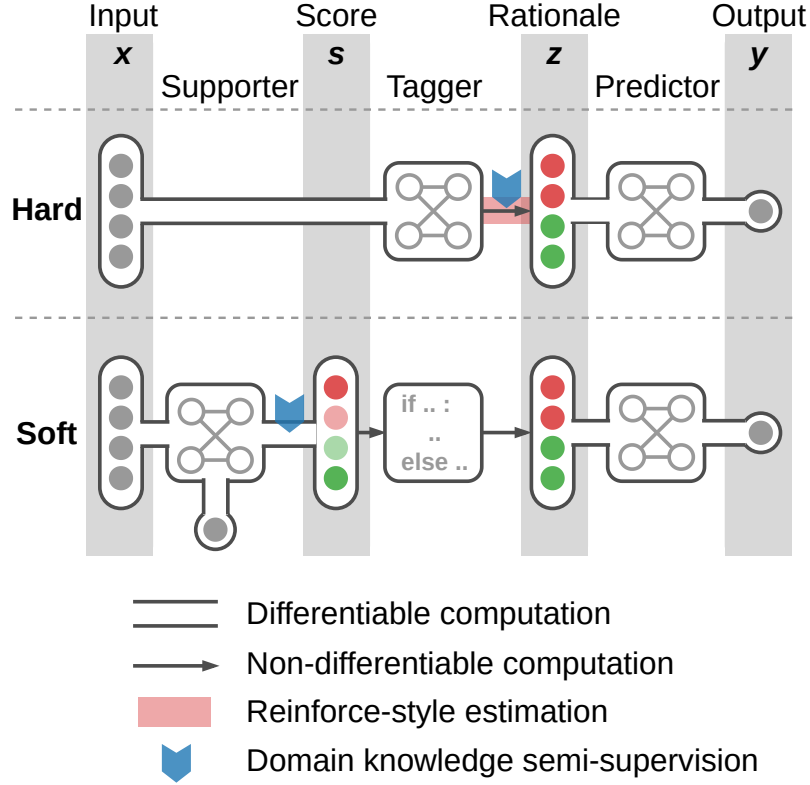


Figure 5.2: **Hard and soft rationalization methods.** Hard rationalization is an end-to-end model that first uses input x to generate rationales z , and then uses unmasked tokens to predict y . Soft rationalization is a three-phased model that first uses input x to predict y and outputs importance scores s , then binarizes s to rationales z , and finally uses unmasked tokens to predict y again as evaluation for faithfulness.

adversarial component [37, 288]. Combining these parts together, the model is trained end-to-end using reinforce-style estimation [278], as sampling rationales is a non-differentiable computation. The modules of hard rationalization are illustrated in Figure 5.2.

Soft rationalization, in contrast, allows a model to first output a continuous version of importance scores $s = [s^i] \in \mathbb{R}^l$, and then binarize it to get z . Initially formalized by [110] as a multiphase method, the model first conducts a standard text classification using a supporter module $\hat{y} = \text{Supporter}(x)$ and outputs importance scores s , then binarizes s using a tagger module, i.e., $z = \text{Tagger}(s)$, and finally uses only unmasked tokens of x to make another prediction \hat{y} to evaluate the faithfulness of selected rationales.³

³The second and third modules were named *extractor* and *classifier* by [110]. I continue using *tagger* and *predictor* to align with the hard rationalization method.

These three modules are trained separately in three phases.⁴ Since the supporter and predictor are standard text classification modules the only loss needed is for the prediction $L_y(\hat{\mathbf{y}}, \mathbf{y})$. This method is more straightforward than the hard rationalization method, as it avoids non-differentiable computations and the instability induced by reinforce-style estimation. The modules of soft rationalization are also illustrated in Figure 5.2.

The popular *attention* mechanism [19] provides built-in access to \mathbf{s} . Although there have been debates on the properties achieved by attention-based explanations [109,230,277], rationales extracted by straightforward rules on attention weights were demonstrated as comparable to human-generated rationales [110]. Additionally, in my use case I only need the rationales themselves as key phrases and do not require them to faithfully predict \mathbf{y} , therefore the last predictor module can be omitted.

Loss functions. For the predictive loss $L_y(\hat{\mathbf{y}}, \mathbf{y})$, I use a common cross entropy loss function.

For the rationale regularization loss $L_z(\mathbf{z})$, it contains two parts as implemented by [288]. The first part is to encourage conciseness:

$$L_{zk}(\mathbf{z}) = \max \left\{ \sum_i z^i - k, 0 \right\},$$

where $\sum_i z^i$ represents the number of selected tokens, and k is a hyperparameter defining a loss-free upper-bound for it. The second part is to encourage contiguity:

$$L_{zl}(\mathbf{z}) = \max \left\{ \sum_i |z^i - z^{i-1}| - l, 0 \right\},$$

where $z^i - z^{i-1}$ denotes a transition between $z^i = 0$ and $z^{i-1} = 1$ or vice versa, therefore $\sum_i |z^i - z^{i-1}|$ represents the number of rationale phrases, and l is another hyperparameter defining a loss-free upper-bound for it.

Combining these two parts together, I can further specify $\lambda_z L_z(\mathbf{z})$ as $\lambda_{zk} L_{zk}(\mathbf{z}) + \lambda_{zl} L_{zl}(\mathbf{z})$.

For domain knowledge weak supervision, I define $L_d(\mathbf{z}, \mathbf{z}_d)$ as:

$$L_d(\mathbf{z}, \mathbf{z}_d) = - \sum_i z^i z_d^i,$$

which decreases loss by 1 if both $z^i = 1$ and $z_d^i = 1$, i.e., selecting a token in the domain knowledge

⁴Tagger is often flexibly designed as a rule-based algorithm, therefore no training is needed.

CHAPTER 5. STORYTELLERS

vocabulary V_d , and has no effect on the loss otherwise. Similarly, I define $L_d(s, z_d)$ as:

$$L_d(s, z_d) = - \sum_i s^i z_d^i,$$

which decreases loss by s^i if $z_d^i = 1$, and has no effect on the loss if $z_d^i = 0$. This encourages the training to increase the importance score s^i on domain knowledge to reduce the loss.

With this implementation, there are five hyperparameters to search for the hard rationalization method: λ_{zk} , k , λ_{zl} , l and λ_d , and only one hyperparameter to search for the soft rationalization method: λ_d .

5.1.2 Domain Knowledge as Weak Supervision

Both hard and soft rationalization methods can be trained with or without supervision w.r.t. rationales z [56]⁵. When rationales are selected in an unsupervised manner, the model would intuitively favor rationales that are most informative to predict the corresponding label as a result of optimizing the loss function. This could result in some undesirable rationales in my case: for example, certain entities like “COVID-19” or “Trump” that are highly correlated with misinformation would be selected as rationales even though they do not suggest any misinformation types. Therefore, I propose to weakly supervise⁶ the rationalizing process with domain knowledge to obtain specific, desired types of rationales.

Assuming a lexicon of vocabulary V_d as domain knowledge, I reprocess the input and generate weak labels for rationales $z_d = [z_d^i] \in \{0, 1\}^l$ where $z_d^i = 1$ (i.e., unmasked) if $x^i \in V_d$ and $z_d^i = 0$ (i.e., masked) otherwise. Then, I include an additional loss item $L_d(z, z_d)$ or $L_d(s, z_d)$ for the hard or soft rationalization method.

Combining the loss items together, the objective for the end-to-end hard rationalization model is:

$$\min_{\theta} L_y(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_z L_z(\mathbf{z}) + \lambda_d L_d(\mathbf{z}, \mathbf{z}_d),$$

where θ contains the parameters to estimate and $\lambda_{(\cdot)}$ are hyperparameters weighting loss items.

Similarly, the objective function for the first phase of soft rationalization is:

$$\min_{\theta} L_y(\hat{\mathbf{y}}, \mathbf{y}) + \lambda_d L_d(\mathbf{s}, \mathbf{z}_d).$$

⁵They are trained with supervision w.r.t. the label \mathbf{y} .

⁶Since there is inherently no ground-truth of misinformation types in fact-check articles.

5.2 Rationalizing Public Datasets

I conduct experiments on public datasets to evaluate the performance of hard and soft rationalization methods, particularly for my needs, and confirm that including domain knowledge as weak supervision helps with the rationalizing process.

5.2.1 Datasets, Domain Knowledge, and Experimental Setup

I first introduce the selection process for public datasets and lexicons as domain knowledge, as well as the experimental setup.

Datasets selection. An ideal dataset for my models should meet the following requirements: **(a)** formulated as a text classification problem, **(b)** annotated with human rationales, and **(c)** can be associated with high quality lexicons to obtain domain knowledge. I select two datasets based on these criteria: the **movie reviews** dataset released by [198] and later annotated with rationales by [289], which contains 2K movie reviews labeled with positive or negative sentiments; and the **personal attacks** dataset released by [283] and later annotated with rationales by [37], which contains more than 100K Wikipedia comments labeled as personal attacks or not.

Domain knowledge. For the sentiment analysis on movie reviews, I use the EmoLex lexicon released by [176], which contains vocabularies of positive and negative sentiments. For identifying personal attacks, I use a lexicon released by [276], which contains a vocabulary of abusive words. With corresponding vocabularies, I generate weak rationale labels z_d for each dataset.

Evaluation metrics. I choose binary precision $\Pr(z)$ to evaluate the quality of extracted rationales, because **(a)** a perfect recall can be trivially achieved by selecting *all* tokens as rationales,⁷ and **(b)** my case of identifying key phrases requires concise rationales. Additionally, I measure the average percentage of selected rationales over the input length $\%(z)$. For predictions, I use macro $F_1(y)$ as the evaluation metric as well as the percentage of information used $\%(x)$ to make the prediction.

Experimental setup. The train, dev, and test sets are pre-specified in public datasets. I optimize hyperparameters for $F_1(y)$ on the dev sets, and only evaluate rationale quality $\Pr(z)$ *after* a model is decided.

⁷I later show that this is the default model behavior if rationale selection is under-regularized.

CHAPTER 5. STORYTELLERS

Module cells. Each module in soft and hard rationalization methods can be implemented with different neural cells. Here, I consider two common types of choices: RNN cells, e.g., LSTM, and transformer cells [261], e.g., BERT [55].

For hard rationalization, the rationale selection process is actively regularized by $L_z(\mathbf{z})$, therefore I simply choose the cell type that optimizes $F_1(\mathbf{y})$ on dev sets, i.e., transformers.

For soft rationalization, the rationale selection process is based on passively generated importance scores (i.e., attention), therefore the inherent behavioral difference between RNN and transformer cells would significantly impact my choice.

In my experiments, I observe that transformer cells often assign strong importance to a single token, but assign near zero weights to its neighboring tokens (possibly as a result of its multi-head attention mechanism), while RNN cells assign strong importance to a single token, but also some residue, fading weights to its neighboring tokens.

Consider the following example, which shows the distribution of importance scores generated by transformer cells, with **darker** text representing higher importance scores and **lighter** text scoring near zero. In the following example, only the token **conspiracy** is selected as rationale:

“...Furthermore, claims that COVID-19 was “manufactured,” or that it “escaped from” this Chinese lab, are nothing more than baseless **conspiracy** theories...”

In contrast, the following example shows the distribution of importance scores generated by RNN cells for the same snippet, i.e., the token **conspiracy** has the strongest importance score, but its neighboring tokens are also assigned some weight above the threshold, and therefore the phrase **baseless conspiracy theories** is selected as rationale:

“...Furthermore, claims that COVID-19 was “manufactured,” or that it “escaped from” this Chinese lab, are nothing more than baseless **conspiracy** theories...”

As I prefer to obtain phrases (i.e., one or more tokens) for rationales, I choose between RNN cells. After optimizing $F_1(\mathbf{y})$ on dev set, I choose bidirectional LSTM initialized with GloVe embeddings [202] for the soft rationalization method.

Hyperparameters. Since the size of dev sets is relatively small in my experiments, a rigorous grid search for hyperparameters might overfit to several instances in the dev set, therefore I tune the hyperparameters manually starting from the hyperparameters released by [288] and [37].

For **movie reviews** [289], the best-performing model for hard rationalization uses $\lambda_{zk} = 5.0$, $k = 240$, $\lambda_{zl} = 5.0$, $l = 10$, and $\lambda_d = 8.0$ with domain knowledge as weak supervision, and the best-performing model for soft rationalization uses $\lambda_d = 0.5$.

		Movie reviews [289]				Personal attacks [37]			
		Pr(z)	%(z)	F ₁ (y)	%(x)	Pr(z)	%(z)	F ₁ (y)	%(x)
h_0	Hard rationalization	0.37	2.7%	0.72	2.7%	0.17	32.5%	0.73	32.5%
h_1	w/ Domain knowledge	0.38	3.7%	0.72	3.7%	0.22	16.9%	0.73	16.9%
h_2	w/o Rationale regularization	0.31	99.9%	0.92	99.9%	0.19	99.9%	0.82	99.9%
h_3	w/ Adversarial components	0.33	2.5%	0.70	2.5%	0.22	14.9%	0.75	14.9%
s_0	Soft rationalization	0.58	3.7%	0.91	100%	0.35	16.9%	0.82	100%
s_1	w/ Domain knowledge	0.62	3.7%	0.92	100%	0.39	16.9%	0.82	100%
s_2	w/ Half rationales	0.64	1.9%	0.92	100%	0.46	8.4%	0.82	100%
s_3	w/ Double rationales	0.55	7.4%	0.92	100%	0.31	33.8%	0.82	100%

Table 5.1: **Evaluation results for hard and soft rationalization methods.** My experiments show that: **(a)** hard rationalization requires a sensitive hyperparameter λ_z to regularize rationales (h_2 to h_0); **(b)** soft rationalization achieves the best F₁(y) overall, but Pr(z) depends on the rationale extraction approach (s_2/s_3 to s_0); **(c)** domain knowledge as weak supervision improves Pr(z) for both hard (h_1 to h_0) and soft (s_1 to s_0) rationalization while maintaining similar %(z) and F₁(y); **(d)** soft rationalization achieves better Pr(z) in a fair comparison (s_1 to h_1).

For **personal attacks** [37], the best-performing model for hard rationalization uses $\lambda_{zk} = 5.0$, $k = 7$, $\lambda_{zl} = 5.0$, $l = 1$, and $\lambda_d = 10.0$ with domain knowledge as weak supervision, and the best-performing model for soft rationalization uses $\lambda_d = 0.5$.

Model size, computing machine and runtime. The number of parameters is 325K for hard rationalization models, and 967K for soft rationalization models. All experiments were conducted on a 12GB Nvidia Titan X GPU node, and finished training within an hour per experiment.

The evaluation results for all my experiments on test sets are reported in Table 5.1, indexed with h_0 - h_3 and s_0 - s_3 .

5.2.2 Comparing Rationalization Design Choices

From Table 5.1, I compare the performance between multiple design choices.

Regularization for hard rationalization. h_0 and h_2 are my re-implementation of [143], varying the rationale regularization hyperparameter λ_z . My experiments show that λ_z is a crucial choice. When a small λ_z is chosen (i.e., rationales are under-regularized), the model has a tendency to utilize all the available information to optimize the predictive accuracy. In h_2 , I set $\lambda_z = 0$ and the model selects 99.9% of tokens as rationales while achieving the best F₁(y) overall, which is an undesirable outcome in my case. Therefore, I increase λ_z so that only small parts of tokens are selected as rationales in h_0 . However, echoing [110], the output when varying λ_z is sensitive and unpredictable,

CHAPTER 5. STORYTELLERS

and searching for this hyperparameter is both time-consuming and energy-inefficient. I also run an experiment h_3 with the additional adversarial component proposed in [37, 288], and the evaluation metrics are not consistently improved compared to h_0 .

Binarization for soft rationalization. s_0 , s_2 and s_3 are my re-implementation of [110]. For soft rationalization, rationales are selected (i.e., binarized) after the supporter module is trained in phase one, therefore s_0 - s_3 utilize 100% of the tokens by default, and achieve the best $F_1(\mathbf{y})$ overall. I implement a straightforward approach to select rationales by setting a threshold t and make $z^i = 1$ (i.e., unmasked) if the importance score $s^i > t$ and $z^i = 0$ (i.e., masked) otherwise. Intuitively, increasing t corresponds to less selected rationales, and therefore increasing $\Pr(\mathbf{z})$. To confirm, in s_2 , I increase t until $\%(\mathbf{z})$ is exactly half of s_0 . Similarly, decreasing t corresponds to more selected rationales, and therefore decreasing $\Pr(\mathbf{z})$. In s_3 , I decrease t until $\%(\mathbf{z})$ is exactly double of s_0 .

Is domain knowledge helpful? h_1 and s_1 include domain knowledge as weak supervision. My results show that domain knowledge improves $\Pr(\mathbf{z})$ for both hard (h_1 to h_0) and soft (s_1 to s_0) rationalization methods and on both dataset, while maintaining similar $\%(\mathbf{z})$ and $F_1(\mathbf{y})$. The improvements are more substantial for soft rationalization.

Hard vs. soft rationalization. To fairly compare hard and soft rationalization methods, I choose the threshold t to keep $\%(\mathbf{z})$ the same for h_1 and s_1 .⁸ My experiments show that soft rationalization weakly supervised by domain knowledge achieves better $\Pr(\mathbf{z})$ on both datasets, and therefore I chose it for rationalizing fact-checks.

5.3 Rationalizing Fact-Checks

After determining that soft rationalization is the most appropriate method, I apply it to extract rationales from fact-checks. In this section, I introduce the dataset I collected from Snopes.com and conduct experiment with fact-checks to structurize misinformation stories.

⁸I can easily and accurately manipulate $\%(\mathbf{z})$ for soft rationalization by adjusting t ; conversely, the impact of adjusting λ_z in hard rationalization is unpredictable.

5.3.1 Fact-Check Data and Domain Knowledge

Snopes.com is a renowned fact-checking website, certified by the International Fact-Checking Network as non-partisan and transparent [212]. I collect HTML webpages of fact-check articles from Snopes.com, spanning from its founding in 1994 to the beginning of 2021.

Preprocess and statistics. I first preprocess collected fact-checks by extracting the main article content and verdicts from HTML webpages using a customized parser, and tokenizing the content with NLTK [31]. The preprocessing script is included in my released codebase.

After preprocessing, the median sequence length of fact-checks is 386 tokens, and 88.6% of fact-checks containing $\leq 1,024$ tokens. [112] found that the most informative content in fact-checks tended to be located at the head or the tail of the article content. Therefore, I set the maximum sequence length to 1,024 and truncate over-length fact-checks.

Next, I label each fact-check with a binary label depending on its verdict: (truthful) information if the verdict is at least **mostly true** and misinformation otherwise, which results in 2,513 information and 11,183 misinformation instances.

Additionally, I preemptively mask tokens that are the exact words as its verdict (e.g., “rate it as false” to “rate it as [MASK]”),⁹ otherwise predicting the verdict would be trivial and the model would copy overlapping tokens as rationales.

Ethical considerations. I consider my case of fact-checks a *fair use* under the US¹⁰ copyright law, which permits limited use of copyrighted material without the need for permission from the copyright holder.

According to [1], I discuss how my research abides the principles that are considered for a fair use judgment:

- Purpose and character of the use: I use fact-checks for noncommercial research purpose only, and additionally, using textual content for model training is considered to be transformative, cf. [15–17].
- Amount and substantiality: I present only snippets of fact-checks for illustrative purpose in the thesis (i.e., several quotes and snippets in text and figures), and only URLs to original fact-checks in my public dataset.

⁹Verdicts from Snopes.com are structured HTML fields that can be easily parsed.

¹⁰Where the authors and Snopes.com reside.

CHAPTER 5. STORYTELLERS

- Effect upon work’s value: I do not identify any adverse impact my work may have on the potential market (e.g., ads, memberships) of the copyright holder.

The end goal of my research aligns with that of Snopes.com, i.e., to rebut misinformation and to restore credibility to the online information ecosystem. I hope the aggregated knowledge of fact-checks from my models can shed light on this road and be a helpful addition to the literature.

Domain knowledge for misinformation types. The domain knowledge comes from two sources: (a) the misinformation types theorized by [271], e.g., misleading or fabricated content; and (b) certain variants of verdicts from Snopes.com such as satire or scam [239]. I combine these into a small vocabulary V_d containing 12 words, in which the first 5 are from [271] and the remaining 7 are from [239]:

“fabricated, manipulated, imposter, misleading, parody, satire, unproven, outdated, scam, legend, miscaptioned, misattributed.”

5.3.2 Experiments on Fact-Checks

I randomly split the fact-checks to 80% train, 10% dev, and 10% test sets, and adjust hyperparameters to optimize $F_1(y)$ on dev set. For initialization, I train word embeddings using Gensim [219] on the entire corpus. The final model achieves $F_1(y) = 0.75/0.74$ on the test set with/without domain knowledge.

Clustering rationales. To systematically understand extracted rationales, I cluster these rationales based on semantic similarity. For each rationale, I average word embeddings to represent the embedding of the rationale, and then run a hierarchical clustering for these embeddings. The hierarchical clustering uses cosine similarity as the distance metric, commonly used for word embeddings [170], and the complete link method [265] to obtain a relatively balanced linkage tree.

The results from the clustering are shown in Figure 5.3. From the root of the dendrogram, I can traverse its branches to find clusters until I reach a sensible threshold of cosine distance, and categorize the remaining branches and leaf nodes (i.e., rationales) to multiple clusters. Figure 5.3 shows an example visualization that contains ten clusters of rationales that are semantically similar to the domain knowledge, and leaf nodes in each cluster are aggregated to plot a word cloud, with the frequency of a node encoded as the font size of the phrase.

CHAPTER 5. STORYTELLERS

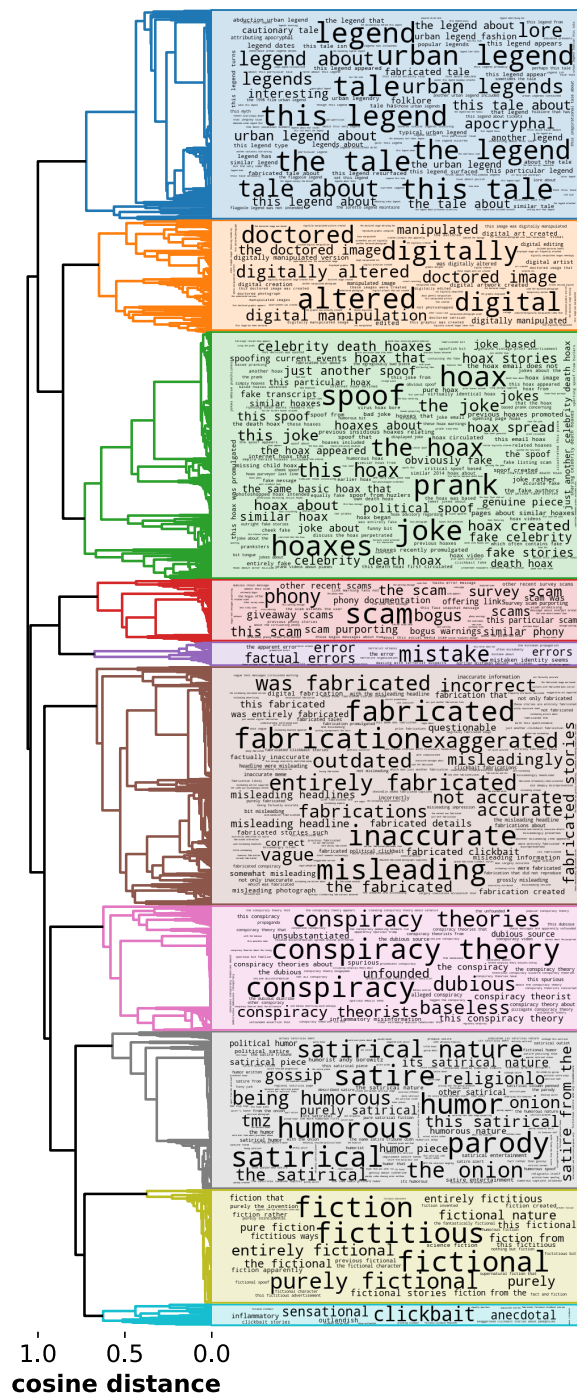


Figure 5.3: Structure of misinformation types. The ten identified clusters (colored) offer empirical confirmation of theorized misinformation types, contain novel fine-grained clusters, and reorganize the structure of misinformation stories.

CHAPTER 5. STORYTELLERS

Note that rationales extracted from soft rationalization are dependent on the chosen threshold t to binarize importance scores. The example in Figure 5.3 uses a threshold of $t = 0.01$. Varying the threshold would affect extracted rationales but mostly the ones with low prevalence, and these rare rationales also correspond to small font sizes in the word cloud. Therefore, the effect from varying t would be visually negligible in Figure 5.3.

Examples of misinformation types. Here are examples of extracted rationales for ten identified misinformation types.

For urban legends and tales ■:

“...the 1930 Colette short story *La Chienne* (The Bitch) has become an **urban legend** in that its plot is often now related as a string of events that...”

For altered or doctored images ■:

“...magazine covers of highest paid people. These **doctored images** have featured celebrities such as John Legend, Chuck Norris, Bob Dylan, Susan Boyle, and...”

For hoaxes and pranks ■:

“...This meme is a **hoax**. Nobody is (or was) licking toilets as a form of protest against Donald Trump. The images shown in the meme were taken from...”

For bogus scams ■:

“...In October 2019, I came across a decidedly bizarre version of **the scam**. This time, Nigerian astronaut Abacha Tunde was reportedly stuck in space and...”

For mistakes and errors ■:

“...noted that reports of missing children (which are typically resolved quickly) are often **mistakenly** confused by the public with relatively rare instances of...”

For fabricated content ■:

“...The Neon Nettle report was “unusual” because it was **completely fabricated**. Bono said nothing during his Rolling Stone interview about “colluding with elites”...”

For baseless conspiracies ■:

“...Furthermore, claims that COVID-19 was “manufactured,” or that it “escaped from” this Chinese lab, are nothing more than **baseless conspiracy theories**...”

CHAPTER 5. STORYTELLERS

For satires and parodies ■:

“...This item was not a factual recounting of real-life events. The article originated with a website that describes its output as being **humorous** or **satirical** in nature...”

For fictitious content ■:

“...However, both of these shocking quotes, along with the rest of article in which they are found, are **completely fictitious**. As the name of the web site implies...”

For sensational clickbait ■:

“...And Breitbart regurgitated some of the pictures as viral **clickbait** under the headline “Armed Black Panthers Lobby for Democrat Gubernatorial Candidate Stacey Abrams”...”

5.3.3 Structure of Misinformation Stories

I make the following observations from the ten clusters of misinformation types identified in Figure 5.3.

First, the clusters empirically *confirm* existing domain knowledge in V_d . Certain theorized misinformation types, such as satires and parodies ■ from [271], are identified as individual clusters from fact-checks.

Second, the clusters *complement* V_d with additional phrases describing (semantically) similar misinformation types. For example, my results add “humor” and “gossip” to the same category as satires and parodies ■ and add “tales” and “lore” to the same category as legends ■. This helps us grasp the similarity between misinformation types, and also enriches the lexicon V_d , which proves useful for subsequent analysis in § 5.4.

Third, I *discover* novel, fine-grained clusters that are not highlighted in V_d . There are multiple possible explanations as to why these misinformation types form their own clusters. Conspiracy theories ■ are often associated with intentional political campaigns [228] which can affect their semantics when referenced in fact-checks. In contrast, digital alteration ■ is a relatively recent misinformation tactic that has been enabled by technological developments such as FaceSwap [133] and DeepFake [275]. Hoaxes and pranks ■ often have a mischievous intent that distinguishes them from other clusters. Other new clusters include clickbait with inflammatory and sensational language ■ and entirely fictional content ■.

Fourth, the clusters *reorganize* the structure of these misinformation types based on their semantics, e.g., fabricated and misleading content ■ belongs to two types of misinformation in [271],

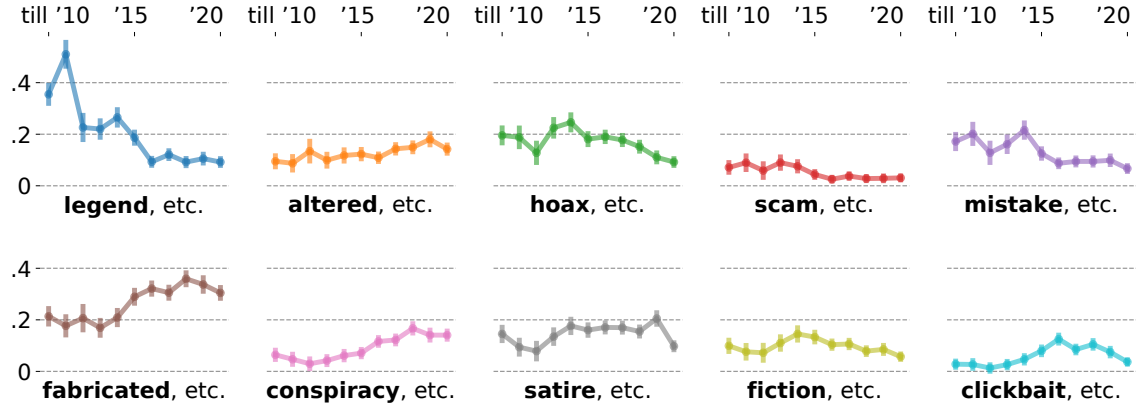


Figure 5.4: **Evolution of misinformation over the last ten years.** Conspiracy theories, fabricated content, and digital manipulation have increased in prevalence. The prevalence of (arguably) less politicized stories (e.g., legends and tales, pranks and jokes, mistakes and errors) has decreased. (95% confidence intervals.)

while in my results they are clustered together. This suggests that the semantic distance between fabricated and misleading content is less than the chosen similarity threshold, at least when these misinformation types are referred to by fact-checkers when writing articles.

Finally, the remaining words in V_d are also found in my rationales. However, due to low prevalence, they are not visible in Figure 5.3 and do not form their own clusters.

5.4 Evolution of Misinformation

In this section, I leverage the clusters of misinformation types identified by my method as a lexicon and apply it back to the my original fact-check dataset. Specifically, I analyze the evolution of misinformation types over the last ten years and compare misinformation trends around major real-world events.







5.4.1 Evolution Over the Last Ten Years

I first explore the evolution of misinformation over time. I map each fact-check article with one or more corresponding misinformation types identified by my method, and then aggregate fact-checks by year from before 2010¹¹ to the end of 2020 to estimate the relative ratio of each misinformation type.

¹¹Since there are relatively few fact-checks before 2010, I aggregate them together to the year 2010.

CHAPTER 5. STORYTELLERS

As shown in Figure 5.4,¹² the prevalence of certain misinformation types on Snopes.com has drastically changed over the last ten years.

Heavily politicized misinformation types, such as digitally altered or doctored images or photographs , fabricated and misleading content , and conspiracy theories  have nearly doubled in relative ratios over the last ten years. In contrast, the prevalence of (arguably) less politicized stories, such as legends and tales , hoaxes and pranks , and mistakes and errors  have decreased.

These trends may be a proxy for the underlying prevalence of different misinformation types within the US. Studies that measure political ideologies expressed online have documented increasing polarization over time [24, 46], which could explain increased ratios of such heavily politicized misinformation. Additionally, the convenience offered by modern digital alteration software and applications [133, 275] provides a gateway to proliferating manipulated images or photographs in the misinformation ecosystem.

Alternatively, these trends may reflect shifts in Snopes.com’s priorities. The website, launched in 1994, was initially named *Urban Legends Reference Pages*. Since then it has grown to encompass a broad spectrum of subjects. Due to its limited resources, fact-checkers from Snopes.com only cover a subset of online misinformation, and their priority is to “fact-check whatever items the greatest number of readers are asking about or searching for at any given time [240].”¹³ Given the rising impact of political misinformation in recent years [291, 292], such misinformation could reach an increasing number of Snopes.com readers, and therefore the website may dedicate more resources to fact-checking related types of misinformation. Additionally, Snopes.com has established collaborations with social media platforms, e.g., Facebook [91], to specifically target viral misinformation circulating on these platforms, where the rising meme culture could also attract Snopes.com’s attention and therefore explain a surge of digitally altered images [152, 269].

5.4.2 2016 vs. 2020 US Presidential Elections

I now compare misinformation types between the 2016 and 2020 elections. To filter for relevance, I constrain my analysis to fact-checks that (1) were published in the election years and (2) included the names of the presidential candidates and/or their running mates (e.g., “Joe Biden” and “Kamala Harris”). This results in 2,586 fact-checks for the 2016 election and 2,436 fact-checks for 2020.

¹²95% confidence intervals.

¹³Users can submit a topic to Snopes.com on its contact page [241], the results from which may affect Snopes.com’s priorities.

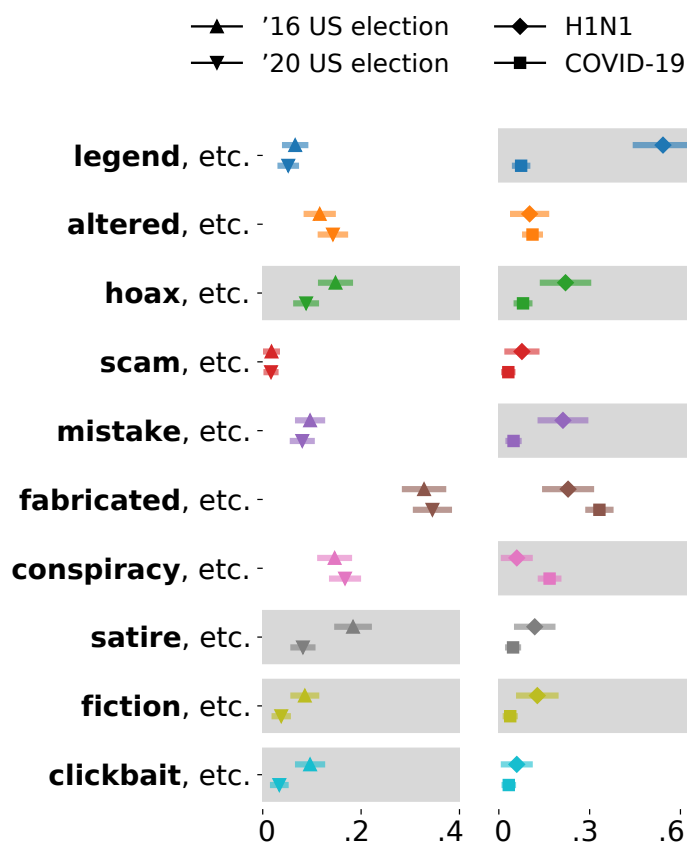


Figure 5.5: **Misinformation between notable events.** The most prevalent misinformation type for both US presidential elections is fabricated content, while the 2016 election has more hoaxes and satires. The H1N1 pandemic in 2009 has more legends and tales, while the COVID-19 pandemic attracts more conspiracy theories. (95% confidence intervals.)

The prevalence of each misinformation type is shown in Figure 5.5. I observe that the relative ratios of many misinformation types are similar between the two elections, e.g., legends and tales ■ and bogus scams ■, while the 2016 election has more hoaxes ■, satires ■, etc. The most prevalent type during both elections is fabricated and misleading content ■, next to conspiracy theories ■.

5.4.3 H1N1 vs. COVID-19 Pandemics

Finally, I compare misinformation types between the H1N1 pandemic in 2009 and the COVID-19 pandemic. For H1N1 related fact-checks, I search for keywords “flu”, “influenza”, and “H1N1” in fact-checks and constrain the publication date until the end of 2012.¹⁴ For COVID-19 related

¹⁴WHO declared an end to the global 2009 H1N1 pandemic on August 10, 2010, yet misinformation about H1N1 continues to spread [250], therefore I extend the time window by two more years.

fact-checks, I search for keywords “COVID-19” and “coronavirus”, and only consider fact-checks published in 2019 or later, which results in 833 fact-checks for the H1N1 pandemic and 656 fact-checks for COVID-19.

The relative ratio of each misinformation type is also shown in Figure 5.5. I observe that the prevalence of some misinformation types are significantly different between two pandemics, e.g., hoaxes ■, mistakes ■. Notably, the H1N1 pandemic has many more legends and tales ■, while COVID-19 has more conspiracy theories ■. The increased prevalence of COVID-19 related conspiracies aligns with recent work measuring the same phenomena [121,259], especially as the COVID-19 pandemic becomes increasingly politicized [98,227,274].

5.5 Summary of Storytellers’ Strategies

This section summarizes the chapter that structurizes misinformation types from storytellers’ strategies.

5.5.1 Research Questions and Answers

In this chapter, I investigated and answered the following RQs:

- **RQ3.1**, *what are the prevalent types of misinformation stories in the US over the last ten years?* I identify ten types of misinformation stories, including urban legends and tales, altered or doctored images, hoaxes and pranks, bogus scams, mistakes and errors, fabricated content, baseless conspiracies, satires and parodies, fictitious content, and sensational clickbait.
- **RQ3.2**, *how has the prevalence of misinformation types evolved over the last ten years?* Heavily politicized misinformation types, such as fabricated and misleading content and conspiracy theories have nearly doubled over the last ten years, while the prevalence of arguably less politicized stories, such as legends and tales, hoaxes and pranks, have decreased.
- **RQ3.3**, *how has the prevalence of misinformation types evolved between the 2016 and the 2020 US presidential elections?* The prevalence of many misinformation types are similar between the two elections, while the 2016 election has more hoaxes and satires. The most prevalent type during both elections is fabricated content and conspiracy theories.
- **RQ3.4**, *how has the prevalence of misinformation types evolved between the H1N1 and the COVID-19 pandemics?* The prevalence of certain misinformation types are significantly

different between two pandemics. Notably, the H1N1 pandemic has many more legends and tales, while COVID-19 has more conspiracy theories.

5.5.2 Limitations

There are several limitations of the study in this chapter.

I adopted a computational approach to investigate my research question, and this method inherently shares common limitations with observational studies, e.g., prone to bias and confounding [26]. Specifically, my corpus contains fact-checks from Snopes.com, one of the most comprehensive fact-checking agencies in the US. Snopes.com covers a broader spectrum of topics than politics-focused fact-checkers (e.g., PolitiFact.com, FactCheck.org),¹⁵ and thus I argue that it covers a representative sample of misinformation within the US. However, Snopes.com may not be representative of the international misinformation ecosystem [3, 77, 125]. In the future, I hope that my method can help characterize misinformation comparatively on a global scale when more structured fact-checks become available.¹⁶ Additionally, fact-checkers are time constrained, as thus the misinformation stories they cover tend to be high-profile. Therefore low-prevalence, long-tail misinformation stories may not be observed in my study. Understanding low-volume misinformation types may require a different collection of corpora other than fact-checks, e.g., a cross-platform investigation on social media conversations [2, 279].

Lastly, the misinformation types I extract from my weakly supervised approach are not validated with ground-truth labels. This is largely due to the lack of empirical knowledge on misinformation types, and therefore I am unable to provide specific guidance to annotators. Although the clusters in Figure 5.3 provide straightforward structure of misinformation stories, in future work, I plan to leverage these results to construct annotation guidelines and obtain human-identified misinformation types for further analysis.

5.5.3 Concluding Thoughts

In this chapter, I identify ten prevalent misinformation types with rationalized models on fact-checks and analyze their evolution over the last ten years and between notable events. I hope that this chapter offers an empirical lens to the systematic understanding of fine-grained misinformation types, and complements existing work investigating the misinformation problem.

¹⁵ Also note that including these additional fact-checkers in the corpus would lead to oversampling of overlapping topics (e.g., politics).

¹⁶ Less-structured and under-represented fact-checks are difficult for computational modeling [112].

Chapter 6

Conclusion

In this final chapter, I summarize the contribution of the thesis, state the impact, and finally discuss limitations and future work.

6.1 Summary of Contributions

6.2 Overview of Limitations

6.3 Concluding Remarks

Bibliography

- [1] 17 U.S.C. § 107. Limitations on exclusive rights: Fair use.
- [2] A. Abilov, Y. Hua, H. Matatov, O. Amir, and M. Naaman. Voterfraud2020: a multi-modal dataset of election fraud claims on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2021.
- [3] B. O. Ahinkorah, E. K. Ameyaw, J. E. Hagan Jr, A.-A. Seidu, and T. Schack. Rising above misinformation or fake news in africa: Another strategy to control covid-19 spread. *Frontiers in Communication*, 2020.
- [4] H. Allcott and M. Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 2017.
- [5] C. O. Alm, D. Roth, and R. Sproat. Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2005.
- [6] E. Alvarez. Youtube ceo talks misinformation, creators and comments at sxsw. Engadget, 3 2018.
- [7] M. A. Amazeen. Revisiting the epistemology of fact-checking. *Critical Review*, 27(1), 2015.
- [8] M. A. Amazeen. Checking the fact-checkers in 2008: Predicting political ad scrutiny and assessing consistency. *Journal of Political Marketing*, 15(4), 2016.
- [9] J. Anderson and L. Rainie. The future of truth and misinformation online. *Pew Research Center*, 2017.
- [10] K. Arceneaux, M. Johnson, and C. Murphy. Polarized political communication, oppositional media hostility, and selective exposure. *The Journal of Politics*, 74(1), 2012.

BIBLIOGRAPHY

- [11] 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 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2017.
- [12] S. E. Asch and H. Guetzkow. Effects of group pressure upon the modification and distortion of judgments. *Groups, leadership, and men*, 1951.
- [13] P. C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12), 2008.
- [14] 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), 2014.
- [15] Authors Guild, Inc. v. Google Inc. 954 f. supp. 2d 282 - dist. court, sd new york, 2013.
- [16] Authors Guild, Inc. v. Google Inc. 804 f. 3d 202 - court of appeals, 2nd circuit, 2015.
- [17] Authors Guild, Inc. v. Google Inc. 136 s. ct. 1658, 578 us 15, 194 l. ed. 2d 800 - supreme court, 2016.
- [18] S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2010.
- [19] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.
- [20] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239), 2015.
- [21] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open information extraction from the web. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 7, 2007.
- [22] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. <http://www.fairmlbook.org>.

BIBLIOGRAPHY

- [23] J. Bastings, W. Aziz, and I. Titov. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- [24] F. Baumann, P. Lorenz-Spreen, I. M. Sokolov, and M. Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 2020.
- [25] L. Becker, G. Erhart, D. Skiba, and V. Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, 2013.
- [26] K. Benson and A. J. Hartz. A comparison of observational studies and randomized, controlled trials. *New England Journal of Medicine*, 2000.
- [27] A. J. Berinsky. Rumors and health care reform: Experiments in political misinformation. *British Journal of Political Science*, 47(2), 2017.
- [28] N. Berman. The victims of fake news, 2017.
- [29] E. L. Bernays. *Propaganda*. Ig publishing, 1928.
- [30] 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.
- [31] S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL Interactive Presentation Sessions*, 2006.
- [32] M. M. Bradley and P. J. Lang. Affective norms for english words (anew): Instruction manual and affective ratings. Technical report, Citeseer, 1999.
- [33] S. Brown. Likert scale examples for surveys, 2010.
- [34] C. Buni and S. Chemaly. The secret rules of the internet. The Verge, Apr. 2016.
- [35] C. Burfoot and T. Baldwin. Automatic satire detection: Are you having a laugh? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2009.
- [36] C. J. Calhoun. *Social Theory and the Politics of Identity*. 1994.

BIBLIOGRAPHY

- [37] S. Carton, Q. Mei, and P. Resnick. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.
- [38] S. Chancellor and S. Counts. Measuring employment demand using internet search data. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.
- [39] 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 *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2016.
- [40] 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. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 1(CSCW), 2017.
- [41] 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. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW), 2018.
- [42] R. Chatterjee and P. Dave. Youtube set to hire more staff to review extremist video content. Independent, 12 2017.
- [43] L. Chen, R. Ma, A. Hannák, and C. Wilson. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.
- [44] 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 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2017.
- [45] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Proceedings of the IEEE/ACM International Symposium on Quality of Service (IWQoS)*, 2008.
- [46] S. Chinn, P. S. Hart, and S. Soroka. Politicization and polarization in climate change news content, 1985-2017. *Science Communication*, 2020.

BIBLIOGRAPHY

- [47] G. L. Ciampaglia, A. Mantzarlis, G. Maus, and F. Menczer. Research challenges of digital misinformation: Toward a trustworthy web. *AI Magazine*, 39(1), 2018.
- [48] 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), 2015.
- [49] J. Constine. Facebook tries fighting fake news with publisher info button on links, 10 2017.
- [50] J. Constine. Facebook reveals russian troll content, shuts down 135 ira accounts. Tech Crunch, 3 2018.
- [51] N. A. Cooke. Posttruth, truthiness, and alternative facts: Information behavior and critical information consumption for a new age. *The Library Quarterly*, 87(3), 2017.
- [52] 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 *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- [53] 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 *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.
- [54] 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.
- [55] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [56] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace. Eraser: a benchmark to evaluate rationalized nlp models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [57] N. Diakopoulos and M. Naaman. Towards quality discourse in online news comments. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2011.

BIBLIOGRAPHY

- [58] J. Dorsey. Twitter: Transparency and accountability. House Energy and Commerce Committee, 9 2018.
- [59] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv*, 2017.
- [60] A. M. Enders, J. E. Uscinski, C. Klofstad, and J. Stoler. The different forms of covid-19 misinformation and their consequences. *Harvard Kennedy School Misinformation Review*, 2020.
- [61] R. Epstein and R. E. Robertson. The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 112(33), Aug. 2015.
- [62] R. Epstein, R. E. Robertson, D. Lazer, and C. Wilson. Suppressing the search engine manipulation effect (SEME). *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 1(CSCW), Dec. 2017.
- [63] A. Esuli and F. Sebastiani. Sentiwordnet: a publicly available lexical resource for opinion mining. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2006.
- [64] D. Evon. Is greta thunberg the ‘highest paid activist’? *Snopes.com*, 2019.
- [65] Facebook. Community standards, 2018.
- [66] Facebook. Investments to fight polarization, 2020.
- [67] D. I. H. Farías, V. Patti, and P. Rosso. Irony detection in twitter: The role of affective content. *ACM Transactions on Internet Technology (ToIT)*, 16(3), 2016.
- [68] R. Farley. Trump said obama’s grandmother caught on tape saying she witnessed his birth in kenya. *PolitiFact.com*, 7 2011.
- [69] J. Farrell, K. McConnell, and R. Brulle. Evidence-based strategies to combat scientific misinformation. *Nature Climate Change*, 2019.
- [70] E. Fast, B. Chen, and M. S. Bernstein. Empath: Understanding topic signals in large-scale text. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2016.

BIBLIOGRAPHY

- [71] E. Fast, T. Vachovsky, and M. S. Bernstein. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2016.
- [72] 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)*, 2017.
- [73] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2012.
- [74] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 59(7), 2016.
- [75] S. Fiegerman. Facebook, google, twitter to fight fake news with 'trust indicators', 2017.
- [76] F. Figueiredo, J. M. Almeida, M. A. Gonçalves, and F. Benevenuto. On the dynamics of social media popularity: A youtube case study. *ACM Transactions on Internet Technology (ToIT)*, 14(4), 2014.
- [77] 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.
- [78] 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. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW), 2018.
- [79] 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), 2015.
- [80] U. Friedman. The real-world consequences of “fake news”, 12 2017.
- [81] V. Gadde and Y. Roth. Enabling further research of information operations on twitter. *Twitter Blog*, 17, 2018.
- [82] 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), 2013.

BIBLIOGRAPHY

- [83] M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57(3), 2019.
- [84] M. Gentzkow, J. M. Shapiro, and D. F. Stone. Media bias in the marketplace: Theory. In *Handbook of media economics*, volume 1. 2015.
- [85] S. Gibbs. Google says ai better than humans at scrubbing extremist youtube content. *The Guardian*, 8 2017.
- [86] E. Gilbert, C. Lampe, A. Leavitt, K. Lo, and L. Yarosh. Conceptualizing, creating, & controlling constructive and controversial comments: A cscw research-athon. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2017.
- [87] T. Gillespie. There’s a reason that misleading claims of bias in search and social media enjoy such traction. *Medium*, 8 2018.
- [88] A. Glaser. Youtube is adding fact-check links for videos on topics that inspire conspiracy theories. *Slate*, 8 2018.
- [89] R. González-Ibáñez, S. Muresan, and N. Wacholder. Identifying sarcasm in twitter: a closer look. In *Proceedings of the ACM Conference on Human Language Technology (HLT)*, 2011.
- [90] Google. Google fact checks feature, 2018.
- [91] V. Green and D. Mikkelsen. A message to our community regarding the facebook fact-checking partnership. *Snopes.com*.
- [92] N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 2019.
- [93] 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.
- [94] M. Gupta, P. Zhao, and J. Han. Evaluating event credibility on twitter. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 2012.
- [95] K. Haglin. The limitations of the backfire effect. *Research & Politics*, 4(3), 2017.

BIBLIOGRAPHY

- [96] 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 *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2014.
- [97] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson. Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2017.
- [98] P. S. Hart, S. Chinn, and S. Soroka. Politicization and polarization in covid-19 news coverage. *Science Communication*, 2020.
- [99] K. S. Hasan and V. Ng. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2013.
- [100] N. Hassan, B. Adair, J. T. Hamilton, C. Li, M. Tremayne, J. Yang, and C. Yu. The quest to automate fact-checking. *world*, 2015.
- [101] J. Hawley. Ending support for internet censorship act, 2019.
- [102] R. Hobbs and A. Jensen. The past, present, and future of media literacy education. *Journal of media literacy education*, 1(1), 2009.
- [103] J. L. Hochschild and K. L. Einstein. *Do facts matter?: Information and misinformation in American politics*, volume 13. University of Oklahoma Press, 2015.
- [104] D. Hu, S. Jiang, R. E. Robertson, and C. Wilson. Auditing the partisanship of google search snippets. In *Proceedings of the Web Conference (WWW)*, 2019.
- [105] 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 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2015.
- [106] B. Hutchinson and M. Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2019.

BIBLIOGRAPHY

- [107] 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.
- [108] B. Jackson. Factcheck, 2018.
- [109] S. Jain and B. C. Wallace. Attention is not explanation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [110] S. Jain, S. Wiegrefe, Y. Pinter, and B. C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.
- [111] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (ToCHI)*, 25(2), 2018.
- [112] S. Jiang, S. Baumgartner, A. Ittycheriah, and C. Yu. Factoring fact-checks: Structured information extraction from fact-checking articles. In *Proceedings of the Web Conference (WWW)*, 2020.
- [113] S. Jiang, L. Chen, A. Mislove, and C. Wilson. On ridesharing competition and accessibility: Evidence from uber, lyft, and taxi. In *Proceedings of the Web Conference (WWW)*, 2018.
- [114] S. Jiang, M. Metzger, A. Flanagin, and C. Wilson. Modeling and measuring expressed (dis)belief in (mis)information. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2020.
- [115] S. Jiang, R. E. Robertson, and C. Wilson. Bias misperceived: The role of partisanship and misinformation in youtube comment moderation. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2019.
- [116] S. Jiang, R. E. Robertson, and C. Wilson. Reasoning about political bias in content moderation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.
- [117] S. Jiang and C. Wilson. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW), 2018.

BIBLIOGRAPHY

- [118] S. Jiang and C. Wilson. Structurizing misinformation stories via rationalizing fact-checks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.
- [119] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang. News credibility evaluation on microblog with a hierarchical propagation model. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, 2014.
- [120] Z. Jin, J. Cao, Y. Zhang, and J. Luo. News verification by exploiting conflicting social viewpoints in microblogs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2016.
- [121] D. Jolley and J. L. Paterson. Pylons ablaze: Examining the role of 5g covid-19 conspiracy beliefs and support for violence. *British journal of social psychology*, 2020.
- [122] K. Joseph, L. Friedland, W. Hobbs, D. Lazer, and O. Tsur. Constance: Modeling annotation contexts to improve stance classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [123] D. Jurafsky and J. H. Martin. *Speech and language processing*, volume 3. Pearson London, 2014.
- [124] B. Kamisar. Conservatives cry foul over controversial group’s role in youtube moderation. The Hill, 3 2018.
- [125] 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. *Social Science Research Network (SSRN)*, 2018.
- [126] L. K. Kaye, S. A. Malone, and H. J. Wall. Emojis: Insights, affordances, and possibilities for psychological science. *Trends in cognitive sciences*, 21(2), 2017.
- [127] M. W. Kearney. Trusting news project report. *Reynolds Journalism Institute*, 7 2017.
- [128] 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 *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2018.

BIBLIOGRAPHY

- [129] D. L. Kincaid. From innovation to social norm: Bounded normative influence. *Journal of health communication*, 2004.
- [130] G. King and R. Nielsen. Why propensity scores should not be used for matching. 2016.
- [131] K. Klonick. Re-shaming the debate: Social norms, shame, and regulation in an internet age. *SSRN Electronic Journal*, 2015.
- [132] K. Klonick. The new governors: The people, rules, and processes governing online speech. (ID 2937985), 2017.
- [133] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.
- [134] T. Kriplean, C. Bonnar, A. Borning, B. Kinney, and B. Gill. Integrating on-demand fact-checking with public dialogue. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2014.
- [135] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the Web Conference (WWW)*, 2016.
- [136] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2015.
- [137] 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), 2013.
- [138] B. Laslett. Unfeeling knowledge: Emotion and objectivity in the history of sociology. In *Sociological Forum*, volume 5, 1990.
- [139] 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), 2016.
- [140] 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.

BIBLIOGRAPHY

- [141] 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*, 2018.
- [142] M. Lechner et al. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3), 2011.
- [143] T. Lei, R. Barzilay, and T. Jaakkola. Rationalizing neural predictions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.
- [144] M. S. Levendusky. Why do partisan media polarize viewers? *American Journal of Political Science*, 57(3), 2013.
- [145] S. Levin. Google to hire thousands of moderators after outcry over youtube abuse videos. *The Guardian*, 12 2017.
- [146] 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), 2012.
- [147] M. Li, Q. Lu, and Y. Long. Are manually prepared affective lexicons really useful for sentiment analysis. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.
- [148] 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 *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2013.
- [149] 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 ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2014.
- [150] 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 ACM Conference on Human Factors in Computing Systems (CHI)*, 2014.
- [151] 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 ACM International Conference on Information and Knowledge Management (CIKM)*, 2014.

BIBLIOGRAPHY

- [152] C. Ling, I. AbuHilal, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini. Dissecting the meme magic: Understanding indicators of virality in image memes. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 5(CSCW), 2021.
- [153] Z. C. Lipton. The mythos of model interpretability. *Queue*, 2018.
- [154] Y. Liu, C. Kliman-Silver, and A. Mislove. The tweets they are a-changin’: Evolution of twitter users and behavior. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2014.
- [155] 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.
- [156] 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 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2016.
- [157] C. Lumezanu, N. Feamster, and H. Klein. # bias: Measuring the tweeting behavior of propagandists. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2012.
- [158] 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 Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.
- [159] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [160] A. Magdy and N. Wanas. Web-based statistical fact checking of textual documents. In *Proceedings of the International Workshop on Search and Mining User-Generated Contents*, 2010.
- [161] G. E. Marcus. The sentimental citizen: Emotion in democratic politics. *Perspectives on Politics*, 2002.
- [162] D. B. Margolin, A. Hannak, and I. Weber. Political fact-checking on twitter: When do corrections have an effect? *Political Communication*, 2017.

BIBLIOGRAPHY

- [163] 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), 1997.
- [164] M. Masnick. Internet content moderation isn’t politically biased, it’s just impossible to do well at scale. *Techdirt*, 8 2018.
- [165] 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.
- [166] H. McCracken. Youtube will use wikipedia to fact-check internet hoaxes. *Fast Company*, 3 2018.
- [167] M. Metzger, A. Flanagin, P. Mena, S. Jiang, and C. Wilson. From dark to light: The many shades of sharing misinformation online. *Media and Communication*, 9(1), 2021.
- [168] D. Mikkelsen. Barack obama birth certificate: Is barack obama’s birth certificate a forgery? *Snopes.com*, 8 2011.
- [169] D. Mikkelsen. *Snopes.com*, 2021.
- [170] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *Workshop Proceedings of the International Conference on Learning Representations (ICLR Workshop)*, 2013.
- [171] J. M. Miotto and E. G. Altmann. Predictability of extreme events in social media. *PLoS One*, 9(11), 2014.
- [172] A. Mitchell, J. Gottfried, J. Kiley, and K. E. Matsa. Political polarization & media habits. *Pew Research Center*, 21, 2014.
- [173] S. M. Mohammad and S. Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2), 2015.
- [174] 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 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010.
- [175] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 2013.

BIBLIOGRAPHY

- [176] S. M. Mohammad and P. D. Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 2013.
- [177] M. Mondal, L. A. Silva, and F. Benevenuto. A measurement study of hate speech in social media. 2017.
- [178] S. Morgan. Fake news, disinformation, manipulation and online tactics to undermine democracy. *Journal of Cyber Policy*, 3(1), 2018.
- [179] D. Z. Morris. Hate speech: Youtube restricts extremist videos. *Fortune*, 8 2017.
- [180] 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 *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2013.
- [181] S. Murray and C. Lima. Trump accuses social media giants of ‘silencing millions of people’. *Politico*, 2018.
- [182] S. News. New youtube recruits to monitor online extremist propaganda ‘wrong approach’. *Sputnik International*, 6 2017.
- [183] NewsBusters. Don’t believe the liberal “fact-checkers”!, 2018.
- [184] C. Newton. Why twitter should ignore the phony outrage over “shadow banning”. *The Verge*, 7 2018.
- [185] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2002.
- [186] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2), 1998.
- [187] R. K. Nielsen and L. Graves. “news you don’t believe”: Audience perspectives on fake news. *Reuters Institute*, 2017.
- [188] M. Nunez. Former facebook workers: We routinely suppressed conservative news. *Gizmodo*, 5 2016.
- [189] B. Nyhan and J. Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2), 2010.

BIBLIOGRAPHY

- [190] 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), 2015.
- [191] B. Nyhan, J. Reifler, and P. A. Ubel. The hazards of correcting myths about health care reform. *Medical care*, 51(2), 2013.
- [192] A. Olteanu, C. Castillo, J. Boy, and K. R. Varshney. The effect of extremist violence on hateful speech online. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2018.
- [193] A. Olteanu, O. Varol, and E. Kiciman. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2017.
- [194] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2011.
- [195] V. Palladino. Youtube to fight fake news with links to real news and context. *Ars Technica*, 7 2018.
- [196] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(10), 2009.
- [197] B. Pang, L. Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 2008.
- [198] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.
- [199] 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 ACM International Conference on Supporting Group Work (GROUP)*, 2016.
- [200] J. Pearl. *Causality*. Cambridge university press, 2009.
- [201] J. W. Pennebaker, M. E. Francis, and R. J. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71, 2001.

BIBLIOGRAPHY

- [202] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- [203] G. Pennycook and D. Rand. Examining false beliefs about voter fraud in the wake of the 2020 presidential election. *Harvard Kennedy School Misinformation Review*, 2021.
- [204] G. Pennycook and D. G. Rand. Assessing the effect of “disputed” warnings and source salience on perceptions of fake news accuracy. 2017.
- [205] S. Phadke, M. Samory, and T. Mitra. What makes people join conspiracy communities? role of social factors in conspiracy engagement. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 4(CSCW), 2021.
- [206] R. Plutchik. Emotions: A general psychoevolutionary theory. *Approches to emotion*, 1984.
- [207] B. Popken. Twitter deleted 200,000 russian troll tweets, 2 2018.
- [208] 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*, 2018.
- [209] N. Y. Post. Youtube committing \$25m to fight fake news. *New York Post*, 7 2018.
- [210] W. J. Potter. *Media literacy*. Sage Publications, 2018.
- [211] 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.
- [212] Poynter. Verified signatories of the ifcn code of principles, 2018.
- [213] D. PreoŃiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar. Beyond binary labels: political ideology prediction of twitter users. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.
- [214] 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 (EMNLP)*, 2011.

BIBLIOGRAPHY

- [215] 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), 2012.
- [216] 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 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.
- [217] RedStateMedia. Donald trump tells jake tapper he won’t denounce david duke or the kkk. YouTube, 2 2016.
- [218] RedStateMedia. Youtube homepage for redstatemedia. YouTube, 2018.
- [219] R. Rehurek and P. Sojka. Gensim - statistical semantics in python. *Gensim.org*, 2011.
- [220] V. Richardson. Conservative project seeks to fact-check the fact-checkers accused of liberal bias, 3 2018.
- [221] M. A. Riordan. Emojis as tools for emotion work: Communicating affect in text messages. *Journal of Language and Social Psychology*, 36(5), 2017.
- [222] 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 on Human-Computer Interaction (PACM HCI)*, 2(CSCW), 2018.
- [223] R. E. Robertson, D. Lazer, and C. Wilson. Auditing the personalization and composition of politically-related search engine results pages. In *Proceedings of the Web Conference (WWW)*, 2018.
- [224] 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*. 2000.
- [225] 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), 1995.
- [226] P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 1983.

BIBLIOGRAPHY

- [227] H. Rothgerber, T. Wilson, D. Whaley, D. L. Rosenfeld, M. Humphrey, A. Moore, and A. Bihl. Politicizing the covid-19 pandemic: ideological differences in adherence to social distancing. *PsyArXiv*, 2020.
- [228] M. Samory and T. Mitra. ‘the government spies using our webcams’ the language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 2(CSCW), 2018.
- [229] 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*, 2014.
- [230] S. Serrano and N. A. Smith. Is attention interpretable? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.
- [231] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer. The spread of misinformation by social bots. *arXiv*, 2017.
- [232] M. Shapiro. Running the data on politifact shows bias against conservatives, 12 2016.
- [233] A. Sharockman. Politifact, 2021.
- [234] Q. Shen and C. Rose. The discourse of online content moderation: Investigating polarized user responses to changes in reddit’s quarantine policy. In *Workshop proceedings of the Annual Meeting of the Association for Computational Linguistics (ALW3 ACL)*, 2019.
- [235] Q. Shen, M. Yoder, Y. Jo, and C. Rose. Perceptions of censorship and moderation bias in political debate forums. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2018.
- [236] S. Sheth. Facebook takes down over 200 accounts and pages run by the ira, a notorious russian troll farm, 4 2018.
- [237] B. Shi and T. Wenering. Fact checking in heterogeneous information networks. In *Companion Proceedings of the Web Conference (WWW Companion)*, 2016.
- [238] 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*, 2017.

BIBLIOGRAPHY

- [239] Snopes.com. Fact check ratings, 2021.
- [240] Snopes.com. How does snopes decide what to write about?, 2021.
- [241] Snopes.com. Submit a topic, 2021.
- [242] R. Snyder. Pro-rubio super pac ad tying trump to kkk misses the mark. *PolitiFact*, 2 2016.
- [243] 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 *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2018.
- [244] K. Starbird, A. Arif, and T. Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*, 3(CSCW), 2019.
- [245] J. E. Stets and P. J. Burke. Identity theory and social identity theory. *Social psychology quarterly*, 2000.
- [246] P. J. Stone, D. C. Dunphy, and M. S. Smith. The general inquirer: A computer approach to content analysis. 1966.
- [247] C. T. Street and K. W. Ward. Improving validity and reliability in longitudinal case study timelines. *European Journal of Information Systems*, 21(2), 2012.
- [248] 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.
- [249] R. Suarez and K. Flynn. Facebook, twitter issue policy changes to manage fake news and hate speech, 2017.
- [250] M. E. Sundaram, D. L. McClure, J. J. VanWormer, T. C. Friedrich, J. K. Meece, and E. A. Belongia. Influenza vaccination is not associated with detection of noninfluenza respiratory viruses in seasonal studies of influenza vaccine effectiveness. *Clinical infectious diseases*, 2013.
- [251] 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*, 2017.

BIBLIOGRAPHY

- [252] H. Tajfel and J. C. Turner. The social identity theory of intergroup behavior. *Psychology of intergroup relations*, 1986.
- [253] M. Tambuscio, G. Ruffo, A. Flammini, and F. Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the Web Conference (WWW)*, 2015.
- [254] 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), 2010.
- [255] F. J. Thoemmes and E. S. Kim. A systematic review of propensity score methods in the social sciences. *Multivariate behavioral research*, 46(1), 2011.
- [256] S. Tschiatschek, A. Singla, M. Gomez Rodriguez, A. Merchant, and A. Krause. Fake news detection in social networks via crowd signals. In *Companion Proceedings of the Web Conference (WWW Companion)*, 2018.
- [257] Twitter. Rules and policies, 2018.
- [258] Twitter. About different types of tweets, 2020.
- [259] J. E. Uscinski, A. M. Enders, C. Klofstad, M. Seelig, J. Funchion, C. Everett, S. Wuchty, K. Premaratne, and M. Murthi. Why do people believe covid-19 conspiracy theories? *Harvard Kennedy School Misinformation Review*, 2020.
- [260] N. Usher. How republicans trick facebook and twitter with claims of bias. *The Washington Post*, 8 2018.
- [261] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
- [262] 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), 2018.
- [263] S. Volkova and J. Y. Jang. Misleading or falsification? inferring deceptive strategies and types in online news and social media. In *Companion Proceedings of the Web Conference (WWW Companion)*, 2018.

BIBLIOGRAPHY

- [264] 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 Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 2, 2017.
- [265] E. M. Voorhees. Implementing agglomerative hierarchic clustering algorithms for use in document retrieval. *Information Processing & Management*, 1986.
- [266] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380), 2018.
- [267] W. Y. Wang. “liar, liar pants on fire”: A new benchmark dataset for fake news detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.
- [268] X. Wang, C. Yu, S. Baumgartner, and F. Korn. Relevant document discovery for fact-checking articles. In *Companion Proceedings of the Web Conference (WWW Companion)*, 2018.
- [269] Y. Wang, F. Tamahsbi, J. Blackburn, B. Bradlyn, E. De Cristofaro, D. Magerman, S. Zannettou, and G. Stringhini. Understanding the use of fauxtography on social media. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2021.
- [270] 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.
- [271] C. Wardle. Fake news. it’s complicated. *First Draft News*, 2017.
- [272] H. Wasserman and D. Madrid-Morales. An exploratory study of “fake news” and media trust in kenya, nigeria and south africa. *African Journalism Studies*, 2019.
- [273] 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), 2015.
- [274] O. Weisel. Vaccination as a social contract: The case of covid-19 and us political partisanship. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 118(13), 2021.
- [275] M. Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 2019.

BIBLIOGRAPHY

- [276] M. Wiegand, J. Ruppenhofer, A. Schmidt, and C. Greenberg. Inducing a lexicon of abusive words—a feature-based approach. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.
- [277] S. Wiegrefe and Y. Pinter. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [278] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.
- [279] T. Wilson and K. Starbird. Cross-platform disinformation campaigns: lessons learned and next steps. *Harvard Kennedy School Misinformation Review*, 2020.
- [280] T. Wood and E. Porter. The elusive backfire effect: mass attitudes? steadfast factual adherence. *Political Behavior*, 2016.
- [281] A. Woodruff, S. E. Fox, S. Roussos-Schindler, and J. Warshaw. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 2018.
- [282] Y. Wu, P. K. Agarwal, C. Li, J. Yang, and C. Yu. Toward computational fact-checking. *Proceedings of the VLDB Endowment*, 7(7), 2014.
- [283] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the Web Conference (WWW)*, 2017.
- [284] Z. Xing, J. Pei, and E. Keogh. A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, 12(1), 2010.
- [285] D. Yang, A. Lavie, C. Dyer, and E. Hovy. Humor recognition and humor anchor extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2015.
- [286] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [287] YouTube. Community guidelines, 2018.

BIBLIOGRAPHY

- [288] M. Yu, S. Chang, Y. Zhang, and T. Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [289] O. Zaidan, J. Eisner, and C. Piatko. Using “annotator rationales” to improve machine learning for text categorization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2007.
- [290] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the ACM Internet Measurement Conference (IMC)*, 2018.
- [291] S. Zannettou, T. Caulfield, B. Bradlyn, E. De Cristofaro, G. Stringhini, and J. Blackburn. Characterizing the use of images in state-sponsored information warfare operations by russian trolls on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 2020.
- [292] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the ACM Web Science Conference (WebSci)*, 2019.
- [293] 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 ACM Conference on Learning@Scale*, 2017.
- [294] X. Zhou, X. Wan, and J. Xiao. Attention-based lstm network for cross-lingual sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.